frontiers | Frontiers in Computer Science

TYPE Methods
PUBLISHED 05 January 2023
DOI 10.3389/fcomp.2022.966319

OPEN ACCESS

EDITED BY
Liang Men,
Liverpool John Moores University,
United Kingdom

REVIEWED BY
Bruce Hunter Thomas,
University of South Australia, Australia
Andrew Robb,
Clemson University, United States

*CORRESPONDENCE
Anthony Steed
A.Steed@ucl.ac.uk

SPECIALTY SECTION
This article was submitted to
Human-Media Interaction,
a section of the journal
Frontiers in Computer Science

RECEIVED 10 July 2022
ACCEPTED 21 November 2022
PUBLISHED 05 January 2023

CITATION
Steed A, Archer D, Brandstätter K, Congdon BJ, Friston S, Ganapathi P, Giunchi D, Izzouzi L, Park GW, Swapp D and Thiel FJ (2023) Lessons learnt running distributed and remote mixed reality experiments. *Front. Comput. Sci.* 4:966319. doi: 10.3389/fcomp.2022.966319

Lessons learnt running distributed and remote mixed reality experiments

Anthony Steed[1]*, Daniel Archer[1], Klara Brandstätter[1], Ben J. Congdon[1], Sebastian Friston[1], Priya Ganapathi[2], Daniele Giunchi[1], Lisa Izzouzi[1], Gun Woo (Warren) Park[1,3], David Swapp[1] and Felix J. Thiel[1]

[1]Department of Computer Science, University College London, London, United Kingdom, [2]Department of Design, Indian Institute of Technology Guwahati, Guwahati, India, [3]Department of Computer Science, University of Toronto, Toronto, ON, Canada

One traditional model of research on mixed-reality systems, is the laboratory-based experiment where a number of small variants of a user experience are presented to participants under the guidance of an experimenter. This type of experiment can give reliable and generalisable results, but there are arguments for running experiments that are distributed and remote from the laboratory. These include, expanding the participant pool, reaching specific classes of user, access to a variety of equipment, and simply because laboratories might be inaccessible. However, running experiments out of the laboratory brings a different set of issues into consideration. Here, we present some lessons learnt in running eleven distributed and remote mixed-reality experiments. We describe opportunities and challenges of this type of experiment as well as some technical lessons learnt.

KEYWORDS

virtual reality, mixed reality, distributed of experiments, avatars, social virtual environments

1. Introduction

Our laboratory at University College London has a long history of running user experiments in virtual reality (VR), augmented reality (AR), and other novel user interfaces. Over the years, we have built up significant infrastructure for running user studies in the lab, but during 2020 and 2021 we had extremely limited access to our labs.

Fortunately, we are on the cusp of the commercialisation of VR and associated technologies. There are now millions of consumer VR systems out there, and this gives us an opportunity to run some of our studies out of the lab. While there are still many reasons for running experiments in the laboratory such as control over the protocol, access to novel display devices or access to monitoring equipment, with careful preparation some studies can be run out of the laboratory or in a distributed manner where the experimenter does not directly supervise participants. Of course the studies would need to be designed for common consumer equipment, run on heterogeneous equipment, and they would need to be self-explanatory.

In this position paper, we elaborate on our evolving practice in developing experiments out of the laboratory, but also discuss our rationale for proposing to continue at least part of our work in this manner. After discussing related work, we give short outlines of four example experiments in some more detail to give an overview of the types of experiment that our laboratory runs. Our laboratory has run over a dozen similar studies over the past few years. Some of these other studies are submitted for publication, were written up in unpublished dissertations or were pilots for larger trials now under way.

In Section 3, we discuss the challenges and opportunities of running experiments out of the laboratory. We then give some technical lessons learnt that we hope are of general utility. The challenges, opportunities and lessons learnt are not derived in a formal way from our studies, but are taken from post-experiment debriefs, notes taken during studies or changes made during study development.

2. Related work

Until the 2010s, most research on immersive systems was undertaken in academic laboratories using high-end equipment and relatively niche software toolkits. While open source toolkits such as VRJuggler (Bierbaum et al., 2001) or commercial toolkits such as WorldViz[1] allowed the distribution of code and applications that would run on many different configurations, it was still difficult to support a broad range of users, and thus we do not know of any large scale studies run in a distributed manner until the advent of large-volume consumer systems. Once these were available, researchers, including ourselves, started to run studies out of the laboratory or "in the wild" (Steed et al., 2016; Mottelson and Hornbæk, 2017). The early justifications for this included exploring the potential of this form of experiment, validating lab-run experiments or simply the exposure of the research work to a broader audience. The COVID-19 pandemic then forced a lot of labs to adapt their work, to either run remotely (Steed et al., 2020b) or switch to more reflective, simulation-based or design-focused work.

The move to online, distributed and unsupervised studies had precedence in a variety of efforts in other science areas, sometimes referred to by the terms *citizen science* (Silvertown, 2009) or *crowd sourcing* (Estellés-Arolas and González-Ladrón-De-Guevara, 2012). The former generally refers to volunteer efforts in science programmes such as data collection or data validation. The latter more typically refers to targeted efforts to recruit paid or volunteer recruits to do a specific experiment. Crowdsourcing programmes can utilise paid services such as Amazon Mechanical Turk[2], Prolific[3], or similar, which provide workers to complete jobs, such as experiments, that can be conducted online. Their use in general human-computer interaction studies has a long history (Kittur et al., 2008). The specific difficulty for MR experiments is that any potential participants must have access to the correct equipment and be familiar with using it themselves (Kelly et al., 2021).

