Towards Self-Guided Remote User Studies - Feasibility of Gesture Elicitation using Immersive Virtual Reality

Madhawa Perera[1], Tom Gedeon[2], Matt Adcock[3] and Armin Haller[4]

Abstract— **Gesture Elicitation Studies (GES) are a widely used empirical method to develop gesture vocabularies, interaction models and methods for gesture-based systems in different contexts. While GES show great promise to identify user-defined gestures, there are inherent problems with current methods used for GES. Especially during the ongoing pandemic, it has been nearly impossible to conduct in-person, in-lab GES, while ensuring the safety and well-being of the participants, and complying with social distancing regulations. Further, with prevailing experiment designs, increasing the number of participants is time consuming, while in-lab environments also limit ecological validity. This study explores an intuitive way of conducting self-guided GES using immersive Virtual Reality (VR), utilizing its capability to simulate various contexts to enhance ecological validity. We present a methodology and a tool set that use an immersive VR environment to conduct ecologically valid GES (as a use case) while requiring minimal involvement by the investigator. We evaluate our method using the case of a smart home environment and measure participant acceptance and discuss opportunities and challenges involved in this method. We believe that this study will help HCI research to move forward with participatory design research, even when lab experiments are difficult to conduct.**

I. INTRODUCTION

Originating from behavioral science, user studies have been an effective empirical research method that was instrumental in developing many interaction theories and models such as Fitts' Law. To date, user studies remain one of the most effective approaches in human-computer interaction (HCI) to make generalizable findings. One of the most popular areas of user studies is to study user defined system inputs. Especially with the proliferation of new human-computer interactions and interfaces, users became central, in research, for eliciting their inputs to analyse user defined interactions. This tendency is evident in the works of 'participatory design' by Schuler et al. [1]. Having emerged from participatory design, Gesture Elicitation Studies (GES) are a popular and commonly used user study method [2] to identify user preferred gestures to interact with a certain referent (desired effect of an action) [3].

As many authors such as Villarreal-Narvaez et al. [4] cite, GES were introduced by the work of Wobbrock et al. [5]. Using this method, in 2009, Wobbrock et al. [6] presented an approach for designing tabletop gestures by eliciting 1080 gestures from 20 non-technical users and found the users' agreement of gestures to given referents. Later Vatavu et al. [7] proposed a refinement for eliciting the best gestures based on agreement rate. Many researchers then adopted Wobbrock and Vatavu et al's work in GES to determine the best gestures when designing systems, applications, devices, automobiles, etc. The main objectives of GES are to collect a gesture set from the users and to understand user behavior [6] while investigating whether there are gesture agreements between users [4]. GES essentially contribute to the design of good gestures that possess ease-of-performance, memorability, discoverability or reliability [8], [3].

Despite its effectiveness, there are notable limitations in the methods of conducting GES which are common in investigator led, in-lab user studies in HCI. This was further highlighted with the restrictions imposed during the pandemic, making it challenging to follow existing methods to conduct user studies.

A. Challenges in conducting GES

There are different techniques that researchers follow to conduct GES (cf. Section II), each with their own challenges. Most GES are conducted with participants being present in-person in lab conditions. There are a few studies that have used web-based tools to interact with participants, such as Amazon Mechanical Turk (AMT) [9] and video conferencing tools, where the investigators use techniques such as Wizard of OZ (WoZ), speak out loud etc. to create the presence of a referent and to guide the users. However, recruiting participants for such methods is difficult and there are inherent issues with tool such as AMT. Madapana et al. [9] have shown that AMT workers do not pay careful attention to user study instructions due to their urge to finish a task and start another one to receive the financial incentives. Hence the quality of the collected data is low. Unfortunately, during the COVID-19 pandemic universities and research labs being closed to non-critical experiments, made in-person GES difficult to conduct. Thus, the HCI community is keen to address this challenge by looking at various methods of conducting user studies (in general) with remote participants.

