An 'In the Wild' Experiment on Presence and Embodiment using Consumer Virtual Reality Equipment

Anthony Steed* University College London, UK
Sebastian Friston† University College London, UK
María Murcia López‡ University College London, UK
Jason Drummond§ University College London, UK
Ye Pan¶ University College London, UK
David Swapp‖ University College London, UK

Abstract—Consumer virtual reality systems are now becoming widely available. We report on a study on presence and embodiment within virtual reality that was conducted 'in the wild', in that data was collected from devices owned by consumers in uncontrolled settings, not in a traditional laboratory setting. Users of Samsung Gear VR and Google Cardboard devices were invited by web pages and email invitation to download and run an app that presented a scenario where the participant would sit in a bar watching a singer. Each participant saw one of eight variations of the scenario: with or without a self-avatar; singer inviting the participant to tap along or not; singer looking at the participant or not. Despite the uncontrolled situation of the experiment, results from an in-app questionnaire showed tentative evidence that a self-avatar had a positive effect on self-report of presence and embodiment, and that the singer inviting the participant to tap along had a negative effect on self-report of embodiment. We discuss the limitations of the study and the platforms, and the potential for future open virtual reality experiments.

Index Terms—virtual reality, consumer equipment, head-mounted display, presence, embodiment, self-avatar

1 Introduction

Consumer virtual reality is becoming more popular, creating both a risk and an opportunity. The risk is that a lot of content is being built that is not following best practices and that consumers' initial experiences of different virtual reality systems will be disorientating and cumbersome because the interfaces for different applications are quite different. One can look at the variety of different types of embodiment, locomotion interfaces and interaction metaphors being used in current titles for the Oculus Rift DK2 as indicative of the range of experiences that will be available at the launch of consumer devices. For example, many applications use a self-avatar, especially if the application is seated, but many others do not. The opportunity is that we can mine data from consumers' experiences of such applications to better inform design going forward.

In this paper we propose to collect performance data and user responses to questionnaires to guide decisions about design features. Iterative design is, of course, a part of the development of any application, but we would seek to establish new guidelines or re-affirm the applicability of existing knowledge or guidelines in consumer applications. Guidelines could then be applied to a broad range of applications with more confidence.

To this end we built a simple application for two of the current consumer virtual reality platforms: Samsung Gear VR and Google Cardboard. These were chosen because of their broad accessibility to consumers rather than developers. The application was a simple seated experience using only head gaze as input to activate questionnaires. Participants would download the application from the appropriate store or link. They could read guidance about the experiment, including statements about ethics and data collection, both online and within the application itself. They could then indicate their permission that data be collected. A short pre-questionnaire was followed by a main experience that involved watching a singer perform in a bar and then a longer post-questionnaire. Data was collected on a server and feedback about experimental conditions given to the user.

This type of data collection 'in the wild' is common in mobile human-computer interaction (HCI) studies and other areas that have engaged citizens in scientific data collection (see Section 2.2). The main issues are the uncontrolled nature of data collection and the potential for non-compliance with protocol. Another issue becomes more relevant with an immersive system: the safety and ethical issues of conducting an uncontrolled study. Thus while data collection is common in commercial applications and systems, due consideration was paid to the issue for this application. In particular the content was not shocking and the choice was made to create a stationary seated experience to avoid issues of motion-induced instability or sickness. The study was approved by University College London (UCL) Research Ethics Committee.

The goals of the study were thus two-fold: to test the feasibility and utility of running a study on virtual reality in the wild; and to test three specific hypotheses about presence and embodiment in immersive virtual reality. The first hypothesis was that having a self-avatar would have a positive impact on self-report of presence. The second hypothesis was that having the singer appear to engage in eye contact with the participant would have a positive impact on self-report of presence. The third hypothesis was related to embodiment and self-representation. At the start of the singer's performance she would invite participants to tap along to the music and the self-avatar (if one existed) would also tap along. The hypothesis was that this would have a positive impact on self-report of embodiment as tested by questions related to body ownership illusions (see Section 2.1). The results showed some support for the first hypothesis, no support for the second hypothesis and support for the negation of the third hypothesis.

The remainder of this manuscript is organized as follows. In Section 2 we briefly review related work on embodiment and presence. Section 3 describes the system implementation, scenario and methodology. Section 4 reports on data collection, filtering and results. Section 5 discusses the results and the potential for such in the wild studies. Section 6 concludes.

*e-mail:a.steed@ucl.ac.uk
†e-mail:sebastian.friston.12@ucl.ac.uk
‡e-mail:maria.murcia.13@ucl.ac.uk
§e-mail:J.Drummond@cs.ucl.ac.uk
¶e-mail:y.pan@cs.ucl.ac.uk
‖e-mail:D.Swapp@cs.ucl.ac.uk

Manuscript received 21 Sept. 2015; accepted 10 Jan. 2016. Date of publication 20 Jan. 2016; date of current version 19 Mar. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2016.2518135

2 Related Work

2.1 Presence and Self Representation

Presence has been a topic of research for many years in virtual reality. The term is loosely defined as the 'sense of being there' where 'there' is the virtual environment created by the virtual reality system (e.g. [7]). A detailed overview of the area is beyond the scope of this paper and we describe a subset of the work that concerns measurement of presence and the impact of the virtual body. A good overview can be found in Scheumie et al. [20].

We will follow the definitions of Slater & Wilbur [26] in separating *immersion* as description of the system and *presence* as a user's response to the virtual environment. The modern smartphone-based head-mounted displays (HMDs) that we are supporting in this experiment are highly immersive in that they are quite high fidelity, resolution and field of view. However they do not provide interaction through the hands and thus the self-avatar cannot match the real body. This might hinder the presence response in highly interactive scenarios that might usefully use kinesthetic cues and other sensorimotor contingencies [22]. However for some situations that rely mostly on passive observations of action, we might expect appropriate levels of presence to occur and be reported or otherwise detected.

The most common method to assess presence is to use some form of subjective questionnaire. Various questionnaires have been proposed such as the Slater, Usoh, Steed questionnaire [25], Witmer and Singer questionnaire [31] and ITC-Sense of Presence Inventory [13]. While these questionnaires have proved effective, they have their limits (e.g. see [28]), and one main constraint in our experiment is that we had to keep the number of questions as low as possible to ensure good numbers of responses from participants. Thus as we will discuss later, we adapted some of the questions from [25].