A number of different ways of running remote and distributed studies have been described (Mottelson et al., 2021; Ratcliffe et al., 2021; Steed et al., 2021; Zhao et al., 2021). For example, Saffo et al. investigate the running of experiments on social VR platforms (Saffo et al., 2021). As with prior work on running experiments on desktop social virtual environments [e.g., Friedman et al.'s (2007) study on spatial behaviour in SecondLife], one problem with the approach is the lack of precise control the experimenter has, especially if they are running the study unsupervised. A second problem is exporting data in a reliable and regulation-compliant way. For example, if the experimenter needs to follow the EU's General Data Protection Regulation (GDPR) they may not be allowed to transfer user data to or from outside the EU. This might prevent them using social VR services that are hosted in part in the USA, for example. Williamson et al. (2021) used the Mozilla Hubs system for their study. Being open source, this allowed the experimenter to closely control the experiment. However, Mozilla Hubs is a relatively complex system that is not straightforward to develop for. Thus, while we do not know of a definitive survey, our observation is that most experiments are built with current game engine systems such as Unity or Unreal, as these have direct support for the consumer VR devices and have a lot of example code to build upon.

Thus, assuming that a MR experiment can be developed, we still have to facilitate its distribution to potential participants. Recently Radiah et al. developed a framework that considered the recruitment of participants, how they would get access to the correct software and how they would run the experiment (Radiah et al., 2021). As discussed in that paper, and in discussion on research-related forum (e.g., the Distributed-VR3DUI Slack[4]), commonly participants are recruited from social media posts, through websites such as XRDRN[5] or through university mailing lists. However, reaching participants with the correct equipment can be challenging.

A distributed and remote experiment might be single-participant at a time or multiple participants. Thus, considerations of the technology of networking for VR come in to play (Singhal and Zyda, 1999; Steed and Oliveira, 2009). Indeed, a very active research area is the affordances of social interaction on such systems (Biocca et al., 2003; Oh et al., 2018). The problem with commercial platforms has already

[1] https://www.worldviz.com/

[2] https://www.mturk.com/

[3] https://www.prolific.co/

[4] https://join.slack.com/t/distributed-vr3dui/shared_invite/zt-1a5zppgmv-1bP0js118NiSz12VUnxfJg

[5] https://www.xrdrn.org/

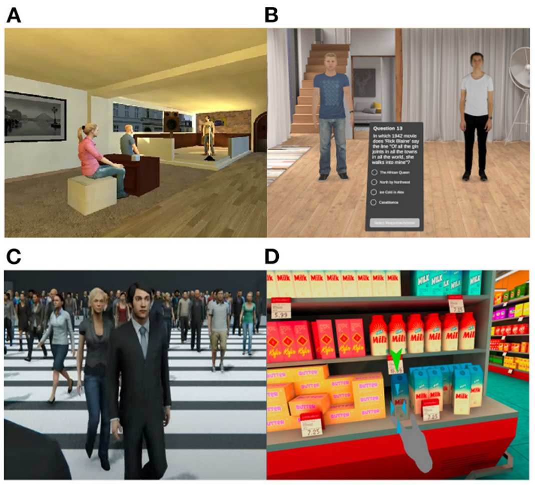

FIGURE 1
Images from four of our recent experiments. **(A)** Our first in the wild study, where the participants watched a singer perform in a bar environment. **(B)** A VR environment simulating a potential AR conferencing situation. **(C)** An experiment on virtual crowds. **(D)** An experiment on seated vs. standing locomotion.

been made, but it is worth noting that there are huge numbers of commercial platforms. Ryan Schulz's blog lists over 160 at the time of writing[6]. Commercial toolkits such as Photon are an option for developing distributed experiments[7], though there are downsides to using commercial back-end services (such as GDPR compliance). Recently some of the authors released Ubiq, an open-source toolkit for the construction of social VR experiences (Friston et al., 2021).

3. Example scenarios

In this section, we give short descriptions of four experiments that we ran out of the laboratory and then a very brief overview of seven other studies. Later we discuss the challenges, opportunities and lessons learnt from these and several other studies. All of our studies are written in Unity. No specific toolkits were used, but some of the studies used assets purchased from the Unity Asset Store. All code was developed by our team. We discuss four of our studies in depth to give an insight into the range of technical and logistical options chosen. We give a main lesson learnt for each of these four studies and in Sections 4 and 5 list further general issues and lesson. For each of these four studies, we refer the interested reader to the associated papers for more specific details of protocols and results. We then give a very brief overview of seven other studies to give a flavour of the range of studies done remotely in our laboratory.