In addition to pandemic restrictions, in-lab GES take considerable amounts of time, both of the investigator and the participants, as experiments are conducted in series (not parallel). Further, the 'Hawthorne effect' [10] documented nearly 60 years ago that participants may behave differently in lab experiments due to the stress of being observed or the rewards offered for participation. Thus, in-lab experiments may not extract a user's typical behaviour is another concern. Also creating ecologically valid physical setups and maintaining them until the completion of the study is time consuming and challenging. Due to these challenges, GES typically secure a small number of participants. As per Koeman's [11]

*This work is supported by Immersive Environments Lab, CSIRO

[1,2,3,4] College of Engineering and Computer Science, Australian National University (firstname.lastname@anu.edu.au), [1,3] CSIRO (firstname.lastname@csiro.au)

review, lab-based research is still common, yet, the majority of these studies have fewer than 20 participants [12]. The inherent lack of diversity of the results and a low statistical power makes generalization of their findings difficult [13]. On practical difficulties, recorded videos of participants with fixed camera angles, that are generally collected in these studies, reduce the flexibility of looking at a gesture in other angles when such examination is required. Furthermore, data visualization and privacy challenges underlie the whole process; visualizing gestures requires a significant effort from researchers and experiment recordings often do not preserve the privacy of the participants, even if they do not wish to be identified.

To mitigate these challenges, we conducted an experimental GES using a fully immersive Virtual Reality (VR) environment to study user behaviours, and gestures in particular. In this paper we describe this self-guided fully immersive Human Device Gesture Interaction (HDGI) study and its design and implementation along with the participants' experience and acceptance of using VR for GES. As highlighted by Steed et al. [14], currently there are no easy-to-use tools to run VR experiments and there are various technical issues with implementing and distributing experiments to consumer devices. Therefore, we believe the present methodology and the prototype we used to collect and visualise user defined gestures in this paper will trigger further research on using immersive VR to conduct remote GES (and user studies in general) by mitigating technical barriers.

II. Related Work

A. Existing methodologies to conduct GES

Methods to conduct GES can be differentiated based on the ways they present referents to participants, and the recording mechanisms of gestures. Vogiatzidakis et al. [15] discuss commonly utilised approaches to present referents to a participant including text on screens or verbal feedback, video presentation or still images and manipulating the actual artefact. Manipulating actual artefacts is the best approach, yet can be challenging in many scenarios. The other methods have lesser ecological validity as participant may behave differently as they realise that they are not interacting with the real device. As for recording gestures, motion capture [16] and questionnaires are sometimes used. Video recording is a common mechanism to record elicited gestures [15], [4].

Looking at GES techniques, the WoZ method [17] which was initially used in natural language interfaces, has become a popular approach to identify user defined gestures. Wobbrock et al.'s [6] initial work has adopted this technique. In this approach, participants were shown the effect of their gesture by an unseen technical wizard (human) manipulating the system. Participants were unaware of the human operator, thus WoZ design makes the participant feel that their interactions are actively being recognized, allowing them to use it as if they are interacting with a real gesture recognition system. This method allows rapid prototyping and has been widely used by researchers to conduct GES [18], [19], [20], [21]. It has been identified that the illusion of system autonomy is of paramount importance to the results in a WoZ design. Henschke's [22] experiments using WoZ, reveal that this illusion will be dispelled if the user becomes aware that the system is not operating autonomously.

Another major criticism against this method is that it always requires the investigator/experimenter (wizard) to be present during the experiment period and thus the experiments are conducted sequentially, unless there are many experimenters and equally set up lab spaces. There are other techniques such as the think-aloud protocol [23] which is used alongside WoZ to overcome some of its pitfalls but it still requires conducting experiments in series. Further, Schieben et al. [24] has used another approach named 'theater system'. This technique again extends WoZ by making the wizard/investigator play a confederate who is no longer hidden. The confederate can play through different use cases with the participants as if they would play a role in a theater [24]. This techniques is used by Mahr et al. [25] and May et al. [26] in elicitation studies conducted to investigate driving related scenarios but it still requires the presence of the investigator. Commonly, these approaches require in-person setups and experiment conduct is chronological.

B. Using immersive VR to conduct user studies

One of the objectives of immersive environments is to create the perception of 'being present' in a different environment or context by immersing the user in a computer generated setting [27]. They have long been used for user studies, especially when creating an ecologically valid environment for the experiment. For example, VR has been used in investigating driving experiences [28], [29]. Weidner et al.'s [30] comparison of VR and non-VR Driving Simulations show that data gathered from VR simulators is similar to stereoscopic 3D or 2D screens and they did not observe significant differences regarding physiological responses or lane change performance. Holzner et al. [31] indicate that a VR approach is a very cost effective way for testing a smart home environment together with a Brain Computer Interface (BCI) system. Further, Spoladore et al. [32] introduce a VR smart home simulator to address a variety of issues involved in the development of Ambient Assisted Living (AAL) solutions. Bates et al.'s [33] work proposed an action recognition and learning system in which researchers can collect examples of human behaviour using a VR application, which overcomes the difficulties associated with capturing performances in physical environments.

