Virtual reality studies outside the laboratory

Mottelson, Aske; Hornbæk, Kasper

Published in: Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology

DOI: 10.1145/3139131.3139141

Publication date: 2017

Document version Publisher's PDF, also known as Version of record

Virtual Reality Studies Outside the Laboratory

Aske Mottelson
Department of Computer Science
University of Copenhagen
amot@di.ku.dk

Kasper Hornbæk
Department of Computer Science
University of Copenhagen
kash@di.ku.dk

ABSTRACT

Many user studies are now conducted outside laboratories to increase the number and heterogeneity of participants. These studies are conducted in diverse settings, with the potential to give research greater external validity and statistical power at a lower cost. The feasibility of conducting virtual reality (VR) studies outside laboratories remains unclear because these studies often use expensive equipment, depend critically on the physical context, and sometimes study delicate phenomena concerning body awareness and immersion. To investigate, we explore pointing, 3D tracing, and body-illusions both in-lab and out-of-lab. The in-lab study was carried out as a traditional experiment with state-of-the-art VR equipment; 31 completed the study in our laboratory. The out-of-lab study was conducted by distributing commodity cardboard VR glasses to participants; 57 completed the study anywhere they saw fit. The effects found in-lab were comparable to those found out-of-lab, with much larger variations in the settings in the out-of-lab condition. A follow-up study showed that performance metrics are mostly governed by the technology used, where more complex VR phenomena depend more critically on the internal control of the study. We argue that conducting VR studies outside the laboratory is feasible, and that certain types of VR studies may advantageously be run this way. From the results, we discuss the implications and limitations of running VR studies outside the laboratory.

CCS CONCEPTS
-Human-centered computing →Virtual reality; User studies; Empirical studies in HCI;

KEYWORDS
Consumer VR; Google Cardboard; User Studies; Crowdsourcing

ACM Reference format:
DOI: 10.1145/3139131.3139141

1 INTRODUCTION

The recent advance in consumer technology has accelerated research in virtual reality (VR). In particular, a host of VR user studies are being conducted. They include both evaluations of the usability and user experience of particular VR applications, as well as behavioral research using VR. The former includes evaluating games and educational applications (e.g., [Bolton et al. 2014; von Zadow et al. 2013]). The latter includes simulating environments to conduct experiments that would otherwise be difficult (e.g., [Pan et al. 2016; Slater et al. 2013]), impossible (e.g., [Banakou et al. 2013; Kilteni et al. 2012; Slater et al. 2010]) or even unethical (e.g., [Slater et al. 2006]) to carry out using classical experimental paradigms.

Conducting VR studies, however, faces similar decisions about practical matters and research validity as running studies using non-VR technology: those decisions and their associated trade-offs are well described (e.g., [Hornbæk 2013; McGrath 1995]). For instance, much planning goes into recruiting people, managing schedules, selecting environments in which to conduct studies, and running the actual studies. Nevertheless, VR studies are almost exclusively done in laboratories using specialized equipment (e.g., for tracking) and few and homogeneous participants (e.g., typically fewer than 25 participants recruited through university mailing lists). In that respect, VR studies are similar to studies from other parts of HCI [Caine 2016; Hornbæk et al. 2014].

For non-VR technologies, many of these studies are now done outside the laboratory, for instance with crowdsourcing or in-the-wild studies. In crowdsourcing, user studies are conducted as micro tasks giving small amounts of payment on crowdsourcing platforms such as Amazon Mechanical Turk or Crowdflower [Kittur et al. 2008]. Research shows that crowdsourcing often give a higher diversity of participants [Mason and Suri 2012; Paolacci and Chandler 2014; Ross et al. 2010] and that they can be done at a low cost [Buhrmester et al. 2011; Kittur et al. 2008; Mason and Suri 2012], reliably [Buhrmester et al. 2011; Crump et al. 2013; Rouse 2015], and quickly [Kittur et al. 2008].

Although out-of-lab experimentation has been applied in many computing areas (e.g., [Carter et al. 2007; Heer and Bostock 2010; Kittur et al. 2008; Mottelson and Hornbæk 2016; Reinecke and Gajos 2015]), it is not clear if that is feasible or valid for VR. Earlier work has suggested that this type of experimental practice is ill suited for tasks that depend on the physical environment [Heer and Bostock 2010]. Also, many VR studies depend on deadlines still not in common use and, sometimes, other equipment that is not widely available (e.g., for tracking or physical stimulation). Finally, participants in unsupervised experiments are not always paying attention, switch tasks frequently [Gould et al. 2016], and may decide to pause an experiment; all of these behaviors could interfere with goals of VR studies, such as generating perception of presence. Goodman et al. [Goodman et al. 2013] stressed that “MTurk participants are less likely to pay attention to experimental materials”, which could reduce the effects of experimental manipulations.
We explore the possibility of conducting out-of-lab VR studies, and compare experiments in uncontrolled settings using commodity VR technology to doing them in the laboratory. We distributed VR cardboard glasses to 57 participants for use with participants’ own smartphones in exchange for their participation in the study involving three canonical experimental VR tasks. The results show that it is a feasible way to conduct affordable, ecologically valid, and large-scale VR studies outside the laboratory. Additionally we discuss potential directions for out-of-lab VR experimentation, and how crowdsourcing is an interesting platform for future VR studies.