3.1. Singer in the bar

The first experiment we discuss is our first experiment run in the wild (Steed et al., 2016). This was developed for Samsung Gear VR and Google Cardboard. Thus it had limited interaction and this was primarily through eyegaze. Recently we have updated this for Oculus Quest 1/2 for which the main modification was to include head and hand-tracking (Steed et al., 2022). The user community in 2015/2016 was still rather small, so invitations were sent out over email. We also did publicity through an Oculus development competition to which the Gear VR version of the study was sent and through social media. The app was made available by SideloadVR (a service for side loading content onto the Android phones used—it is now defunct but SideQuestVR performs a similar function for recent devices) or by downloading the application as an APK from our website. Aside from completely unsupervised use, we also showed the application to visitors at regular open events. In this case the participant was shown the device and told how to operate the controls, but no further advice was given. The experiment was designed to be used unsupervised, and thus the full instructions were given inside the application. Consent to collect data was also sought inside the application. If participants did not consent, or could not consent because they were not adults, then they could still visit the main scene. In the main scene, the participants sat in the location of the pink-shirted avatar in Figure 1A, and watched a singer on the stage. There were three binary factors for the experiment, leading to eight versions of the experience: being embodied or not in the avatar (there were male and female self-avatars), whether the singer looked directly at them or not, and whether they were instructed to tap along to the beat or not. Of 115 participants who started the application and consented to data collection, 85 completed the study and of these, 59 data sets were usable. Data was exported from the application by simply posting locally written log files to a web server at UCL. This web server was a very simple PHP script that we authored. The server didn't enforce any balancing of conditions, so participants were simply allocated random conditions.

As a first experience of running an experiment in a mostly unsupervised manner, this experiment was a good success technically, but the overall experimental results were modest as only one of the manipulations made a significant impact (having a self avatar led to self-rating of response to a very minor threat in the environment). The main lessons learnt were the difficulty of building an application for unsupervised use, and the need to have checks for outlier data. We discuss these further below.

3.2. VR simulations of AR scenarios

The second experiment we describe is a study on avatar representation and trust. The study uses a protocol that we

[6] https://ryanschultz.com/list-of-social-vr-virtual-worlds/

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

have used in prior work (Pan and Steed, 2016). In the study, a participant is asked challenging trivia questions and can seek assistance from one of two advisors (see Figure 1B). This prior work used a physical robot and a video projected on to a screen to provide different advisor representations. We were interested in recreating this in higher fidelity with an AR headset. Typically we would invite participants in to the lab to run through the study with our own AR devices. As this was no longer possible, and the potential pool of participants with their own AR devices was small, we aimed to replicate the scenario as closely as possible in VR.

The study was developed in Unity for the Oculus Quest 1 and 2. The experiment was submitted on XRDRN and on SideQuest. The incentive for participation was entry into a lottery. The application was designed to function entirely remotely; instructions and consent were baked in to the app itself. The experiment included three different advisors, of which two would be present. One advisor was an expert and would mostly answer questions confidently and correctly; the other was a non-expert and would mostly answer incorrectly. This gave us six conditions. We also controlled for the position of the expert (to the left and right of the user) and which actor was used to represent the expert. In total, we had 24 variants. A server was used to distribute the variant code to the participant's device to balance conditions, and to record the participant's responses while in-app.

After the study, participants completed a short questionnaire in-app on the advisors, and a generic demographics questionnaire outside of the app in the Oculus browser. It was possible to launch this out-of-app questionnaire with a link from the app, so this was straightforward for the participant. The out-of-app demographics questionnaire also included an auto-filled randomly generated user identity number for the participant, and an optional field for an email address, used only for entering the lottery.

Forty-five participants completed the study; our server did not record partial completions. Of those, 28 results were usable. To filter out unusable results, we checked for entries where no advisor or only one advisor was asked for help.

The primary lesson learnt was on experimental protocols for remote experiments. In the original experiment, participants were able to ask neither advisor for help if they wished. In our experiment, we adopted the same strategy, but found a large number of participants never asked either advisor for help. This could be because these participants never realised it was an option; the ability to ask advisors was only described in the instructions page of the app, which participants may have skipped through. Also, some participants could have perceived asking for advice as failing in some way, or as a last resort to be used sparingly.

It is straightforward to filter out responses from participants that selected no advisors. However, many of the participants who responded in this way may have more fully engaged with the experiment given better encouragement. We could have potentially avoided this issue with a change in protocol, by requiring participants to choose an advisor. Alternatively, we could have increased the incentive to seek the correct answers to questions. The original study employed a performance-dependent variable reward scheme, although this is a challenging option to use for remote studies.

3.3. Virtual crowds

The third experiment we describe is a VR study on simulated crowds by Giunchi et al. (2021) (see Figure 1C). The simulations were developed in Unity and run on untethered devices: Oculus Quest 1 and 2. We ran a within-subjects experiment with limited interaction to evaluate the perception of the crowd trajectories that were coming from a synthetic generator or real data. An additional condition included the point of view of the participant within the virtual scenario.

In this study, we did two experiments: the first was completely remote, and it asked to evaluate some videos of crowd simulations, while the second was in a virtual environment with a controlled setup for a group of participants and a remote group of participants. For the remote participants in the second user test, we submitted our experiment on the XRDRN website. We provided the app in the form of a downloadable APK on a web page created specifically for the experiment, accompanied by an information sheet. We also provided internal instructions on how to perform the experiment and a consensus panel for data collection before the beginning of the experiment in VR. Without giving a consensus, the user was placed in an empty room. We did not pay for the participation. The user performed four sessions in which they had to observe a crowd simulated in different conditions: points of view (ground level or top view at 30 meters of elevation) and source of the crowd trajectories (real data or synthetically generated). We randomised the sequence of the sessions to balance ordering effects. At the end of each session, an internal questionnaire was used, and data was saved remotely by using the Firebase server. We reorganised such questionnaire-base tool and extended it to mixed reality into a different product called MR-RIEW toolkit (Bovo et al., 2022b), a toolkit for designing remote immersive experiment workflows.