Many other methods of assessing presence exist [11]. Some behavioural cues might be accessible to an experiment done with consumers. Certainly high-level behaviours in interactive environments such as navigation routes through environments would be easily accessible. However techniques that rely on a stressful responses (e.g. [27]) are arguably not appropriate due to the ethics of collecting data in public studies, see below. Others that require verbal responses might be available if audio was recorded from the device; most devices being used in this study could accomplish this (e.g. [24] uses verbal reports of break in presence, [11] suggests a continuous verbal presence report). This is a promising avenue for future studies.

The impact of the virtual body or self-avatar has been extensively investigated. A recent paper by Biocca reviews the potential benefits and drawbacks of using a self-avatar [1]. A very common demonstration of virtual reality involves a 'virtual pit' style environment [27]. Lin et al. have shown that a self-avatar that is gender and height matched made participants less likely to step off a ledge [14] .

Several authors have looked at the impact of a self-avatar on task performance and interaction. The general thrust of the work indicates that self-avatars are important (e.g. [18]), and that animation of the avatar can improve the effect of the self-avatar [15]. These results are very important for the study we present, in that we provide a self-avatar, but it cannot be animated to follow the participant's movements because we have a single point of orientation-only tracking on the head.

Aside from the work on impact of self-avatar, several authors have looked at the mechanisms by which the self-avatar embodies the user, so that the user can experience the self-avatar as their own body [21]. The basis of much of this work is rubber hand illusion demonstration from Botvinick and Cohen [3], where a participant is induced into the illusion that a rubber hand is part of their body. To set up the illusion, the participant's real hand is obscured and a rubber hand placed roughly in the visual line between the eye and obscured hand. The induction involves the participant's real hand being touched at the same time as they see the rubber hand being touched. After a few seconds the participant has a strong sense that the rubber hand is their own, to the extent that if the rubber hand is threatened, the participant might react by withdrawing their real hand. Variations of this demonstration have been made in virtual reality [19]. In particular, in [33] it was shown that a participant can 'self-induce' a similar illusion by taking part in an interactive virtual reality session where the participant extensively engages in interaction tasks with their tracked hands. We will attempt to induce a similar illusion, but by having the participant tap along to an song, not by engaging in an interactive task.

2.2 Data Collection in the Wild

The notion of collecting data 'in the wild' has a long history. Various organisations have engaged the public in data gathering for decades (e.g. bird or butterfly surveys, meteorological data gathering, etc.). This has been given impetus by the ubiquitous use of personal and mobile computing. A famous example is the Fold@Home software, where games players solved a protein folding problem [12]. The term 'citizen science' is sometimes used to refer to such public studies. Such studies serve a dual purpose: both collection of data for scientific purposes and raising the profile of the science programme in a field [2].

The term 'in the wild' within HCI also has the connotation of performing studies with real users in uncontrolled environments rather than in the laboratory [4]. There are obvious challenges to conducting studies out of the lab, such as the reliability of data gathering and ensuring the control of conditions. However it is argued that the amount of data that can be gathered is larger and thus reliability can be achieved in different ways [5, 6].

Specifically for mobile devices, the prevalance of app stores has meant that researchers can distribute applications easily to a vast population, removing many of the worries about installation and reliability of apps on diverse platforms [8]. Thus many studies in mobile computing have been able to draw on extremely large populations of users (e.g. [10, 9]). While the install base of consumers with virtual reality devices is relatively small at the moment, we can hope to achieve such scale in the future.

A key issue in accessing data from consumer applications is the ethics of data collection. McMillan et al. review the general ethical guidance for experimental studies and apply it to large-scale mobile HCI experiments [16]. That group's later work suggests that a key requirement is to be very clear about what data is collected and thus to visualise the data being collected [17]. While in this study we have followed such guidelines and interpreted them as necessary for the immersive situations (see Section 3.5), we only notify the participants through text posters and audio explanation about the data collected. The suggestion to show the participants the data collected seems eminently applicable to virtual reality studies that, for example, record positional tracking data. This is something we plan to pursue in future studies.

3 System and Method

In this section we discuss the system design and implementation of the experiment application. The application was coded in the Unity system, and the main devices supported were the Samsung Gear VR and Google Cardboard devices on iOS and Android.

Note that throughout this section and later in the paper, we will use the term "user" to refer to anyone using the application, whereas we will use the term "participant" to refer to those users who agreed to submit data. Users need not agree to data collection, in which they are allowed to go through the main experience without completing unnecessary questionnaires and without any data collection taking place.

Note that the application is freely available at http://vr.cs.ucl.ac.uk/vrjam. A side effect of doing the experiment in the wild is that the full application and protocol are available to experience.

3.1 Scenario Design

This initial experiment was designed as a mostly passive experience. Because of the constraints of the target platforms, only head orientation was assumed and no button input. Although the Samsung Gear VR has a small touch pad and one button, most Google Cardboard devices have only a single button and some similar devices do not

Fig. 1. Virtual bar, singer and second spectator.

have any buttons. For this reason, and issues of safety and ethics, attempting a highly interactive experience was ruled out for this initial experiment.

The main experience decided upon was to have the user sit in a bar watching a singer perform. A simple bar scenario was based on the model "French Pub Interior" by Enozone that is available on the Unity Asset Store. The size and complexity of this model had to be reduced in order to be able to meet frame rate goals on the target devices. Some objects were replaced with simplified versions using textures to replace geometric features. Others were removed. The static lighting was recalculated. We used a small number of light probes for real-time lighting.

The user was seated facing a stage, see Figure 1. In front of the user was a table. On the other side of the table another male avatar was also watching the singer. On the table was a small box (see Figure 2 Top).

After a brief initial silent period (13s) where the singer appears to be waiting, there is a short instrumental intro to the song and then the singer starts. The song lasts approximately 150s, after which there is some applause from the other male avatar. This scene lasts 170s in total.

During the song the male avatar continually faces towards the singer. He does occasionally shuffle his chair and this knocks the table. On the third shuffle the box slides off the table, hits the knee of the user's avatar (or the space where the avatar's knee would be in conditions without an avatar) and then rolls under a chair.