2 RELATED WORK

VR studies have been organized in a variety of ways, including analytic evaluation techniques (e.g., [Bach and Scapin 2010; Sutcliffe and Gault 2004]) as well as empirical ones (e.g., [Banakou et al. 2013; Kilteni et al. 2012; Slater et al. 2010]). The literature also contains evaluations of usability issues and user experience with VR headsets [McGill et al. 2015], and Marsh [Marsh 1999] discussed some issues in evaluating the usability of VR. Here we focus on empirical user studies using VR and first discuss those briefly. Then we review types of user studies conducted outside of laboratories, and outline the potential of that methodology for VR.

2.1 VR Studies

Virtual reality research has for long been engaged in user studies, and with the availability of consumer oriented VR technology a host of VR studies are being conducted. Those studies include evaluations of the usability and user experience of particular VR applications, such as games (e.g., [Bolton et al. 2014]) and educational applications (e.g., [von Zadow et al. 2013]). Another line of behavioral research uses VR to study phenomena relating to body perception and body schema. These often employ body ownership illusions, which are studies where participants perceive non-bodily objects, or alterations of their own body to be parts of their own body [Kilteni et al. 2015]. These illusions are usually made feasible by means of synchronous stimulation of the virtual and physical body [Slater et al. 2008, 2009]. With this, perceptions of objects sizes have been shown to be influenced by hand size alterations [Linke nauger et al. 2013], and racial attitudes are found to be influenced by ownership of an other skin-toned body [Maister et al. 2013]. Also, ownership of a child body has been shown to cause faster identification of child-like attributes [Banakou et al. 2013].

Researchers in virtual reality studies face many of the same concerns that go into doing any user study, such as recruiting people, managing schedules, selecting environments, and running experiments. Other concerns relate to validity and research methodology; similar to concerns about user studies in other domains (e.g., [Hornbaek 2013; McGrath 1995; Shadish et al. 2002]). While locomotion has historically been a critical topic within VR research [Slater et al. 1995; Usoh et al. 1999], many VR studies are conducted stationary (e.g., sitting, standing, lying). Most consumer oriented VR applications are also used stationary: the most popular consumer VR system that supports walking for locomotion (HTC Vive), reports standing as the most common configuration amongst its users [STEAM 2017]. This could indicate that many VR user studies could be straightforward for participants to conduct without the guidance of a human evaluator. If the user studies conducted in VR research bear at least some similarities to user studies conducted in other parts of HCI, might a shift in experimental practice from laboratory to out of laboratory (which is widely successful in other parts of HCI) be beneficial for VR studies?

2.2 Out-of-lab Studies in HCI

External validity concerns whether a causal relationship holds over persons, settings, treatments, and outcomes [Shadish et al. 2002]; that is, to which extent findings are generalizable to a broader domain. In an attempt to increase the external validity of a study’s results, researchers may conduct their research outside of a laboratory. Unlike observational research such as field studies, some out-of-lab research practices allow researchers to control experimental conditions and manipulate independent variables. These unsupervised experimental practices, such as crowdsourcing and in-the-wild experiments, have been an ongoing endeavor within HCI for a while [Brown et al. 2011; Carter et al. 2007; Kjeldskov and Skov 2014].

Crowdsourcing. In crowdsourcing, user studies are conducted as micro-tasks giving small amounts of payment on crowdsourcing platforms such as Amazon Mechanical Turk or Crowdflower [Kittur et al. 2008]. Crowdsourcing has shown to be a valuable experimental practice that allows for fast and low-cost experimentation, with high diversity of participants [Buhmester et al. 2011; Crump et al. 2013; Kittur et al. 2008; Mason and Suri 2012; Paolacci and Chandler 2014; Ross et al. 2010; Rouse 2015].

In-the-wild. Conducting in-the-wild experiments has long been a research agenda within the ubiquitous computing community, and as such a variety of protocols for conducting unsupervised experimental research has evolved. An alternative to the popular micro-task platforms includes LabintheWild [Reinecke and Gajos 2015], a highly scalable way of conducting studies with widespread, uncompensated, and unsupervised participation. The authors created an online experimental platform that provides participants with information about themselves in exchange for their participation in studies. In-the-wild mobile experiments have also been conducted, Henze et al. [Henze et al. 2011] for instance did in-the-wild mobile experiments, with mobile app store distributed gaming-based user studies.

2.3 Potential of out-of-lab studies for VR

When is high external validity key to VR research? For some VR studies, high external validity is of less concern than others. This is the case when VR is being used to mitigate arachnophobia [Garcia-Palacios et al. 2002], estimate general practitioners’ susceptibility to prescribing antibiotics [Pan et al. 2016] - and in general studies with homogeneous participants and few experimental settings, especially for within-subjects designs. For studies concerning a heterogeneous population, using subtle differences between conditions, with more experimental settings, the external validity is of much higher concern for the integrity of the research. This is often true when employing a between-subjects design. Could out-of-lab experimentation be a worthwhile methodology to consider for such
studies? Unfortunately, it is not clear if it is feasible to use widespread out-of-lab experimental practices to conduct VR studies, or whether these approaches are valid. While crowdsourcing could give larger samples of more varied participants, the widespread adoption of VR consumer devices has yet to happen which makes it difficult to recruit participants for a crowdsourced VR study.