With this study, we understood the difficulty of designing a self-contained application that functions as an end-to-end unsupervised user test session with questionnaires included. In particular, the part related to the internal questionnaires and the need to place such stages in the middle of trials forced us to study and design a toolkit that could be used in different situations. Such a tool should have provided a stable method of displaying information, made available a simple interaction for selecting the answers, and stored data transparently and in an anonymous form.

3.4. Seated vs. standing locomotion

The fourth experiment we describe is a VR study on a novel interaction technique meant to combine the advantages of both sitting and standing posture (Ganapathi et al., 2022). A majority of VR applications are designed with a standing user in mind, but a seated posture might be preferred by some users. The reasons for this vary from injuries and disabilities to simply wishing for less physical strain and more comfort. However, the seated posture has the disadvantage of a lower eye height in the virtual world which results in an inferior overview of the environment and difficulties interacting with objects (see Figure 1D). To combine the advantages of both postures we proposed the floating technique which would alter the user's virtual sitting eye height to match the eye height of a standing user. We performed a user study to investigate the efficiency of our floating technique comparing it with sitting and standing postures.

The experiment was conducted remotely at first with a follow-up in our lab. For the remote study, participants ($N = 18$) that already owned an Oculus Quest device were recruited through online communities such as Reddit and Facebook. The app and instructions were provided online and the installation was done *via* SideQuest. Consent for data collection was obtained through an online form. At the start of the application, a unique key for each participant was generated before the training session. This key was used as an identifier to connect the results from the online questionnaires with the data collected by the app while retaining the anonymity of the participants. Instructions were presented in the virtual environment, including instructions for moving, rotating and grabbing target items. After the completion of a training session, the participants performed the same task, searching for items in a supermarket, in a standing, sitting, or floating posture. The three techniques were presented in a randomised order that was locally generated. After each condition, the participants were asked to fill in online questionnaires on motion sickness and presence. The three conditions were concluded with a post-experiment questionnaire and a free-text field for qualitative feedback. During the experiment, tracking data was recorded and on completion submitted to a server at UCL. Tracking data was recording from head and hands at 2 Hz. The uploading of this data was not a concern in this study. In total, the length of the experiment was 40 to 50 min including questionnaires. Out of 40 data sets collected remotely, only 18 were found to be usable as many had either incomplete questionnaires or inconsistencies in the telemetry data such as described in Section 5. Due to the low compliance, a follow up lab-based study ($N = 18$) was conducted under the supervision of an experimenter. When comparing the results of the follow-up with the ones from the remote study, no significant differences were found, but the compliance was better. Compared to the 22 rejected data sets of the remote study, only 13 were rejected in the local follow-up. We observed that the floating technique had no detrimental effect in comparison to the standing technique and had a slight benefit over the sitting technique.

The key lesson learnt from this experiment was the importance of comprehensive logging and clear and simple instructions when performing an experiment unsupervised in an unknown environment. While it can not fully replace an observing researcher, comprehensive telemetry data can help find cases of non-compliance while better instructions can reduce the number of cases where participants deviate from the protocol unintentionally.

3.5. Other studies

We ran at least seven other studies at least partly out of the laboratory during the past two years. These range enormously in complexity and style. They include very short studies (5 min) through to long studies (60+ min, e.g., Bovo et al., 2022a). They involve testing of story-based scenarios through piloting of larger studies (Thiel and Steed, 2021) to large-scale testing of new interaction techniques. Most studies used Oculus Quest 1 or Oculus Quest 2. One study required Oculus Quest 2 and also used an EVU TPS Wearable to measure galvanic skin response, temperature and heart-rate. For this study the HMD and wearable device were delivered to remote participants. One study required a tethered HMD, and thus used the participants used an Oculus Rift (two users) and Oculus Quest *via* Link (two users). One study was an update and re-run (Steed et al., 2022) of the study mentioned in Section 3.1. The number of participants varied from 6 (4 complete data sets) through to 37 (only 15 complete datasets). Other studies achieved 100% completion rates with, for example, 20 participants. Some studies ran additional participants in the laboratory once it was re-opened. The designs of the environments varied from abstract visualisations through to depictions of real places. One study used a full avatar, but the most common choice was hands-only representations, with two studies choosing the standard Oculus Avatars. Other than those cited, one study is forthcoming at the time of writing and the others were un-published student projects.

4. Challenges and opportunities

4.1. Challenges: Experiment protocols

The first challenge is that the protocols have to be self-running, self-documenting and robust to user behaviour. In the lab, the entire process, from the moment a participant arrives until they leave at the end of the session, is highly regimented, e.g., we offer them water, discuss the experiment, answer any questions they have about participant information and get their consent. This control and uniformity of participant experience allows for strong internal validity—for any given experimental condition, each participant has the same experience, albeit moderated by their own actions and the (intended) system responses to these actions. With the "out of the lab" model, some of this control is inevitably lost. Participants can still ask questions (e.g., *via* email) but the added overhead makes this less likely. The ability to directly observe participants during a pre-trial tutorial phase and correct noncompliance (e.g., due to misunderstanding of instructions) is absent, and so experiment designs need to be tailored to reduce the scope for such errors of misunderstanding. For lab-based studies, each trial takes place in the same uncluttered, reasonably spacious and quiet space, whereas we have no control over the space in which remote experiments are conducted. In the event of deviation from the protocol, due to hardware or software malfunction, a lab-based experiment can be briefly paused while corrections or adjustments are made. However, in uncontrolled environments obstacles that interrupt the flow of the experiment stand a much greater chance of the participant ending the experiment prematurely. This is especially true of capture systems running concurrently to the immersive experience, such as bio data-gathering apps *via* Bluetooth wearables. Often the signal can be interrupted or drop entirely, resulting in data loss that a user would be unaware of until they complete the experiment.