In earlier studies Brooks et al. [34] examined the potential for using VR in memory rehabilitation. This is an indication of researchers attempting to utilise VR environments to bring ecological validity to their user studies. Kourtesis et al. [35] have developed VR-EAL (Virtual Reality Everyday Assessment Lab) to conduct an ecologically valid examination of everyday prospective memory. In comparison to the traditional method, immersive VR has shown a higher validity [36]. Huygelier et al.'s [37] study on acceptance of immersive head-mounted virtual reality in older adults strengthens its ability to reach diverse demographics for user studies. Borrego et al.'s [38] study evinces that immersive environments could enable the navigation and exploration of real-life sized virtual environments, without any notable adverse effects. Radiah et al. [39] and Ratcliffe et al.'s [40] review of challenges and opportunities on remote VR studies show that the pandemic has increased the use of VR by consumers, and that users are open to new uses of VR. Mottelson et al. [13] have validated VR experimentation outside a lab environment and have shown that it is feasible to get reliable data. Thus, VR provides a promising direction to help continue HCI research and participatory design user studies.

III. Self-Guided, Remote Gesture Elicitation

The state of the art shows the potential as well as interest in the HCI community to conduct remote user studies using immersive VR. Our methodology for conducting GES addresses the challenges with current GES stated in section I-A while providing added benefits such as self guidance for users in the experiment setup, ability to conduct GES in parallel with multiple participants, preserving privacy of the user as we do not collect any video of real user actions or their voice that would potentially make them personally identifiable. Further, the method allows us to conduct elicitation studies in an ecologically valid setup compared to a stereotypical lab setup and allows the investigator to visualize gestures at any required angle.

We selected a smart home environment as it is difficult to create such a setup in a lab environment as it may require several smart devices and complicated physical setup to create ecological validity for the experiment. In our GES, we collected user defined gestures to control a selected set of smart apparatus in a room space. The environment setup is easily customizable as required by researchers. Finally we evaluated the participant experience and acceptance for using VR for GES.

A. Apparatus

We used the Unity game engine to develop the VR application. Our application is capable of running on both Oculus Quest 1 and 2 VR HMDs. We selected Oculus Quest as it used to conduct industry employee training (e.g., Walmart [41]) and currently leads the VR HMD sales rankings [42]. Additionally Oculus was the first to introduce in-built hand tracking on consumer grade VR HMDs and has the highest share of Steam (online game platform) users with a VR headset worldwide as of February 2021 [43].

B. Participants Recruitment

As we are conducting the GES remotely, we aim to collect a diverse set of participants' data from different demographics. Hence, participants joined this experiment in two different ways. The first option was, if a participant owned a VR HMD, they could visit the HDGI study instruction page and follow the instructions and complete the user study whenever they want (users with VR devices could download and install it onto their devices and conduct the experiment). The second option was that participants could visit our lab at a time allocated via a booking system and use one of the available VR HMD to conduct the experiment on their own, without the requirement of the investigator being present in the lab. Multiple participants could do the study in parallel based on the available HMDs. There were participants with previous VR experience and without. We conducted the study in a short period of time (three weeks) within which we collected gestures from 53 participants, who originated from 15 countries (27 females, 23 males, 1 non-binary and 2 participants who did not disclose their gender). Participants were distributed from age brackets 21 - 25 years being the first and 61 - 65 years being the last; highest of 26.4% of participants were in the age group of 26 - 30 year bracket. Since the VR consumer market is rapidly growing [14], we now recognize that there is a potential to reach out to many participants through this method.

C. VR application

The VR application was developed by considering reusability, better participant experience, and ease of use for the investigator after collection of the data. The application is extensible to different contexts and usages for other researchers. The app is modularized and includes three main modules: the environment setting module, the gesture recording module and the gesture visualization module. The modularization creates reusability as other HCI researchers could utilize these modules in their own VR development with Unity, by customizing the application to fit their research purposes. Additionally, the app was built as a standalone application allowing direct download, instead of publishing on readily available VR app stores. The VR application file was hosted on a web server for participants to download and to execute on their own.