While the potential of conducting VR studies out-of-lab is promising, the associated questions are severe. Although increasing the generalizability of VR research has been an ongoing agenda (e.g., [de Kort et al. 2003]), we are only aware of one study that conducted out-of-lab VR experimentation; Steed et al. [Steed et al. 2016] ran a mobile app-based experiment, gathering user stories from owners of household VR devices such as Google Cardboard and Samsung Gear to study presence and embodiment in a bar with a singer. The authors did a between-subjects study with eight conditions, exploring among other things the prevalence of a self-avatar on presence, hand-tapping, and eye contact with a virtual person. While this study is a valuable example of using consumer VR technology to conduct studies using mobile app stores, it does not conclude on the validity or feasibility of using that approach. Also, the authors employed an untried procedure, which makes it difficult to separate effects from the experimental procedure and the methodology.

2.4 How to study VR outside the laboratory

We surveyed crowdworkers to understand the types of VR equipment at their disposal. We asked 250 people (for 0.05$ USD pay, 92% validated using a verifiable control question) about their ownership of computer equipment using a randomized ordered checklist of household computer technology. We found that at the time of writing, 3% of crowdworkers own one or more devices capable of VR. In particular participants reported ownership of the following: Google Cardboard (2.2%), Samsung Gear (1.3%), HTC Vive (0.9%), and Oculus Rift (0.4%). In comparison, 83.4% of the respondents reported ownership of an Android smart phone. The effective size of the active MTurk population has been estimated to be about 7300 workers [Stewart et al. 2015]. If our sample is representative of the MTurk population, we should expect at most 226 crowdworkers to own a VR device. Thus, the modest share of crowdworkers who own VR equipment at the time of writing make it unrealistic to use popular micro-task markets to crowdsourced VR experimentation. Because of the widespread adoption of consumer smart phones combined with Google Cardboard as a cheap alternative to other VR technology, we see a contemporary opportunity for inexpensive large-scale out-of-lab VR experimentation. To provide insights about the feasibility and validity of conducting out-of-lab VR studies, we propose a study protocol where participants are recruited online, pre-screened prior to participation, and provided with commodity cardboard VR glasses to participate in the study.

Crowdsourcing is arguably not the correct term for the approach tried in this work, because we pre-screen participants and require them to visit our premises. Accordingly, the term in-the-wild seems inaccurate, since the experimental nature of our setup will enforce an artificial controlled setting, not expected to occur completely in-the-wild (we did not expect any voluntary participation from regular app store users). When HMDs become more prevalent, it will be possible to completely crowdsourc VR studies without requiring participants to visit the premises. In the remainder of this work, we employ the term out-of-lab to describe the method of equipping pre-screened participants with cardboard VR glasses, and have them conduct experiments in non-controlled settings. This is opposed to when we speak about in-lab studies, which covers experimentation in controlled settings, at our research facilities. We conducted experiments using canonical VR paradigms, all previously verified in laboratories. We study the implications of conducting out-of-lab VR experiments by directly comparing the participants’ differences across settings and technology.

3 EXPERIMENT

The purpose of the experiment was to validate the potential of conducting VR studies outside the laboratory, and to do so we compare in-lab to out-of-lab VR experiments over a range of VR phenomena. The studies were set up to be representative of how VR studies are usually done (i.e., laboratory VR studies usually do not use commodity technology).

3.1 Participants

We posted the invitation to participate in our experiment on large group for locals on Facebook, in addition to sending invitations using our internal e-mail list. Participants signed up online for either the in-lab or out-of-lab study. In both cases, participants came to our premises; either to pick up a set of cardboard glasses, or to participate in our lab-study.

In-lab. Thirty-one people, aged 20-50 (SD = 10.3 years) participated in the laboratory study and were reimbursed with a gift worth the equivalent of 15$ US. Our regular minimum rate for lab-study participation. Of these participants were 12 male.

Out-of-lab. 100 participants were given a Google cardboard for participating in the study. Fifty-seven participants, aged 20-40 (SD = 7.0 years) completed the study within 20 days; of these were 35 males, with 34 using iOS and 23 using Android.

3.2 Apparatus

We developed the VR tasks using Unity 5.3. The VR application was identical for in-lab and out-of-lab, except that the VR equipment held by the avatar was substituted to match the visuals of the actual VR equipment. The applications would send relevant user metrics to a server application written in Python. The application contained on-screen instructions that blended in with the VR environments.

In-lab. The VR studies in-lab used an HTC Vive.

Out-of-lab. We deployed the VR application with the Google Cardboard SDK, distributed at relevant application stores for both Android (version ≥5.0) and iOS (version ≥7.0).

3.3 Design

Participants conducted three independent tasks; one without any experimental variation, and two with a between-subject two-condition design (see Table 1). Participants were randomly assigned to the experimental conditions on a per-task basis. The three tasks were administered in a randomized order. The study procedure was carried out both in-lab and out-of-lab with different participants.
with four IDs (range 2-4), and two repetitions per ID. Every experiment had expected experimental effects that differ between multiple conditions.