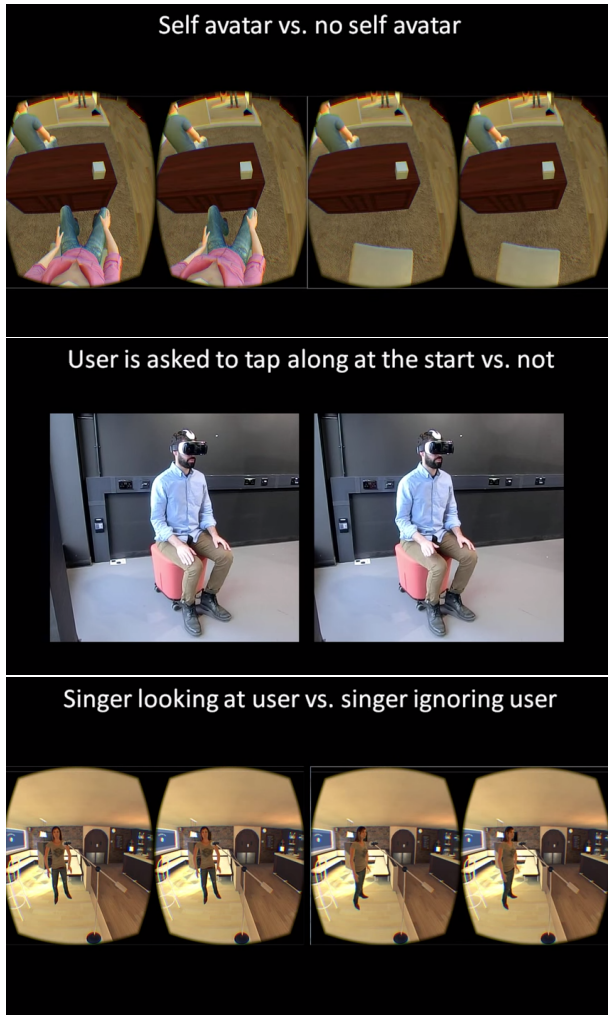

Fig. 2. The experimental conditions. Top: self-avatar versus no self-avatar. Middle: induction versus no induction. Bottom: singer looking at user versus singer not looking.

3.2 Experiment Conditions

There were eight conditions forming three pairs: self-avatar versus no self-avatar (*Avatar*), induction versus no induction (*Induction*), singer looking at user versus singer not looking at user (*LookAt*). Each is illustrated in Figure 2.

We provided both male and female self-avatars. Users self-selected the gender in the pre-questionnaire (see Section 3.4). If they indicated that they preferred not to say what gender they were, they were assigned a random avatar if they were in the self-avatar condition.

The induction condition involves the singer saying the words "Please tap along to the beat" immediately before the song introduction. The self-avatar then appears to tap along to the beat for about 20s into the song. In the no-induction condition, the singer says nothing at this point. The animations of the singer are the same.

In the singer looking at user condition, the avatar's animation was played out as captured, with the singer's head motion targeted at the correct azimuth and elevation such that her eye gaze appeared to be at the user. In singer not looking at user condition, the avatar was rotated in azimuth 40° anti-clockwise so that she appeared to be singing into empty space.

3.3 Equipment and Programming

The scenario was programmed in Unity 5. For the GearVR the Oculus Mobile SDK 0.6 was used. For the Android devices, the Google Android SDK April 2015 version was used. Both of these SDKs provide drop-in assets and libraries to facilitate scene construction in Unity.

Five scenes were generated. The first was a simple UCL logo, which loaded very quickly once the application was launched. The second scene was the pre-questionnaire (see below). The third scene was the main bar experience. The fourth scene was the post-questionnaire and the fifth scene was a results and information screen. Transitioning from first to second and third to fourth scenes was triggered on a timer. Transitioning from second to third and fourth to fifth scenes was done on completion of the corresponding questionnaire. Exiting the fifth screen terminated the application. On Gear VR this returned the user to the Oculus Home screen. On Google Cardboard this returned the user to the main screen.

Interaction with the Pre-and Post-Questionnaires was done by a look at and dwell interaction technique. All questions were answered on either simple multiple choice, or Likert scales. An example is shown in Figure 3. The dwell time was 1.5s. Users or participants could go back in each questionnaire if required.

The motion for the singer was captured using an OptiTrack motion capture system during one performance of the song. The facial expression of the singer was captured using FaceShift during a second performance of the song. It was necessary for the singer to restrict her movement in order to remain within the capture volume for the Kinect in order to ensure a reliable data capture in FaceShift. Face and

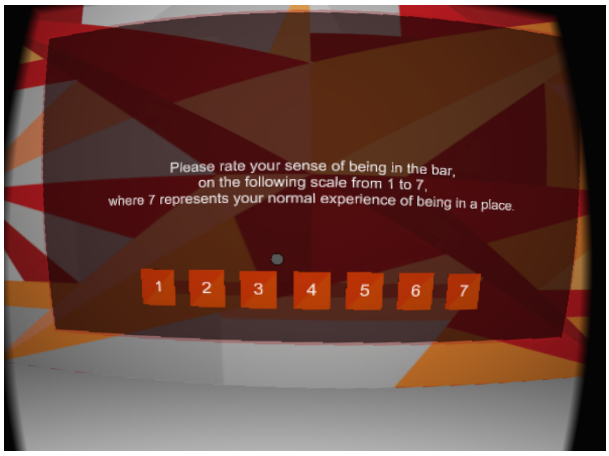

Fig. 3. An example question from the post-questionnaire. The grey circle above the '3' button indicates the current look direction.

Table 1. Pre-questionnaire. Users who answered no to data collection only answered PQ1 so that the environment could be generated with an appropriate avatar

Label	Question
PQ0	I confirm that I am over 18 years old and that I consent to take part in the study of effectiveness of virtual reality. Yes, No
PQ1	Please indicate your gender to configure the environment. Male, Female, Prefer Not to Say
PQ2	How much time do you spend playing computer or video games each week? None, Less than 1 hour, 1-2 hours, 2-5 hours, 5-10 hours, More than 10 hours
PQ3	Have you experienced virtual reality before? Never, Once or Twice, 3-10 Times, Frequently

body motion were processed in MotionBuilder to synchronise them to the audio. A simple motion for the self-avatar for conditions with induction was motion captured and processed in the same way.