In contrast to a majority of approaches taken by researchers to conduct VR remote studies, we used a mix of written (textual) information and voice. As our design aimed for self-guidance, unlike in Radiah's et al. [39] method of displaying instruction to participants, we included a virtual voice assistant supplementing the available textual instructions. During the trial and at the end of the experiment, this virtual agent guides the users through the experiment.

In our VR app, unlike in the majority of studies described in our related work section, we could not use the VR controllers as the main purpose of this study was to collect user-defined hand gestures. Thus, we forced participants to use their hands by making hand interaction mandatory for application loading. Participants can only enter the app once they rest their controller. Otherwise, the application will not start. We used Oculus Quest's in-built hand tracking capabilities to record users' performed gestures. Gesture recording and users' head movement data were recorded in JSON files. Recording frequency was customizeable and set during app assembly time. Recorded JSON files included rotation data (in quternions) of hand bones and user's head position (as a 3D vector). A sample gesture data file can be seen here[1].

[1] https://madhawap.github.io/vr-ges/json_viewer.html

Once the experiment is completed, these gesture data files were pushed to a secured server. We indicated to users that even when we are collecting the virtual hand and head data, we only do so for 5 seconds for each affordance. Further, we did not collect any video footage of users' hands or their physical space, only the hand motion data.

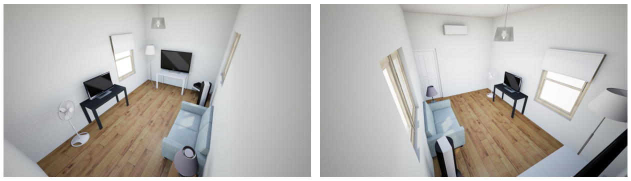


Fig. 1: VR Room setup

D. Experiment Design

The experiment is divided into five stages: a user information guide, a pre-questionnaire, a self-guided experiment trial, the actual experiment, and a post experiment questionnaire. Each participant goes through each of these stages on their own, and in this order, to complete the full study. The sequence of stages were planned as per a typical HCI user study design in Lazar et al's. [44] guide on designing HCI experiments.

During the study we evaluate the participant's comfort with the presented features of our VR application. This would help to investigate what further alterations are required in the VR application design for GES purposes. To make the experiment self-guided, we followed the independent study method proposed by Radiah et al. [39]; that is, the study was conducted asynchronously, where participants ran the study on their own at any time they wanted. No experimenter was present during the study, whether it was completed in their home or in our lab.

Firstly, participants were asked to go through the information guide to get an overall understanding of the study and then to install and setup the VR application. Once a participant successfully installed the application, they were asked to answer a pre-questionnaire which covered basic demographic data and their previous experience with Augmented Reality (AR)/VR HMDs. Once completed, they entered the VR application.

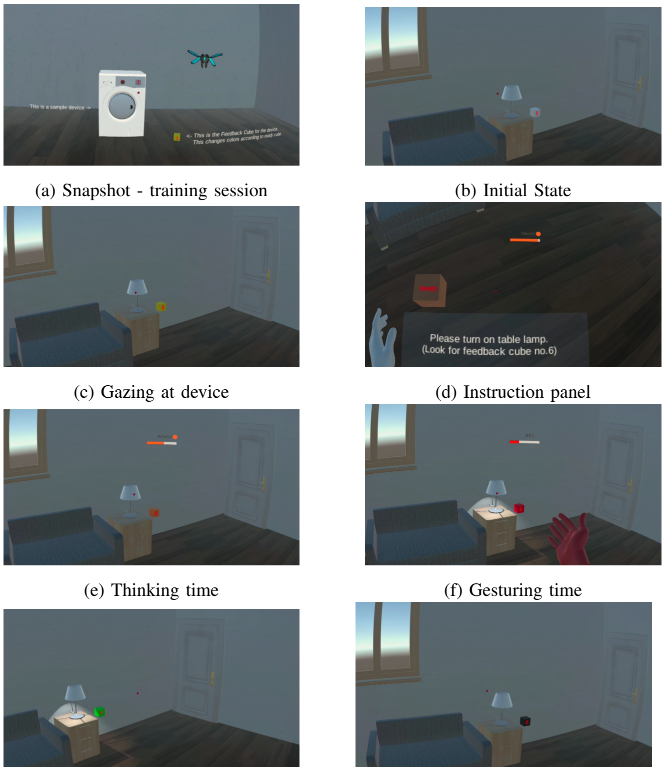
We followed a within-group design as GES tasks could have large individual differences. However, to reduce the impact of the learning effect [44] we balanced the order of the tasks that participants performed and provided sufficient time for training. Therefore, we let users go through a very short self-phased and self-guided trial setup to learn and practice before they enter the real elicitation study. Research has shown that providing sufficient training time for users to get acquainted with the study can greatly reduce the learning effect during the actual task sessions. In the trial scene, participants were guided to interact with an affordance which was not part of the main study. This step is equivalent to a participant being briefed by an investigator, and it also provided them with a chance to perform a trial and familiarize with the setup as suggested by Lazar et al. [44].