4.2. Challenge: Experiment scope

Another challenge in running studies remotely is the difficulty of supporting specific custom interface or capture technologies. For example, some of our experiments require one-off systems (e.g., Steed et al., 2020a which uses a custom haptic system) or use combinations of monitoring equipment (e.g., Yuan and Steed, 2010 is one of several studies that use galvanic skin response sensors (GSR)). Some things can be lent to users, but this means that we need to at least engage with the potential participants through a delivery service, thus reducing the potential to get more numerous or more diverse participants, see below. While we have lent out additional monitoring devices on their own or with consumer VR equipment, the main implication of this restriction is that experiments are mostly constrained to work within the limitations of the existing interface hardware.

4.3. Challenge: Ethics

Dealing with ethical issues that are raised by distributed experiments has been challenging for the community. Our laboratory has a lot of experience in running experiments and had operated under a "blanket" authorisation to run certain classes of experiments with certain measurements for many years. Following an early experiment on running a study out of the lab in 2016 (Steed et al., 2016), we had already received blanket ethics approval for remote studies though this was much narrower in scope than our blanket ethics approval for in-lab studies. This approval covered secure data capture, how to frame the experience within a stand-alone application (see below), etc. When COVID-19 hit, this second blanket approval was sufficient for many, but not all, of the studies we now wanted to run outside the lab. This was particularly salient in experiments that required the capture of bio data *via* a sensor. Ordinarily in a lab setting, the complex capture process would run locally *via* dedicated hardware. But in distributed experiments, mobile solutions with an easier initialisation process are required, since the participant is responsible for their setup. These solutions tend to rely on third party servers (often based in the US) to parse, process and return the data, which can further complicate the process due to GDPR regulations.

4.4. Challenge: Publishing experiments and recruitment

There are three ways that we have run or are planning to run studies out of the laboratory. The first is just publishing an application online. In this case, it must run on consumer equipment (e.g., Steed et al., 2016). The second is a small variant, in that we publish an application but have other experimenters run it because the equipment is largely constrained to research labs (e.g., on relatively uncommon systems such as HoloLens, see Steed et al., 2020b). The third is sending out specific sets of equipment to users, including, for example, a consumer head-mounted display (HMD) with some other tracking equipment (e.g., Moustafa and Steed, 2018).

4.5. Opportunity: External validity

Distributed experiments provide an opportunity to reach a different participant pool which is potentially more diverse. In the lab, although we use various participant pools and recruitment services, most participants are students or friends of students. They have specific motivations: they want to try VR, their friend tried it, they have a spare hour between classes, or they get paid. Out of the lab experiments have the potential to reach a much wider participant demographic, potentially enhancing the ecological validity of results obtained. However, for distributed participants we should still acknowledge some potential biases. Participants who own VR hardware could be biased towards being early adopters and certainly have some experience with virtual reality and games/social experiences. This population might skew in age or gender. However, for certain types of experiment we desire participants who are already VR (or AR) users. We no longer plan any experiment where we expect all participants to be naive and thus it is more important to measure prior experience in VR. Some of the studies we discuss started to include a simple question such as "How many times have you experienced VR before?" to be answered on a scale of "None, 1–3 Times, 4–6 Times, 6–10 Times, 10+ Times". While this allows us to distinguish in-experienced from more experienced users, on its own, it doesn't probe the range the experience types the person has had, or even what they consider to be "VR". This is certainly an area where a better, more standard questionnaire would be useful.

4.6. Opportunity: Reproduction and openness

A second opportunity, and one that motivated some of our early efforts in this area, is that sharing experiments can enable people to understand our experiments and reproduce them, because they can experience them. Our Singer in the Bar Experiment was available on a number of platforms for a few years, but those platforms (Gear VR, Cardboard) are now deprecated (Steed et al., 2016) (an updated version is now available, see Steed et al., 2022). A slightly different angle on this was not a distributed experiment, but used commercial content, in that the experiment was based on the experience "We Wait" produced by the BBC which is still available (Steed et al., 2018). Aside from making experiments available, we hope to start sharing full experiment code. This is partly enabled by Microsoft open sourcing the RocketBox avatars (Gonzalez-Franco et al., 2020b) as many of our demonstrations use those assets.

4.7. Opportunity: Scale and flexibility

There is also a larger opportunity: recruitment of very large numbers of participants allowing different types of experiment with more conditions, or more open-ended participant engagement. To reach this, we obviously need to make the experiment attractive to run on its own. It is unlikely that large scale experiments would be compensated.

We would highlight another advantage to participants: they can run experiments in their own time and thus engagement might be higher as they can schedule themselves. This potentially reduces the risk of no-shows, which has been a particular problem for us when running multi- user experiments.

5. Technical lessons

In this section we cover, in no particular order, some technical insights from our experience of running experiments out of the lab.

We have adopted two strategies to ensure participants are properly instructed and give informed consent: instructions are fully on the web as part of the download experience and/or instructions are given from within the app. The first is relatively straightforward, but means that we need to control distribution of the application so that participants have read the instructions. Downloading acts as confirming consent. The second is more suitable to app stores. Our ethics approval allows for short-form instruction and consent in-app as long as participants can also access a long-form version online if they wish.