The motion for the other spectator, table and box were built in MotionBuilder. The four avatars used (singer, spectator and two versions of the self-avatar) came from Mixamo. They were rigged in 3DS Studio Max. The two self-avatars were modified to remove the head geometry.

The user was not able to translate during the experience due to lack of position tracking on the target devices. They could look around based on the orientation sensing. The camera position was thus set to a generic height. The seating position helped ensure that it was not obvious that the eye height would not be correct.

3.4 In-Scene Questionnaires

3.4.1 Pre-Questionnaire

In the second, pre-questionnaire, scene, a voiceover invites the participant to read posters about the study, and then requests that the participants consent to take part in data collection (question PQ0) in Table 1. Those who do not consent to take part answer PQ1 only, in which case the answer is not recorded (the voiceover indicates this) and it is used only for configuring the environment. Those who consent to data collection are then asked questions PQ2 and PQ3.

3.4.2 Post-Questionnaire

In the fourth, post-questionnaire, scene, there are two sets of questions, see Table 2. Questions Q1-Q8 are answered by all participants. Questions Q9-Q13 are only answered by those participants in the body

Table 2. Full post-questionnaire. Subjects without a body only answered Q1 to Q8

Label	Question
Q1	Please rate your sense of being in the bar on the following scale from 1 to 7, where 7 represents your normal experience of being in a place.
Q2	To what extent were there times during the experience when the virtual reality became the 'reality' for you,and you almost forgot about the 'real world'in which the whole experience was really taking place? 1 indicates not all, 7 indicates all the time.
Q3	During the time of the experience, which was strongest on the whole,your sense of being in virtual bar, or of being in the real world?1 being the real world, and 7 virtual bar.
Q4	During the time of the experience, did you often think to yourself that you were actually just sitting in a room wearing a helmet or did the virtual reality overwhelm you?
Q5	When the box fell, how much did you feel that your hand might be hurt? 1 being not at all, 7 being I felt my hand might be hurt.
Q6	When the box fell, did you react by moving your hand? 1 being not at all, 7 being I moved my hand.
Q7	During the experience did it feel as if your hand disappeared? 1 being not at all, 7 being my hand disappeared.
Q8	During the experience did it feel as if you moved across the bar? 1 being not at all, 7 being very much.
Q9	How strong was the feeling that the body you saw was your own? 1 being not at all, 7 being very much.
Q10	How much did you feel that you were looking at your own body? 1 being not at all, 7 being very much.
Q11	How much did you feel that your real body was becoming virtual? 1 being not at all, 7 being very much.
Q12	How much did you feel feel as if you had two bodies? 1 being not at all, 7 being very much.
Q13	How much did you feel as if the virtual body became to look like your real body? 1 being not at all, 7 being very much.

condition. Note that Q8 is a control question designed to identify participants who are not reading the questions and answering repeatedly on the same score.

Q1-Q4 are based on standard presence questions from the Slater-Usoh-Steed (SUS) questionnaire [25]. Q5-Q7 and Q9-Q13 are based on a body-ownership illusion questionnaire, which originated in [3] and was altered to suit virtual environments in [23]. In both cases fewer questions were utilised.

For the body ownership illusion questions, Q5, Q6, Q9 and Q10 are expected to indicate that the participant has the illusion of the virtual arm being their own. Q7, Q11, Q12 and Q13 are considered control questions that should not be affected by a body ownership illusion. Note that Q5, Q6 and Q7 are asked of all participants as they can be answered by them without confusion.

3.5 Online Materials and Ethics

The study was approved by the UCL Research Ethics Committee after discussion with a departmental ethics review facilitator. The main concerns were the health and safety of the participant in the experiment but also other users of the application. For this reason an obvious stress re-

sponse such as a virtual pit-style experience [14] was ruled out in case a participant in the experiment, or even a user of the application who had not agreed to data collection, suffered an adverse reaction. With a vertigo-inducing experiment, there might, for example, be a risk of a fall. Although such experiences and other more extreme "jump scares" are common in publicly available demonstrations, we have a duty of care to participants and users because we have invited participants and users to take part in an experiment and the experience is not designed mainly for entertainment.

Ethics information was available online prior to the experience for any participant to read. Shortened versions of this information were available in the second scene within the app, and participants were invited to read this information by the voiceover message (see Section 3.4).

After completing the second questionnaire the participants transitioned to the fifth scene, which contained a short debriefing statement and a poster indicating what the eight conditions were and which condition they had experienced. There was an invitation to read a longer online statement about the experiment. Correspondingly, a web page was available online that explained the eight conditions. In this scene, if the user had not agreed to data collection they saw the same information about the condition they had seen.

3.6 Data Collection

From participants who agreed to data collection we collected the following information:

1. Answers to pre-questionnaire
2. Device model name
3. Anonymous unique identifier from the device.
4. Head tracking information at 10Hz
5. Answers to post-questionnaire

Items 1 to 4 were uploaded to a data collection server at the end of the third scene (the main bar experience). Item 5 was uploaded at the end of the fourth scene (the post-questionnaire). Data was uploaded to a secure server at UCL.

We offered participants the ability to have their data removed from the study. This was done by recording an anonymous identifier for each device. Participants were invited to email the identifier to a specific email address. It was planned that emails to this address would be deleted in order that no relationship between device and user could be constructed. In practice no-one requested data removal.

4 Results

4.1 Filtering

A first version of this application for Samsung Gear VR for Note 4 was made available during the VRJam Mobile competition in Q2 2015. Later a version was made available for Google Cardboard. Subsequently a version was made for Gear VR for S6 and Note 4 via SideLoadVR [30]. We have approximate numbers of installs from the relevant app stores (150 to 8th September for Cardboard for iOS, 38 for Cardboard for Android, 91 for Gear VR via SideLoad VR) but we do not have numbers for the direct installable for Gear VR from the VRJam. As the direct installable from the VRJam competition was the first available installable, we estimate that the total number of installs is in the region of 400.

The application needed to be online to record data, so anecdotally we know that some people ran the experiment, but data was not recorded because the device was offline. We did not record any information if the participant did not agree to data collection. We also know anecdotally that some people did not agree to data collection.