Table 1: The three tasks employed with their corresponding experimental conditions and dependent variables.

<table>
<thead>
<tr>
<th>Task</th>
<th>Conditions</th>
<th>Dependent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-lab</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pointing</td>
<td></td>
<td>MT, accuracy, TP</td>
</tr>
<tr>
<td>3D Tracing</td>
<td>Dynamic</td>
<td>Duration, estimations</td>
</tr>
<tr>
<td>Body Ownership</td>
<td>Consistent</td>
<td>BO, presence</td>
</tr>
<tr>
<td>Illusion (BOI)</td>
<td>Inconsistent</td>
<td></td>
</tr>
<tr>
<td>Out-of-lab</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The intention of this design was to combine tasks where absolute performance values could be compared (pointing task) and tasks with expected experimental effects that differ between multiple conditions (3D tracing, body ownership illusion). For the latter we are interested in comparing the outcomes of experimental conditions for in-lab and out-of-lab. To validate the feasibility of conducting VR experimentation outside the laboratory, we hypothesize that out-of-lab studies yield similar effects and effect sizes to those conducted in-lab.

3.4 Tasks

The participants were presented with three VR tasks as below. The intention was to get insights of VR experimentation in-lab and out-of-lab in a broad range of tasks; we therefore employed three different complexities of VR experimentation: a pointing task, a 3D tracing task, and a body illusion task. The tasks are based on established experimental protocols about performance aspects of HCI and perception in VR. The tasks do not require the participant to travel across physical space, making them suitable to conduct virtually anywhere and without the need of a human evaluator.

Pointing Task. The goal of this task was to show the feasibility of collecting performance metrics from elementary VR navigation. We studied participants’ performance with 2D navigation within a VR environment, using a common Fitts’s Law task (see Figure 1a). Most aspects of the task adhered to Soukoreff and Mackenzie [Soukoreff and MacKenzie 2004], but to alleviate fatigue we used 15 targets with four IDs (range 2-4), and two repetitions per ID. Thus every participant pointed at 120 targets, excluding a warm-up round. Translational movements were ignored for this task.

3D Tracing Task. This task measured participants’ performance in judging depth and navigating in VR. The task by Arthur et al. [Arthur et al. 1993] compares users’ performance on distinguishing 3D objects in different viewing conditions. Two trees composed of straight lines were placed next to each other (see Figure 1b). Each tree consisted of three levels of branches, resulting in 27 branches for each tree excluding the root. For each trial, one of the trees would contain a yellow leaf (the leaf was placed on the branch with the x coordinate nearest to the center), and the participant then distinguished which of the trees the leaf belonged to. Participants were randomly assigned to either of two conditions: (1) participants could inspect the 3D spatial properties of the trees dynamically by moving their head around, or (2) participants were presented with a static view, requiring them to determine the origin of the leaves having seen the trees from one angle only. Textual feedback (correct/incorrect) was provided after each selection for one second. This task was intended to test if out-of-lab participants could use the 3D spatial capabilities of VR to increase depth judgment accuracy. For each trial of the 40 trials, a leaf was randomly placed on a branch belonging to either the left or right tree.

Table 2: Body ownership illusion task post-questionnaire, from [Banakou et al. 2013; Slater et al. 1994].

<table>
<thead>
<tr>
<th>Q#</th>
<th>Question</th>
<th>Scale</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>How much did you feel that the virtual body you saw when you looked down at yourself was your own body?</td>
<td>Not at all</td>
<td>Body Ownership</td>
</tr>
<tr>
<td>Q2</td>
<td>How much did you feel that the virtual body you saw when you looked down at yourself in the mirror was your own body?</td>
<td>Very much</td>
<td>Body Ownership</td>
</tr>
<tr>
<td>Q3</td>
<td>How much did you feel that your virtual body resembled your own (real) body in terms of shape, skin tone or other visual features?</td>
<td>Not at all</td>
<td>Body Ownership</td>
</tr>
<tr>
<td>Q4</td>
<td>How much did you feel as if you had two bodies?</td>
<td>Very much</td>
<td>Body Ownership</td>
</tr>
<tr>
<td>Q5</td>
<td>During the experience, did it feel as if you moved across the bed room?</td>
<td>Not at all</td>
<td>Control</td>
</tr>
<tr>
<td>Q6</td>
<td>Please rate your sense of being in the bedroom, where 1 represents the normal experience of being in a place.</td>
<td>1</td>
<td>Presence</td>
</tr>
<tr>
<td>Q7</td>
<td>We will enter with three minutes during the experience when the virtual reality became the ‘real’ for you, and you almost forget about the ‘real world’ in which the whole experience was really taking place?</td>
<td>All the time</td>
<td>Presence</td>
</tr>
<tr>
<td>Q8</td>
<td>During the time of the experience, which was strongest on the whole, your sense of being in the virtual room, or of being in the real world?</td>
<td>Real world</td>
<td>Presence</td>
</tr>
<tr>
<td>Q9</td>
<td>How much did you feel as if you had two bodies?</td>
<td>Very much</td>
<td>Body Ownership</td>
</tr>
</tbody>
</table>