Our apps are designed to be easy to use. They must include their own tutorials. We use standard interaction techniques such as teleportation and grabbing, and do not overload the user with instructions. User interfaces, especially if they involve questionnaires or other data entry, need extensive testing. For example, it is useful if a questionnaire system includes a back-button to cope with situations with inadvertent clicks. Alternatively, if there is need for keyboard entry, then this needs testing on different HMDs. One recent activity of the group has been to try to standardise some simple user interface and data entry components as part of our Ubiq-Exp toolkit (Steed et al., 2022).

We direct participants to debriefing and often provide a simple summary of what happened within the applications at the end. A form of real-time feedback on user interaction during this tutorial phase is recommended so that participants can be guided back to the protocol, should they deviate from it. Simply denying them the ability to continue with an experiment can prove jarring and result in premature termination, while allowing them to continue risks compromising the experimental data collection.

Questionnaires can be administered either in-app after each trial, or we can instruct the user to remove the headset and complete the questionnaire online, before putting the headset back on for the next stage of the experiment. We have found there to be pros and cons for both methods. Making the user remove the headset to complete questionnaires at intermediate stages complicates the protocol and risks non-compliance (e.g., any inattentive users may click through to the next stage without doing the questionnaire). However, this method may create better engagement with tasks in VR since fatigue due to prolonged wearing of the headset is less likely. Doing the questionnaires in-app reduces risk of non-compliance, since the user is led through the process linearly.

Distribution of applications to participants is not straightforward. On PC participants rely on our word that the application is not dangerous. We have made extensive use of the SideQuest platform to deliver applications because although it is not simple for end-users, the instructions are clear and they might appreciate having access to other demonstrations on that platform. While SideQuest functions similarly to an app store, participants are still required to "sideload" applications. As a result, users are required to setup their device as a developer and have Oculus developer accounts. This introduces biases to the participant pool and we might expect that participants gathered this way might tend to be more experienced.

A problem we have dealt with is compliance with the protocol. There are two main ways of dealing with this. In our Singer in the Bar experiment (Steed et al., 2016), we put in various measures to monitor participants to make sure that they were active and looking in plausible directions. We also filtered the answers that were given on Likert scales to remove participants who answered too quickly or who answered the same value for each question. We also balanced some questions, such that some opposing ratings might be expected.

In a study investigating virtual posture manipulation for seated users (Ganapathi et al., 2022), non-compliance took on two main forms—height discrepancies and timing discrepancies. For different stages of the experiment, participants were required to perform tasks either in a seated or standing posture. Non-compliance with these instructions could be identified by examining headset height data for these trials and data showing height discrepancies were eliminated. Timing discrepancies occurred when participants stopped and took off their headset mid-experiment. Reasons for this are unclear, but we speculate that in an unsupervised domestic setting, they may have stopped to answer phone calls or doorbells. This type of non-compliance was identified by discrepancies between the experiment time recorded (Unity time, which is paused when the application is interrupted) and the real-world time (Unix time) that had elapsed. In subsequent experiments we have advised participants that they should attempt to complete the experiment in one sitting.

Many of the experiments have been designed for seated participation. The reason for this is that we don't know the amount of space that is available to the user. While we could monitor the chaperone or guardian system, there is no guarantee that the participant has this correctly configured.

We extensively test our applications and generally avoid anything that could cause simulator sickness. Thus, while not all travel techniques cause simulator sickness, we have tended to avoid enabling travel techniques unless necessary. This has included making the virtual space smaller so that the participant does not have to move.

To satisfy data protection requirements we tend to log data to a secure server at UCL during the experience. No data is left on the device. However, we have adopted a process of keeping a simple count of the number of times the experiment has been run on the device. We can't prevent a participant running the experiment multiple times, but we can ensure that we iterate through different conditions. Another workaround, especially when the use of additional peripheral devices is involved (such as for gathering biometric data), is to send a smartphone that pairs to the said device and stores the data *via* a custom mobile application. As mentioned earlier, the disadvantage in that case is that participants are unable to monitor the smartphone app while inside VR in the event of it crashing or the connection dropping during the experiment.

If participants are going to be paid, there needs to be a negotiation between a server and the application software. The easiest way we have found is that the application generates a unique code on completion that can be redeemed if emailed to us. This unique code can be matched to the data we received. Gift card codes could be provided in an automated way. However, while time consuming, the intervention of researchers helps to prevent abuse. The same approach can be used for both direct reward and lottery rewards. The motivation for lottery rewards is that we didn't always know exactly how many participants the study would attract. It also makes sure that the distribution of compensation is a fixed cost to the experimenter, in that there are only a fixed number of awards to distribute. We also considered it to be appropriate if the study was relatively short in duration (e.g., <10 min) as otherwise the participant would be compensated a relatively small amount. While this was attractive to some participants we did receive a small number of negative comments stating that the potential participants (only potential because we don't know if they completed the study or not) would rather do the study for a fixed amount.

Ensuring that each participant is a unique user and represents a new source of data is a significant challenge. Users may share a device, which is a legitimate use case for running the study multiple times. Uniquely identifying a device is in any case difficult. It is possible to log IP and MAC addresses, but this does not constitute a foolproof method as both can be changed. We keep rewards small to reduce the incentive for abuse of this kind. We also distribute rewards over email and can identify duplicate email addresses, which is an additional small impediment to fraud.

Because we require participants to be online to log data, for some experiments we have adopted a strategy of having a server distribute any necessary condition configuration so as to balance the number of participants in conditions. On connection, the participant is provided a condition code matching the condition with the least complete participants.