In total, 115 people agreed to data collection and successfully uploaded at least one of the log files. Of these, 85 formed complete sets. That is 30 did not complete the second questionnaire, stopped the application at that point or before data upload was complete, or the data upload failed. Of the 85 data sets, 26 were rejected for the following reasons: being known test versions before public release, participants scoring highly on the control question (Q8), or making the same score repeatedly in the majority of the post-questionnaire.

Table 3 gives a summary of the numbers of results and the numbers of participants in each condition. Note that participants are randomly assigned and are not matched in any way. Thus the conditions are not balanced.

4.2 Presence

Participant responses are measured with Likert scales. Although the responses are ordinal values and thus an ANOVA is not strictly appropriate, as is common we originally analysed the responses to Q1-Q7 using a three-way ANOVA. Significant results were obtained, but the residual models were not normal so the results could not be treated as reliable. The results are very similar to the those obtained in the following analyses, with the same main effects.

No main effects were found for LookAt for any question. Thus, while we include LookAt in the main models for completeness, we do not present further analysis of this factor. The discussion ventures a hypothesis about why this factor had no impact.

Questions 1-4 were answered by all participants over three conditions. The results were analyzed by an Ordinal Logistic Regression (OLR) with three factors, Avatar, Induction and LookAt. For all OLR analyses, the assumption of homogeneity of odds ratios was checked using the *omodel* package for STATA [32]. Figure 5 gives an overview of the responses to Q1-Q7 over the Avatar factor. Figure 4 gives an overview of the responses to Q1-Q7 over the Induction factor.

No effects are found for Q1, Q2 and Q4. Thus the only presence question that drew out a distinction was Q3. The model for Q3 with all three factors is not significant, ($LR\chi^2(3) = 5.73, p = 0.125$), but the expected factor Induction is significant ($z = -2.25, p = 0.024$). The coefficient for Induction is -1.071, std. err. 0.475, indicating that Induction condition reduces the score on this question compared to NoInduction condition.

4.3 Box Falling and Hand Disappearing

The two questions Q5 and Q6 ask about the participant's reaction when the box fell. The model for Q5 with all three factors is not significant, ($LR\chi^2(3) = 6.15, p = 0.1044$). The factor Avatar is significant ($z = 2.30, p = 0.022$). The coefficient for Avatar is 1.183, std. err. 0.515, indicating that self-avatar condition increases the score on this question compared to no self-avatar condition.

The model for Q6 with all three factors is significant, ($LR\chi^2(3) = 13.48, p = 0.1257$). The factor Avatar is significant ($z = 2.37, p = 0.018$). The coefficient for Avatar is 1.447, std. err. 0.610, indicating that self-avatar condition increases the score on this question compared to no self-avatar condition. The factor Induction is significant ($z = -2.39, p = 0.017$). The coefficient for Induction is -1.461, std. err. 0.610, indicating that induction condition reduces the score on this question compared to no induction condition.

The model for Q7 (hand disappearing) is not significant.

4.4 Further Questions on Body Ownership

Q5 and Q6 are initial indicators of a body ownership illusion as the participant reports that they felt as if they should, or did actually, withdraw their hand.

Q9 and Q10 can be considered an indication of a body ownership illusion, whereas Q11-Q13 should show no effect. These questions are modelled only for Induction and LookAt factors as they are not answered by participants in the no self-avatar condition.

The model for Q9 with both factors is close to significant, ($LR\chi^2(2) = 5.91, p = 0.052$), but the expected factor Induction is significant ($z = -2.31, p = 0.021$). The coefficient for Induction is -1.769, std. err. 0.766, indicating that Induction condition reduces the score on this question compared to NoInduction condition.

There is no significant model for Q10, which was not expected. There is no significant model for Q11-Q13 as expected.

Table 3. Counts of the numbers of completed participants in each condition.

	LookAt and Induction Conditions			
	NoLookAt		LookAt	
Avatar Condition	NoInduction	Induction	NoInduction	Induction
No Self-Avatar	9	6	6	10
Self-Avatar	6	7	8	7

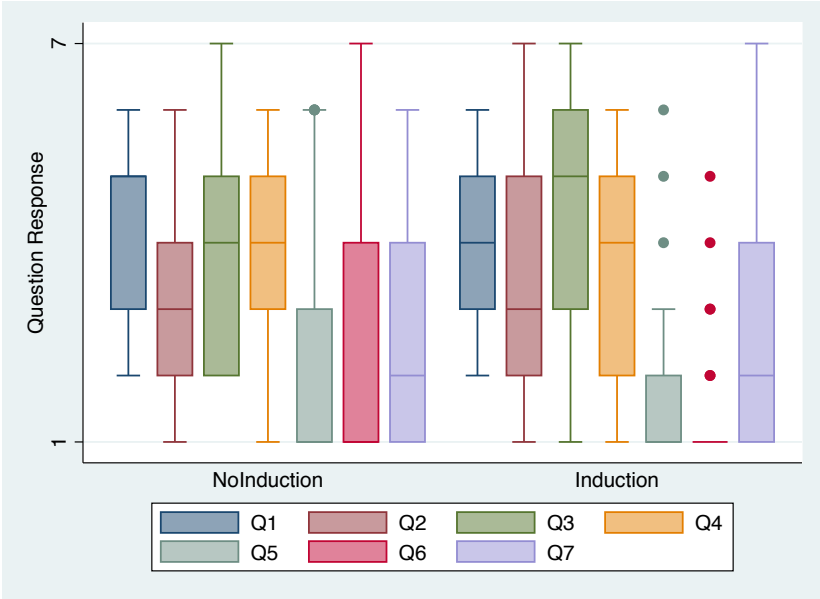

Fig. 4. Responses to Q1-Q7 for all participants, over Induction condition.

4.5 Other Data

We recorded some basic demographic data on games playing and virtual reality experience. We have not found an impact of these data on the models generated.

We also recorded head motion data, but have not analysed this further than plotting the data to spot obvious mistakes in data collection or unusual behaviours.

5 Discussion

5.1 Main Results

The results from the experiment indicate that the Avatar factor and Induction are having an impact. Some caution needs to be taken because expected effects were not found on all the presence questions and all the body ownership questions. However we take the result as a whole as good evidence that these factors are important, and also that this form of experiment can return useful results.