Figure 1: The tasks. (a) Pointing: participants targeted the red spheres as fast and accurate as possible by moving their head. (b) 3D tracing: participants selected which tree a yellow leaf belonged to. Participants could either inspect the trees dynamically by moving their head around, or could only see the trees from one angle. (c) Body Ownership Illusion: participants were immersed in a virtual bedroom, with their bodies substituted with sex-matched avatars. In half of the cases avatars mapped the participants’ movements real-time. A mirror that reflected the virtual body was present in the bedroom as shown.
Body Ownership Illusion Task. Body ownership illusions refer to the class of illusions where participants perceive virtual bodies to be their own [Kilteni et al. 2015]. The illusion of ownership of virtual bodies has been shown feasible by means of consistent stimulation of the virtual and physical body (e.g., using a rod) [Slater et al. 2008, 2009]. To study the feasibility of conducting out-of-lab VR tasks involving more complex VR phenomena, we designed a body ownership illusion task inspired by Banakou et al. [Banakou et al. 2013]. The intention was to induce the illusion of body ownership, and the feeling of being present in a virtual room. The participants were asked to look around and take notice of the room for two minutes. The participants could look down and see a sex-matched virtual body. Additionally, a mirror was present where the participant’s avatar could be seen (see Figure 1c). The task employed two conditions: consistent and inconsistent visuomotor stimuli. In the consistent condition, participants’ movements were mapped real-time to the avatar, both when looking down at the virtual body and when looking in the mirror. With participants hands fixed in a binocular pose, we only mapped the upper torso using a simple inverse kinematics system. In the inconsistent condition, the avatar’s body did not reflect participants’ movements. We expected the consistent condition to result in higher degrees of body ownership and presence, as previous studies (e.g., [Banakou et al. 2013; Sanchez-Vives et al. 2010]). We employed a questionnaire that quantified body ownership and presence on a [−3, 3] Likert scale.

3.5 Procedure

All tasks started with a textual description of the task. Participants would have to trigger a begin button to initiate a task. All button selections were done by dwelling two seconds at a target using a cross hair that triggered visual feedback.

In-lab. Participants first signed an informed consent form, and were then placed standing in the middle of a 4×4m room. To minimize effect of disturbing noise from our laboratory, a noise cancellation headset was put on, and together with the HMD placed on the participants’ heads by the evaluator. To minimize effects of body posture compared to the out-of-lab study, we asked the participants to keep a posture similar to that when using a Google cardboard (arms and hands in binocular pose) during the entire study. In the same fashion, although VR applications for the HTC Vive are usually controlled using hand-carried controllers, we employed a dwell-based head controlling system to gather comparable data to the out-of-lab cardboard study. An evaluator stayed in the room with the participant during the study and made sure directions were adhered to.

Out-of-lab. Participants were invited to come by our premises to pick up their Google cardboard (if their phone was capable of Android version ≥5.0 or iOS version ≥7.0). Together with the commodity VR equipment, participants were provided with instructions on how to acquire the experimental application and carry out the experiment, in addition to descriptions of the extent of the data collection and associated privacy concerns.

3.6 Ethical Concerns

The immersiveness of VR combined with participants conducting the study protocol on their own give rise to a number of concerns, and out of ethical concerns, some VR experiments should be avoided as out-of-lab studies. Additionally collecting data for publicly open studies requires some consideration. First, opposite to a laboratory-based experiment, an evaluator is not present to help subjects, for instance in case of motion sickness or falling over objects. Second, our experimental application also took two photos we used to analyze participants’ surroundings, and to confirm that the phone was in fact correctly placed in the cardboard VR glasses. How to ethically log user data from mobile experiments depends on several factors, and a simple solution to this question does not seem to exist. We however followed directions proposed by Henze et al. [Henze et al. 2011] and informed users prior to participation about the logging and thus implicitly get consent from the user by continuing use of the application.

3.7 Data Validation

As in other studies, attention and compliant participation is key. This is especially difficult to verify in out-of-lab studies because of the absence of a human evaluator during the study. In addition to exclusion criteria based on earlier work [Kittur et al. 2008; Steed et al. 2016], we used the front-facing camera of participants’ phones to take a photo that was later used to determine whether the phone was accurately placed inside the cardboard equipment (this was mentioned in the experiment invitation). In summary, we used the following criteria to disqualify participants:

- Errorneous placement of phone in VR equipment (front camera)
- Zero variance in questionnaire responses
- High response to control question (Q5) (> 2)
- Too slow completion time (> M + 3 SD)
- Too fast response to questionnaire (< M − 3 SD)

Four participants were discarded from the in-lab study, one because of zero variance in the questionnaire answers, three because of the control question (Q5); thus 27 participants remained. For the out-of-lab study, we also discarded four participants, one for not placing the phone in the VR glasses, one from taking too long, and two because of the control question; 53 participants remained.

4 RESULTS

The overall purpose of the result section is to present differences in dependent variables due to experimental conditions, and to do comparison between the in-lab study and the out-of-lab study. We report results from each of the three tasks, divided by the type of analysis. Where nothing else is indicated, statistical tests were done with a one-way ANOVA.