The experiment can be conducted in any venue that has at least a 1m x 1m space. During the actual experiment each participant interacts with 42 affordances spread across 12 smart objects (cf. Table I) in a smart home environment. Each participant interacts with all 42 affordances presented to them in an order balanced fashion. Compared to elicitation studies in the literature, our study contains a higher number of affordances, yet keeps the study well below recommended HCI study times. It is generally suggested that the appropriate length of a single experiment session should be 60-90 minutes or shorter [45]. In real life situations, smart home users will interact with multiple devices instead of a single device. Thus we assumed this will enhance the ecological validity of the elicitation study and allow to observe if users perform gestures differently in such settings. We were aware of the problem of fatigue that could be caused by multiple experimental tasks. In order to address this, we designed experiment tasks economically and made the participant aware that they are spending 15 seconds on each affordance, only.

Figure 1 shows the room setup with the device placements. When a participant is asked to interact with a certain device affordance, they are given a timed 15 seconds, which is split into 10 seconds to think and prepare and 5 seconds to perform the gesture. Since Vatavu et al's [46] study shows that there is a negative correlation between agreement rates for a gesture and thinking time, we decided to keep the thinking time less than the average time given in Vatavu et al.'s [46] study, which is 20.5 seconds. Further, to address another challenge in gesture elicitation studies which is to identify when the gesture started and when the gesture ended and whether it is a gesture that a user performed, we used the following steps. When participants were given an affordance to interact with, the thinking time and gesture performance times were visualized with a colour-changing progress bar. Additionally, when it is time for the participant to perform a gesture, their virtual hands change color to indicate that the system is recording the gesture at that time. See Figure 2e and 2f for the thinking time and recording time visual feedback mechanisms. This process continues until a participant completes all of the affordances.

Devices	**Affordances**
1. Bladeless Fan	Turn on | Increase Temperature | Decrease Temperature | Enable swing mode | Turn off
2. Chandelier	Turn on | Turn off
3. Double Window	Open the window | Close the window
4. Floor Lamp	Turn on | Turn off
5. Hung Window	Open the window | Close the window
6. Inverted Air conditioner	Turn on | Increase Temperature by 1 unit | Decrease Temperature by 5 units | Turn off
7. LED TV (Model 1)	Turn on | Go to next channel | Go to previous channel | Turn off
8. LED TV (Model 2)	Turn on | Go to next channel | Go to previous channel | Turn off
9.Security Camera	Turn on | Turn Right | Turn Left | Turn off
10. Speaker System	Turn on | Increase Volume | Decrease Volume | Play the next track | Increase volume by two units | Turn off
11. Table Lamp	Turn on | Turn off
12. Ventilator	Turn on | Increase speed | Decrease speed | Turn off

TABLE I: Affordances List


This is a sample device ->

<- This is the Feedback Cube for the device. This changes colors according to ready cube.

Please turn on table lamp. (Look for feedback cube no.6)

(a) Snapshot - training session

(b) Initial State

(c) Gazing at device

(d) Instruction panel

(e) Thinking time

(f) Gesturing time

(g) Device feedback

(h) End of interactions

Fig. 2: A few snapshots from our VR application

Figure 2b to 2h show a sequence of interactions with a device (table lamp). Every device in the scene was labeled with a numbered cube next to the actual device. This cube was named 'feedback cube' and it was introduced during the trial to participants. Feedback cube colour changes provided visual feedback to participants about which interaction stage they were at for that device. Further the instruction panel displayed the feedback cube number so that participants could find the device that carried the given affordance (Figure 2d). Once the participant completed a gesture, the virtual device activated the relevant affordance, instead of any textual or voice feedback (like in many GES). For example, the table lamp turned on, the speaker system played a music track or changed the track, or the window opened etc. after the gesture was performed.