The application is designed to understand and make use of the code. At this stage, the server temporarily increments its internal count of completed experiments. Should the participant not complete the experiment within a 24 hour interval, the completed count for the condition is decremented. This helps to avoid conditions being over-served when multiple participants connect at once.

6. Conclusions

Going forward, we hope new platforms emerge that facilitate the running of experiments, to enable more studies to be run by a broader range of researchers. To this end we are making some of our own software available, starting with the Ubiq platform (Friston et al., 2021). One of the main areas that we believe needs development is the support of diverse and flexible systems for avatars for experiments. The RocketBox/MoveBox effort is a key step in this direction (Gonzalez-Franco et al., 2020a,b), which we have supported in Ubiq (Izzouzi and Steed, 2022). Much more can be done to make tools available for customising such avatars to give more variety and perhaps make them resemble participants. Additionally there is a need for easy-to-use system for animating behaviours for full-body avatars. Another area is robust logging and data analysis for these systems. In other work, we have demonstrated methods for analysing distributed systems (Friston et al., 2018). It would be useful to integrate such tools into real-time tools for analysing transient issues in systems to understand how to engineer them more effectively. Finally, aside from open software, we hope that more platform services emerge that facilitate some of the tasks in experiment operation, such as solicitation of participants, screening questionnaires, software distribution and secure logging.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by University College London Research Ethics Committee. The studies were approved the requirement of written informed consent for participation.

Author contributions

AS coordinated the paper and led the write-up. The related studies were conducted with UCL as the project lead. All authors contributed their experiences.

Funding

This work was partly funded by UK EPSRC project CASMS (Grant Reference EP/P004016/1), UK EPSRC project Graphics Pipelines for Next Generation Mixed Reality Systems (Grant Reference EP/T01346X/1), EU H2020 project RISE (Grant Number 739578), EU H2020 project PRIME-VR2 (Grant Number 856998), and EU H2020 project CLIPE (Grant Number 860768).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bierbaum, A., Just, C., Hartling, P., Meinert, K., Baker, A., and Cruz-Neira, C. (2001). "VR Juggler: a virtual platform for virtual reality application development," in *Proceedings IEEE Virtual Reality 2001* (Yokohama), 89–96. doi: 10.1109/VR.2001.913774

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: review and suggested criteria. *Presence Teleoperat. Virtual Environ.* 12, 456–480. doi: 10.1162/105474603322761270

Bovo, R., Giunchi, D., Costanza, E., Steed, A., and Heinis, T. (2022a). "Shall I describe it or shall I move closer? Verbal references and locomotion in VR collaborative search tasks," in *ECSCW 2022* (Coimbra).

Bovo, R., Giunchi, D., Steed, A., and Heinis, T. (2022b). "MR-RIEW: an MR toolkit for designing remote immersive experiment workflows," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (Christchurch: IEEE), 766–767. doi: 10.1109/VRW55335.2022.00234

Estellés-Arolas, E., and González-Ladrón-De-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *J. Inform. Sci.* 38, 189–200. doi: 10.1177/0165551512437638

Friedman, D., Steed, A., and Slater, M. (2007). "Spatial social behavior in second life," in *Intelligent Virtual Agents*, eds C. Pelachaud, J. C. Martin, E. Andre, G. Chollet, K. Karpouzis, and D. Pele (Berlin; Heidelberg: Springer), 252–263. doi: 10.1007/978-3-540-74997-4_23

Friston, S., Griffith, E., Swapp, D., Marshall, A., and Steed, A. (2018). "Profiling distributed virtual environments by tracing causality," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Tübingen; Reutlingen), 238–245. doi: 10.1109/VR.2018.8446135

Friston, S. J., Congdon, B. J., Swapp, D., Izzouzi, L., Brandstätter, K., Archer, D., et al. (2021). "Ubiq: a system to build flexible social virtual reality experiences," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology, VRST '21* (New York, NY: Association for Computing Machinery). doi: 10.1145/3489849.3489871

Ganapathi, P., Thiel, F. J., Swapp, D., and Steed, A. (2022). "Head in the clouds - floating locomotion in virtual reality," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (Christchurch: IEEE), 668–669. doi: 10.1109/VRW55335.2022.00185

Giunchi, D., Bovo, R., Charalambous, P., Liarokapis, F., Shipman, A., James, S., et al. (2021). "Perceived realism of pedestrian crowds trajectories in VR," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (Osaka), 1–5. doi: 10.1145/3489849.3489860

Gonzalez-Franco, M., Egan, Z., Peachey, M., Antley, A., Randhavane, T., Panda, P., et al. (2020a). "MoveBox: democratizing MoCap for the microsoft rocketbox avatar library," in *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)* (Utrecht: IEEE), 91–98. doi: 10.1109/AIVR50618.2020.00026

Gonzalez-Franco, M., Ofek, E., Pan, Y., Antley, A., Steed, A., Spanlang, B., et al. (2020b). The rocketbox library and the utility of freely available rigged avatars. *Front. Virtual Real.* 1, 561558. doi: 10.3389/frvir.2020.561558

Izzouzi, L., and Steed, A. (2022). "Integrating rocketbox avatars with the ubiq social VR platform," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (Christchurch), 69–70. doi: 10.1109/VRW55335.2022.00025

Kelly, J. W., Cherep, L. A., Lim, A. F., Doty, T., and Gilbert, S. B. (2021). "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)* (Lisboa), 687–694. doi: 10.1109/VR50410.2021.00095