4.1 Pointing Task

Movement time. While both in-lab and out-of-lab participants finished the pointing task without issues, in-lab participants finished significantly faster: in-lab: \(M = 98s\) (SD = 14s), out-of-lab: \(M = 149s\) (SD = 62s), \(F(1,78) = 17.5, p < .001\). We plotted the linear fits of participants’ movement times as a function of ID (see Figure 2), which shows that Fitts’ Law (Shannon formulation) is indeed a very accurate model for pointing in VR (both in controlled and
uncontrolled settings). An ANCOVA shows a significant effect of ID and place on movement time, but no significant interaction $F(1, 316) = 0.72, ns$, hence the slopes, or the influence on movement time by ID is comparable for in-lab and out-of-lab VR pointing. The distribution of movement times between two targets were similar for in-lab and out-of-lab, both ranging between roughly 500 and 2000 milliseconds (see Figure 2). The distribution hints that participants in both studies did not encounter any difficulties nor took any noticeable breaks in the middle of trials.

**Table 3: Completion times for the 3D Tracing task, ±SD.**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Dynamic</th>
<th>Static</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-lab</td>
<td>240.9s ± 62.4</td>
<td>247.8s ± 79.0</td>
<td></td>
</tr>
<tr>
<td>Out-of-lab</td>
<td>292.4s ± 92.1</td>
<td>269.8s ± 91.2</td>
<td></td>
</tr>
</tbody>
</table>

**Correct Estimations.** Both the in-lab and out-lab study showed that participants in the dynamic condition estimated the origin of the leaves significantly better (see Table 4). In-lab participants increased estimations by 10.2%; Cohen’s $d = .35$; out-of-lab participants increased estimations by 9.1%; Cohen’s $d = .39$. The effect of the experimental condition was significant for both studies; for in-lab this effect was $F(1, 25) = 13.1, p < .01$, for out-of-lab it was $F(1, 51) = 16.0, p < .001$.

**Table 4: Correct estimations for the 3D Tracing task, ±SD.**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Dynamic</th>
<th>Static</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-lab</td>
<td>86.8% ± 8.4</td>
<td>76.6% ± 4.6</td>
<td>**</td>
</tr>
<tr>
<td>Out-of-lab</td>
<td>83.1% ± 10.6</td>
<td>74.0% ± 9.5</td>
<td>***</td>
</tr>
</tbody>
</table>

**In-lab vs. Out-of-lab.** The rates of correct answers to which tree the leaves originated were negligible between in-lab and out-of-lab. In-lab participants on average estimated 81.7% correct; out-of-lab 78.6%. This difference was not significant. Confidence intervals for the effect sizes for this task show that the effects due to experimental conditions are within the same range: out-of-lab, $d = 1.11, 95\%$ CIs [.45, 1.80]; in-lab, $d = 1.40, 95\%$ CIs [.40, 2.77].

**Summary.** Results from both the in-lab and out-lab study showed that participants increase their rate of correctly estimating the origin of a leaf in VR, given the ability to inspect trees spatially. The data show that the effect of the experimental condition was the same between in-lab and out-of-lab, and with comparable effect sizes. Also, in-lab and out-of-lab participants did not perform significantly different in terms of speed. The data therefore provide evidence for the validity of conducting VR studies that entail 3D navigation outside the laboratory.

### 4.3 Body Ownership Illusion Task

**Body Ownership.** Body ownership is different sensory cues unified into the perception of my body [Kilteni et al. 2015]. We asked four questions from [Banakou et al. 2013] to quantify body ownership (see Table 2): VRBody (Q1), Mirror (Q2), Features (Q3), and Two-Bodies (Q4). Figure 3 gives an overview of the responses. Based on previous work (e.g., [Banakou et al. 2013]), we expected the visuomotor consistency to increase the degree of body ownership.

For the in-lab study, all means but TwoBody were higher in the consistent condition; for the out-of-lab study all means but VRBody were higher in the consistent condition. Using a Wilcoxon rank-sum test on these questions across the consistent and inconsistent conditions, we found that only Mirror showed a significant difference, and only for the in-lab study, $Z = -1.92, p = .05$.

**Presence.** Presence is the sense of being there, distinguished from immersion, as being the participants’s response to the environment, and not relating the fidelity of technology used [Slater and Wilbur 1997]. We asked three questions originating from [Steed et al. 2016], relating to presence (see Table 2): BeingThere (Q6), VRExp (Q7),

### 4.2 3D Tracing Task

**Completion Time.** The manipulation did not have a significant effect on task completion times (see Table 3). There was not a significant difference on completion time between in-lab and out-of-lab either.
As evident from Figure 4, the differences between conditions for the two studies were found for the two conditions using a Wilcoxon rank-sum test. Steed et al. [Steed et al. 2016] did not find any differences in degrees of presence across conditions, except for synchronous tapping which had the opposite effect to the expected (presence decrease).

5.1 Participants

Twenty-two people, aged 20-40 (SD = 5.8 years) participated in this laboratory study, and were reimbursed with a gift equivalent of 158 US. None of the participants had previously participated in any of our studies. Seven of the participants were male. One participant was discarded due to the control question. The experimental design and apparatus followed the out-of-lab condition of the first study.