The presence of a self-avatar had no effects on reported presence. However, it did significantly increase the feeling that the participant might be hurt and their reporting that they reacted to the box falling. This is very encouraging as it suggests that the self-avatar is having an important impact on the participant's experience. In particular it is making the simple animated effect of the box falling become important to the participant. Whether or not they report on this scale higher because they actually have a sense of danger, or because they saw the self-avatar hit and thus thought that they should report a higher value is interesting, but both interpretations indicate that the self-avatar alters the perception of the effect.

Tapping along to the music (Induction condition) was expected to increase body ownership. However it had the opposite effect and reduced reported presence on one question (Q3) and reduced the reported indicators of body ownership illusion on Q5, Q6 and Q9. The Induction condition thus appears to be counterproductive. In retrospect, we observe that the self-avatars tapping along to the music is not necessarily synchronous with the participant's own motion. Anecdotally we know that some users reported that their movement and the avatar's movement were out of synchronisation. Thus the induction might be more like the asynchronous control condition in rubber hand illusion tests (e.g. see [23]). In a traditional rubber hand illusion induction, the real hand is synchronously tapped as the participant sees the virtual hand tapped. This is contrasted with a condition where the real tapping is asynchronous with the virtual rendering of the tapping. In the asynchronous condition it is not expected that the rubber hand illusion is induced. Our results seem to go further and suggest that the induction is counterproductive and emphasises that the virtual world is not consistent with the real world.

The lack of any effects of the LookAt factor was disappointing given that we might expect an effect. In retrospect, two observations can be made. The first is that because the singer was performing it was arguably not an interactive social situation and participants might not have been expecting the singer to engage with them. There were only two audience members, but the participant might have a variety of expectations about what the singer might do. A second observation is that although the singer is only approximately 5 metres away from the participant, it would have been very hard to see the facial expression given the resolution of the devices used. Anecdotally, we know that participants were able to tell that the singing animation was lip-synced to the facial animation and could tell whether the singer was looking at them or not, but otherwise from our own experience we know that seeing the singer's eye gaze was not possible. These limitations may be a hindrance to social presence experiences on the first generation of consumer head mounted displays.

5.2 Experimental Protocol

Although the results are promising, we can make several observations and suggestions for studies that intend to take a similar approach.

An obvious advantage of doing a study in such a manner is the po-

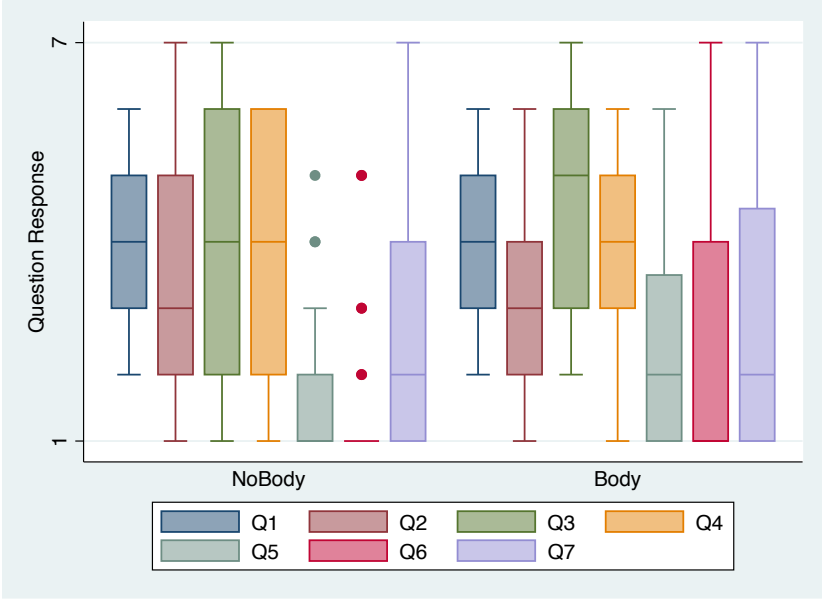

Fig. 5. Responses to Q1-Q7 for all participants, over Avatar condition.

tential number of participants that can be attracted. We do not know the total number of installs of the application, but the return of 59 completed sets of data would be a rate of approximately 15% if each device reported one data set. This bodes very well for larger scale experiments when the consumer devices are more widely available. This study was only publicised narrowly on our own web pages. There may be an opportunity to recruit interested consumers to undertake more detailed experiments, or to undertake a series of experiments for different parties. We can expect that people would volunteer for such efforts given experience with various 'citizen science' projects that have benefitted enormously from consumer engagement and participation.

In our case, the investment of effort to build and run the experiment software itself was more than a traditional lab-based environment. Developing the content and optimising it for a mobile-based platform was more challenging than for our standard lab equipment which uses top of the range GPUs and CPUs. We had to do a lot more testing of the software and stream-lining of the experience in order to make the experience reliable. For example, in a lab setting we would normally use verbal responses to questionnaires or web forms and implementing this in-app was quite time consuming. The server logging infrastructure had to be built and maintained. Some of the infrastructure and tools can be re-used for future experiments.

Related to the issue of time investment in making content, is the quality of content production that is required. Many of our lab-based experiments are simple experiences where the participant is expected to do a fixed task under supervision. We do not tend to spend a lot of effort on making the environment attractive, or give the user a variety of experiences. This has two important implications. The first is related to the issue noted above about the larger cost of developing content: we can't control what the participant does in the experiment. They can choose not to follow instructions, or might not have understood them completely. In this experiment there was no interaction with the world other than looking around. In an interactive environment a lot of care will need to be taken to make sure the participant performs tasks in a way that is reasonable and measurable. We might expect participants to find creative ways of completing tasks that were not anticipated by the developers. Logging and examining behaviour data will be necessary. The second issue is the quality of content that is displayed. When users are selecting experiences to download and try on their consumer devices, we will face an issue of competition with other apps and experiences. Thus it may be a challenge to create experiences that are attractive to consumers to try. As noted above we might expect that someone users will be motivated in order to contribute to the development of virtual reality, however many will not have encountered such requests from developers before. Fortunately many high quality assets are available to make sure that at least the experiment apps are not immediately dismissed as being visually unattractive.