Kittur, A., Chi, E. H., and Suh, B. (2008). "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08* (New York, NY: Association for Computing Machinery), 453–456. doi: 10.1145/1357054.1357127

Mottelson, A., and Hornbæk, K. (2017). "Virtual reality studies outside the laboratory," in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology, VRST '17* (New York, NY: Association for Computing Machinery), 1–10. doi: 10.1145/3139131.3139141

Mottelson, A., Petersen, G. B., Lilija, K., and Makransky, G. (2021). Conducting unsupervised virtual reality user studies online. *Front. Virtual Real.* 2, 681482. doi: 10.3389/frvir.2021.681482

Moustafa, F., and Steed, A. (2018). "A longitudinal study of small group interaction in social virtual reality," in *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology, VRST '18* (New York, NY: Association for Computing Machinery). doi: 10.1145/3281505.3281527

Oh, C. S., Bailenson, J. N., and Welch, G. F. (2018). A systematic review of social presence: definition, antecedents, and implications. *Front. Robot. AI* 5, 114. doi: 10.3389/frobt.2018.00114

Pan, Y., and Steed, A. (2016). A comparison of avatar-, video-, and robot-mediated interaction on users' trust in expertise. *Front. Robot. AI* 3, 12. doi: 10.3389/frobt.2016.00012

Radiah, R., Makela, V., Prange, S., Rodriguez, S. D., Piening, R., Zhou, Y., et al. (2021). Remote VR studies: a framework for running virtual reality studies remotely via participant-owned HMDs. *ACM Trans. Comput. Hum. Interact.* 28, 46:1–46:36. doi: 10.1145/3472617

Ratcliffe, J., Soave, F., Bryan-Kinns, N., Tokarchuk, L., and Farkhatdinov, I. (2021). "Extended reality (XR) remote research: a survey of drawbacks and opportunities," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–13.

Saffo, D., Di Bartolomeo, S., Yildirim, C., and Dunne, C. (2021). "Remote and collaborative virtual reality experiments via social VR platforms," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–15. doi: 10.1145/3411764.3445426

Silvertown, J. (2009). A new dawn for citizen science. *Trends Ecol. Evol.* 24, 467–471. doi: 10.1016/j.tree.2009.03.017

Singhal, S., and Zyda, M. (1999). *Networked Virtual Environments: Design and Implementation*. Reading, MA: Addison Wesley.

Steed, A., Archer, D., Congdon, B., Friston, S., Swapp, D., and Thiel, F. J. (2021). Some lessons learned running virtual reality experiments out of the laboratory. *arXiv preprint arXiv:2104.05359*. doi: 10.48550/arXiv.2104.05359

Steed, A., Friston, S., Lopez, M. M., Drummond, J., Pan, Y., and Swapp, D. (2016). An "in the wild" experiment on presence and embodiment using consumer virtual reality equipment. *IEEE Trans. Visual. Comput. Graph.* 22, 1406–1414. doi: 10.1109/TVCG.2016.2518135

Steed, A., Friston, S., Pawar, V., and Swapp, D. (2020a). "Docking haptics: extending the reach of haptics by dynamic combinations of grounded and worn devices," in *26th ACM Symposium on Virtual Reality Software and Technology, VRST '20* (New York, NY: Association for Computing Machinery), 1–11. doi: 10.1145/3385956.3418943

Steed, A., Izzouzi, L., Brandstatter, K., Friston, S., Congdon, B., Olkkonen, O., et al. (2022). Ubiq-exp: a toolkit to build and run remote and distributed mixed reality experiments. *Front. Virtual Real.* 3, 912078. doi: 10.3389/frvir.2022.912078

Steed, A., and Oliveira, M. F. (2009). *Networked Graphics: Building Networked Games and Virtual Environments*. Burlington, MA: Elsevier.

Steed, A., Ortega, F. R., Williams, A. S., Kruijff, E., Stuerzlinger, W., Batmaz, A. U., et al. (2020b). Evaluating immersive experiences during COVID-19 and beyond. *Interactions* 27, 62–67. doi: 10.1145/3406098

Steed, A., Pan, Y., Watson, Z., and Slater, M. (2018). "We Wait"—the impact of character responsiveness and self embodiment on presence and interest in an immersive news experience. *Front. Robot. AI* 5, 112. doi: 10.3389/frobt.2018.00112

Thiel, F. J., and Steed, A. (2021). "Lend me a hand"—Extending the reach of seated VR players in unmodified games through remote co-piloting," in *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (Lisbon), 214–219. doi: 10.1109/VRW52623.2021.00047

Williamson, J., Li, J., Vinayagamoorthy, V., Shamma, D. A., and Cesar, P. (2021). "Proxemics and social interactions in an instrumented virtual reality workshop," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13 (New York, NY: Association for Computing Machinery). doi: 10.1145/3411764.3445729

Yuan, Y., and Steed, A. (2010). "Is the rubber hand illusion induced by immersive virtual reality? in *Virtual Reality Conference (VR)* (Boston, MA: IEEE), 95–102. doi: 10.1109/VR.2010.5444807

Zhao, J., Simpson, M., Sajjadi, P., Wallgrën, J. O., Li, P., Bagher, M. M., et al. (2021). "CrowdXR - pitfalls and potentials of experiments with remote participants," in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (Bari). doi: 10.1109/ISMAR52148.2021.00062