5.2 Pointing Task

While the first study showed that a simple pointing task can easily be completed by participants both in-lab and out-of-lab, significant differences in performances were reported. The results from the second study provide evidence that this performance gap is most likely due to fidelity of the employed VR technology. Figure 5 shows comparable performances between the out-of-lab and in-lab/low-tech conditions, with significant differences to the in-lab performances. Neither movement time, accuracy, or throughput varied significantly between the out-of-lab and in-lab/low-tech conditions. This shows that the differences observed in the first study for this task, are most likely due to the employed apparatus.

5.3 3D Tracing Task

In the first study we observed that the experimental condition had the same effect across study places: both participants in-lab and out-of-lab estimated origins of leaves better using a dynamic view. We did not observe significant differences in completion times. As Table 5 shows, the experimental manipulation did not cause changes to completion time in the in-lab/low-tech study, as with
Table 5: Completion times for the 3D Tracing task, ±SD.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Dynamic</th>
<th>Static</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-lab</td>
<td>240.9s ± 62.4</td>
<td>247.8s ± 79.0</td>
<td></td>
</tr>
<tr>
<td>Out-of-lab</td>
<td>292.4s ± 92.1</td>
<td>269.8s ± 91.2</td>
<td></td>
</tr>
<tr>
<td>In-lab/low-tech</td>
<td>277.0s ± 155s</td>
<td>276.0 ± 72.6</td>
<td></td>
</tr>
</tbody>
</table>

The experimental manipulation caused better estimations for both in-lab and out-of-lab. As evident from Table 6 this pattern was also true for the follow-up study; the dynamic condition caused significantly better estimations of origins, $F(1, 20) = 14.4, p < .001$.

Table 6: Correct estimations for the 3D Tracing task, ±SD.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Dynamic</th>
<th>Static</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-lab</td>
<td>86.8% ± 8.4</td>
<td>76.6% ± 4.6</td>
<td>**</td>
</tr>
<tr>
<td>Out-of-lab</td>
<td>83.1% ± 10.6</td>
<td>74.0% ± 9.5</td>
<td>***</td>
</tr>
<tr>
<td>In-lab/low-tech</td>
<td>87.0% ± 7.5</td>
<td>75.0% ± 7.0</td>
<td>***</td>
</tr>
</tbody>
</table>

The in-lab/low-tech condition did not significantly differ from the other conditions. Confidence intervals for the effect sizes for the in-lab/low-tech condition showed comparable to the first study: $d = 1.66, 95\% CI[61, 2.56]$, showing that the intervals for out-of-lab and in-lab/low-tech are contained in the in-lab interval.

5.4 Body Ownership Illusion Task

The third task had the most complex VR phenomena of the three. It studied if an avatar’s motor consistency with the participant varies the participant’s sense of body ownership and presence.

Figure 6: Box plots of questionnaire results on (a) body ownership ($Q1 - Q4$), and (b) on presence ($Q6 - Q8$). The results fit the laboratory condition from the first study.

Body Ownership. The mean of all body ownership scores were higher in the condition with visuomotor consistency (see Figure 6a), consistent with the first study. Exactly as with the in-lab condition for the first study, only Mirror ($Q2$) showed to significantly differ between the conditions, $Z = -2.45, p = .01$. This tells us, that even though we observe the same overall trends in body ownership throughout the studies, the two laboratory studies did cause more similar ownership responses.

Presence. As with the first study the observed differences in presence due to experimental manipulation are very negligible, making it difficult to make any inferences of the experimental variation on presence scores.

Summary of Follow-up Study. The follow-up study showed that the differences in simple performance metrics (e.g., accuracy, speed of pointing), between and the in-lab and out-of-lab conditions of the VR study were likely due to the hardware used. High-end VR equipment often deployed for in-lab studies caused faster interactions, with higher accuracy, compared to commodity VR technology.

Conversely, in the more delicate spectrum of dependent variables for the VR studies; the place of study seems to be of more concern than the fidelity of technology. This is indicated by our results, in the body ownership illusion task; results from the follow-up study, although by and large follow the same trend as both previous condition, match the in-lab better. That is, that the mean body ownership showed higher for the condition with visuomotor consistency for most questions, but the same question ($Q2$) showed statistically different by experimental manipulation in the in-lab and in-lab/low-tech conditions.

6 OTHER DATA

In addition to the tasks’ dependent variables we logged other measurements to get insights in the uncertain factors of conducting studies without a human evaluator. We here look at the physical surroundings and the differences in technology.

6.1 Setting

During the out-of-lab study, participants’ phones stored a photo using the back-facing camera, to provide insights into to the contexts in which participants carried out the study. We later printed all photos (see example photos in Figure 7), and categorized them. The categorization of the photos resulted in five non-exclusive groups:

- Place: where participant was during the study
- Locale: type of place (home, public or office)
- Barriers: near surroundings contained physical obstacles
- Activity: surroundings showed signs of co-occurring activity
- Social: other humans were present

Figure 7: Examples of out-of-lab study settings: outside (a), inside (b,c,d), home (a,b,c), office (d), standing (a,b), sitting (c,d). Photos printed with permission from the participants.

Note that thirteen of the photos did not contain much information, for instance when showing a close-up wall, and hence did not provide further insights other than it was taken inside. Additional, nine photos were indecipherable, for instance very blurry or dark.