A strongly related issue is that of user selection. The form of the experiment might be more or less attractive to different users. The fact that it is an experiment, and needs to be labelled as such for ethical reasons, might attract a specific type of participant. This is a key issue that depends on the motivations of the researchers. In the short term, the market for virtual reality is early adopters. A legitimate motivation for a study would be to optimise virtual reality for this population. However, there may be a risk that optimising for this population may reduce accessibility to others. If the study was more applied in nature, such as exploring cognitive biases or spatial memory, one would hope to get a more diverse set of participants. In our experiment we did not collect much demographic data, but in an applied study this would be more valuable. The downside with collecting more data is that the experiment will take longer and thus compliance may be lower. A way around this may be to build profiles of users who will contribute to a series of studies. We would note that one significant positive of an in the wild study is access to a potentially broader set of participants than might typically be available for a lab-based study.

Another related issue is that of replication. Although technically our study can be replicated relatively easily because the materials can be accessed openly, there were a relatively small number of participants in our study. We would hope that this would be less of an issue for future studies that could attract many more users and thus many more complete sets of data from participants.

Finally we note that as a community, we need to understand the ethical implications of more complex studies. As noted in Section 3.5, we chose a non-threatening environment partly because of the risk of scaring an individual who might react negatively or even injure themselves when stepping away from a threat. To some extent, participants are adults and are self-selecting to undertake the experiment. Further, given this is consumer equipment and that there will be heavily publicised content on each platform, any experiment is unlikely to be the first virtual reality experience a participant will have. Thus, one might make the argument that the experiment is no more risky than other experiences. However, there are still risks. Thus studies in this format might be more approriate for understanding interaction and rep-

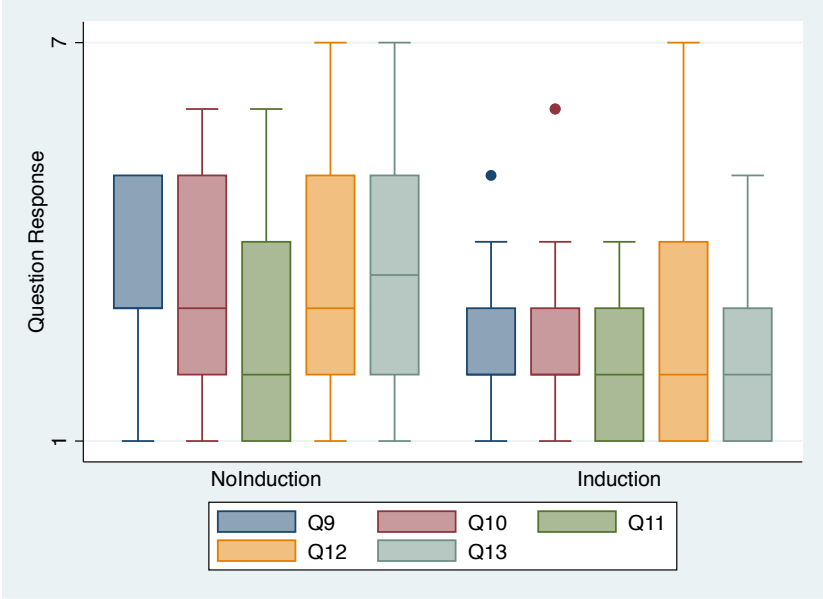

Fig. 6. Responses to Q9-Q13 for participants in self-avatar conditions, over Induction condition.

resentation issues for virtual environments, rather than more complex psychological phenonema such as stress responses.

5.3 Specific Critiques

In Section 4 we noted some anecdotal feedback about the application. Further feedback that we received included the opinion that the song was quite long and there was not much to do. We might hypothesize that in our case, some participants who did not complete the second questionnaire had actually become bored and stopped. This is related to the issue above of having the content in the experiment be comparable to other content that the users will experience. We did try to make the experience interesting: relatively few early apps have full body animation and the song and performance were very good. We already made the performance quite short by shortening the song. Thus, with different content, it may be possible to achieve a higher rate of installs to returned data than the 15% that we achieved.

The dwell to select technique attracted some criticism, but some similar technique was necessary because of the lack of button input. Many developers have adopted similar techniques. One quite common feature that we did not have is visual feedback about the progression of the dwell towards select. We would suggest the dwell selection technique in the games Lands End as an example of good feedback [29], but there are several others.

We did have feedback that the second questionnaire was quite long, especially for those in the self-avatar condition. From our results, there appear to be questions that could be deleted in order to simplify the questionnaire for this particular set-up. However, this is not transferable to other studies and we do not recommend adopting only those questions we found useful.

6 Conclusion

In this paper we have presented an in the wild experiment on virtual reality. The study was undertaken by unknown participants using consumer virtual reality equipment, in particular the Samsung Gear VR for Note 4 and S6, and various Google Cardboard devices running Android and iOS. We received 59 completed sets of data. This was a rate of return of 15% if each device the experiment app was installed on reported one data set.

The results of the experiment were not very strong, but very promising for a study of this scale. The presence of a self-avatar seemed to alter the response to the virtual box falling. Even if this is solely because the participant sees the box hit the self-avatar and thus a collision occurred, rather than reacting because they thought that the box would hit them, this is interesting.

The attempt to generate a body ownership illusion failed, but in doing so highlighted an important potential guideline for application developers: that the self-avatar should not be animated if the participant cannot be tracked. The lack of hand tracking, particularly on smartphone-based HMDs, is thus a problem both because it hinders interaction and also because it seems to hinder the types of body ownership illusion that can be achieved with immersive virtual reality systems. It remains to be seen whether external position tracking of some sort is required, or whether camera or HMD mounted devices can track the hands and body sufficiently well to allow both interaction and engagement with a self-avatar.

There are many potential avenues for next steps. It would be interesting to validate the results in a lab-based experiment, especially the idea of self-induction which we could implement properly with hand tracking. An obvious route for in the wild studies would be to probe individual differences in presence response, to identify what factors might broaden the appeal of virtual reality, or to compare different types of system. We do not have sufficient data yet to probe whether presence or body ownership illusion were higher on Gear VR or Google Cardboard, but once a broader range of mobile and desktop PC-based systems are available, the experiment can be re-run or extended to make such comparisons. We also hope to extend the procedure to interactive experiments to compare locomotion and interaction strategies. We expect that individual differences will become important in more interactive scenarios. It could be that by contributing data to a larger study, participants can collectively highlight both good and poor designs to inform the developer community. However, we expect that the design space will be extremely large, and thus perhaps the more interesting opportunity is to give feedback to users about what preferences they might select in new applications based on their feedback about specific test environments. For example, a user may have a preference for particular locomotion technique. Thus developers might be able to rely on a profile of preferences for a user, rather than having the user re-select their preferences in every application.