6.2 Out-of-lab Technology

We analyzed the effect of two parameters relating to participants’ equipment: phone brand and screen size, to see if the technology used had an impact on the performance of the participants. We looked at all the dependent variables for each task (see Figure 1). Of the 53 participants, 31 participated with an iOS device and 22 with an Android device. We found no significant difference on any of the dependent variables: $F(1, 51) = [0.1, 2.7], p = [0.11, 0.92]$. We compared area of screen to the same parameters, and also here...
We have explored if it is feasible to conduct virtual reality (VR) user studies outside the laboratory. Potentially, this would give access to more varied physical and social settings, and higher participation, which in turn could give VR studies higher external validity at a lower cost. In particular, across three tasks we investigated if the performance parameters obtained (e.g., task completion times) compare to a laboratory condition and if the findings of experimental comparisons compared across in-lab and out-of-lab. For the two tasks containing experimental conditions, we did find significant effects (i.e., not null results). We found that similar effect sizes can be found using an easier and cheaper study method.

We believe that this gives a good first indication of how VR could be crowdsourced. Because VR equipment is currently not widely available we ran the out-of-lab experiments by giving cardboards to participants. We decided against mailing out cardboards because popular crowdsourcing platforms currently do not have the population required for an out-of-lab VR study, and we are not in a country with a significant cardboard population. Nonetheless, the results show that valid data can be acquired from VR studies without supervision, across a range of VR phenomena and complexity.

The literature contains numerous comparisons of in-lab and out-of-lab studies (e.g., [Buhrmester et al. 2011; Crump et al. 2013; Germine et al. 2012; Mason and Suri 2012; Paolacci and Chandler 2014; Ross et al. 2010; Rouse 2015]); to our knowledge, this paper is the first to do such comparison for VR studies. Steed et al. [Steed et al. 2016] provided a first exploration of whether VR studies could be conducted in-the-wild, in this paper we have explored whether the results of in-lab and out-of-lab studies are comparable and indeed whether out-of-lab is a valid methodology for VR.

Our results showed that the absolute differences in performance between the in-lab and out-of-lab study were substantial; participants in the laboratory study performed better in most of the absolute performance metrics (throughput, accuracy, completion time, and depth estimation). A follow-up study however revealed that this difference is likely due to technology used, and to a lesser extent due to the experimental design.

Data from all tasks confirms the feasibility of out-of-lab VR; there were no significant differences between effects of experimental conditions for tasks when comparing the in-lab and out-of-lab studies. We show that even complex VR phenomena entailing body ownership are possible to conduct out-of-lab with comparable results to in-lab studies, although the effects indicate that levels of body ownership were likely higher for the laboratory-based studies.

7.1 Recommendations
Although our setup is limited in a number of ways to be discussed, we can still provide a first set of recommendations on VR studies outside the laboratory.

- Pre-screen participants for the technology accessible to them to avoid recruiting unqualified people
- Expect roughly half of the participants to complete the study
- 15 minutes seems like the maximum tolerable duration for keeping the pose required to use the VR cardboard system
- Validate the integrity of participants, for instance using verifiable control questions, context photos, or user performance.
- Design experiments well-suited for both standing and sitting
- Expect simpler dependent variables (speed, accuracy, throughput) to vary with technology, but complex phenomena (body ownership, presence) to depend more on internal control

7.2 Open Questions
The results in this paper were obtained with Google cardboards. First of all, it raises the question of how to achieve the large-scale participation typically seen in out-of-lab research. We believe it is possible to mail cardboards directly to participants who have signed up online or possibly have them buy them and be reimbursed. The current rate of cardboard adoption (about 2%-3% of the crowdworkers surveyed) makes recruiting on crowdsourcing platforms infeasible for anything but small studies. Second, of course it raises the question what happens when more crowdworkers have high-end equipment (e.g., Oculus, HTC Vive). We do not see that as imminent but the differences in settings, movement of participants, and absolute performance values would be interesting to observe.

The current approach, mainly due to how cardboard VR glasses work sets several limitations on the task design. People must hold the same posture (binocular pose) during the entire study, and it is therefore infeasible to actively use the hands for anything, as you normally would in more advanced VR immersions. Additionally, the use of vibrotactile feedback (such as synchronous stimulation with a rod on virtual and physical body), as seen in many body ownership illusion studies (e.g., [Slater et al. 2006, 2009, 2010]), is infeasible. We foresee that these limitations could be resolved in the future, due to advances in commodity VR and wearable technology. This will also open up to longer studies, where fatigue will not be a factor in task design.

7.3 Conclusion
This paper shows that for VR studies concerning a heterogeneous population, out-of-lab experimentation is a worthwhile and valid methodology to consider. We compared VR tasks concerning pointing, 3D tracking and body ownership illusions, both as in-lab and out-of-lab studies. We showed that it is feasible to get reliable data by conducting VR user studies outside the laboratory, across a range of tasks and VR phenomena. This study is the first to validate VR experimentation outside the laboratory, and provides a first set of suggestions on how to crowdsource VR user studies.

ACKNOWLEDGMENTS
The work was supported by the European Research Council, grant no. 648785.
REFERENCES