To conclude, our study shows that collecting data in the wild is feasible for virtual reality systems. We anticipate that this process can be very valuable in developing or testing community guidelines about best practice for application development. Our study also makes a first contribution to this best practice by confirming that the virtual body has a potentially important impact on presence and embodiment, and

that animation of the avatar that is uncoordinated with the participant's own motions may impair their presence and sense of embodiment in the virtual reality.

ACKNOWLEDGEMENTS

Thanks to Kelvin Wong and Meg Hollands for performing the song, No Guarantee by Bosco and Peck. The song was licensed from beatpick.com. Thanks to Ben Metsers for audio design and recording.

REFERENCES

[1] F. Biocca. Connected to my avatar. In *Social Computing and Social Media*, pages 421–429. Springer, 2014.

[2] R. Bonney, C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11):977–984, 2009.

[3] M. Botvinick, J. Cohen, et al. Rubber hands' feel'touch that eyes see. *Nature*, 391(6669):756–756, 1998.

[4] B. Brown, S. Reeves, and S. Sherwood. Into the wild: challenges and opportunities for field trial methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1657–1666. ACM, 2011.

[5] H. Cramer, M. Rost, and F. Bentley. Special issue: An introduction to research in the large. *International Journal of Mobile Human Computer Interaction*, (Special issue), 2011.

[6] A. Evans and J. Wobbrock. Taming wild behavior: the input observer for obtaining text entry and mouse pointing measures from everyday computer use. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1947–1956. ACM, 2012.

[7] C. Heeter. Being there: The subjective experience of presence. *Presence: Teleoperators and virtual environments*, 1(2):262–271, 1992.

[8] N. Henze and S. Boll. Push the study to the app store: Evaluating off-screen visualizations for maps in the android market. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '10, pages 373–374, New York, NY, USA, 2010. ACM.

[9] N. Henze, M. Pielot, B. Poppinga, T. Schinke, and S. Boll. My app is an experiment: Experience from user studies in mobile app stores. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 3(4):71–91, 2011.

[10] N. Henze, E. Rukzio, and S. Boll. 100,000,000 taps: analysis and improvement of touch performance in the large. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 133–142. ACM, 2011.

[11] W. A. IJsselsteijn, H. de Ridder, J. Freeman, and S. E. Avons. Presence: concept, determinants, and measurement. In *Electronic Imaging*, pages 520–529. International Society for Optics and Photonics, 2000.

[12] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.

[13] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff. A cross-media presence questionnaire: The itc-sense of presence inventory. *Presence*, 10(3):282–297, 2001.

[14] Q. Lin, J. J. Rieser, and B. Bodenheimer. Stepping off a ledge in an hmd-based immersive virtual environment. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '13, pages 107–110, New York, NY, USA, 2013. ACM.

[15] E. A. McManus, B. Bodenheimer, S. Streuber, S. de la Rosa, H. H. Bülthoff, and B. J. Mohler. The influence of avatar (self and character) animations on distance estimation, object interaction and locomotion in immersive virtual environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, APGV '11, pages 37–44, New York, NY, USA, 2011. ACM.

[16] D. McMillan, A. Morrison, and M. Chalmers. Categorised ethical guidelines for large scale mobile hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1853–1862, New York, NY, USA, 2013. ACM.

[17] A. Morrison, D. McMillan, and M. Chalmers. Improving consent in large scale mobile hci through personalised representations of data. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14, pages 471–480, New York, NY, USA, 2014. ACM.

[18] B. Ries, V. Interrante, M. Kaeding, and L. Anderson. The effect of self-embodiment on distance perception in immersive virtual environments. In *Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology*, VRST '08, pages 167–170, New York, NY, USA, 2008. ACM.

[19] M. V. Sanchez-Vives, B. Spanlang, A. Frisoli, M. Bergamasco, and M. Slater. Virtual hand illusion induced by visuomotor correlations. *PLoS ONE*, 5(4):e10381, 04 2010.

[20] M. J. Schuemie, P. Van Der Straaten, M. Krijn, and C. A. Van Der Mast. Research on presence in virtual reality: A survey. *CyberPsychology & Behavior*, 4(2):183–201, 2001.

[21] U. Schultze. Embodiment and presence in virtual worlds: a review. *Journal of Information Technology*, 25(4):434–449, 2010.

[22] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557, 2009.

[23] M. Slater, D. Perez-Marcos, H. H. Ehrsson, and M. V. Sanchez-Vives. Towards a digital body: the virtual arm illusion. *Frontiers in human neuroscience*, 2, 2008.

[24] M. Slater and A. Steed. A virtual presence counter. *Presence*, 9(5):413–434, 2000.

[25] M. Slater, M. Usoh, and A. Steed. Depth of presence in virtual environments. *Presence*, 3(2):130–144, 1994.

[26] M. Slater and S. Wilbur. A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and virtual environments*, 6(6):603–616, 1997.

[27] M. Usoh, K. Arthur, M. C. Whitton, R. Bastos, A. Steed, M. Slater, and F. P. Brooks, Jr. Walking >>walking-in-place >>flying, in virtual environments. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 359–364, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[28] M. Usoh, E. Catena, S. Arman, and M. Slater. Using presence questionnaires in reality. *Presence*, 9(5):497–503, 2000.

[29] Ustwo Games. Lands End. `http://www.landsendgame.com/`, 2015. Accessed 6-December-2015.

[30] VR Bits. SideloadVR. `http://http://sideloadvr.com//`, 2015. Accessed 9-September-2015.

[31] B. G. Witmer and M. J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and virtual environments*, 7(3):225–240, 1998.

[32] R. Wolfe. Omodel: Stata modules to perform tests on ordered probit and ordered logit models. `https://ideas.repec.org/c/boc/bocode/s320901.html/`, 2015. Accessed 9-September-2015.

[33] Y. Yuan and A. Steed. Is the rubber hand illusion induced by immersive virtual reality? In *Virtual Reality Conference (VR), 2010 IEEE*, pages 95–102. IEEE, 2010.