Once a participant completes the study, the recorded gesture data is pushed to a remote server. While the system finalises this task, a virtual agent appears and reminds the participant to complete the post questionnaire. Our pre- and post-questionnaires were designed to understand the aforementioned critical design aspects in a self-guided and remote GES using VR. We further measured participants' satisfaction with how realistically they felt that the virtual hand mimics their real hand movements, experiment completion time, clarity of the instruction, and preferred mode of instruction on performing the study (voice, written or both). Further we asked users to rate the helpfulness of the visual feedback and the device feedback mechanisms we embedded in the study along with their satisfaction with given thinking and gesture performance time. Finally, we asked users to rate the likelihood of them choosing a typical GES versus self-guided and remote VR GES. We used a 5-points Likert scales, closed questions, 1-5 ordinal scales and ranking questions in our pre- and post-questionnaire to obtain users' subjective experience.

E. *Gesture visualization*

Once the data is collected and pushed to a remote server, investigators can use these data files to visualize the user performed gestures for each device affordance. We developed another prototype to rebuild the participants' gestures (3D hand animation) for each affordance separately to view the gesture from any desired angle. This helped us to better classify and describe gestures. Additionally, the collected gesture data was a numeric representation of a human hand's bones. Thus, investigators could run a classification model with these data files if they wish to automate the classification process, which would ease the gesture labelling and the agreement score calculation process. As an additional benefit, the tool helps to capture snapshots of user gestures which makes it easier when presenting and describing the gesture vocabularies.

IV. Results

For the evaluation, we considered all participants who completed all five stages of the experiment from installation to post-questionnaire, i.e., N = 53. All the statistical analyses were conducted using the R programming language in RStudio.

Firstly, we tested whether there was a difference between VR application completion time with users' who have or have not had previous experience with VR. Out of all participants, 41.5% had not used VR HMD before and the rest had at least one or more previous experiences using VR HMD. Since each participant went through the study only once, the samples were independent. As assessed by Shapiro-Wilk normality test and quantile-quantile (q-q) plot on both participant samples; experienced (p-value = 0.20) and non-experienced (p-value = 0.23) participants, we assumed that the distribution is normal. From the F-test (df = 21, denom df = 30, p-value = 0.70) comparison variances of two participant samples, we assumed that variances were equal. Therefore, with the null hypothesis of there being no difference in VR app completion time between experienced and non-experienced participants, we conducted a two-tailed, two-sample t-test. The 22 participants who did not have previous VR experience (mean = 30.441) compared to the 31 participants with previous VR experience (mean = 29.429) demonstrated no statistically significant difference in the experiment completion time (t = 1.77, df = 51, p-value = 0.08). Therefore, this method for self-guided and remote GES can be conducted with both experienced and non-experienced participants.

Further, we investigated participants' choice between remote VR GES and in-lab experiments. According to post-questionnaire data, 83.0% participants preferred self-guided remote VR GES when they were asked to select their most preferred option to participate in GES out of the two options:

in-lab experiment versus VR GES. With an intent to further analyze this decision, we asked participants to provide the likelihood [on a 5-point scale from 1 (Highly unlikely) to 5 (Highly likely 5)] in participating in each of these two types of studies. Overall, the results show a neutral response mean of 2.72/5 for the in-lab experiment, while the self-guided and remote VR GES shows a highly likely rating with a mean of 4.94/5. Then we conducted a Pearson's Chi-squared test of independence to determine if there is a relationship between participants' likelihood for choosing one of these studies with their previous experience in VR. The test results (X-squared = 2.17, df = 2, p-value = 0.34) showed that there was no significant association between these two variables, which indicates that previous VR experience and likelihood of selecting self-guided remote GES are independent from each other. Therefore, participants, regardless of previous experience in VR, have shown a much higher likelihood of participating in self-guided remote VR GES than in an in-lab experiment.

We investigated the user ratings of likelihood for selecting in-lab experiments as well. Here the Pearson's Chi-squared test of independence results (X-squared = 13.41, df = 4, p-value < 0.05 (0.00944)) shows that there is a significant relationship between the two variables; likelihood of selecting in-lab and previous VR experience. Therefore, we conducted a non-parametric Spearman's rank correlation for further analysis. The results show a correlation coefficient of -0.485 with a significant p-value < 0.05 (0.000236) which means non-experienced participants showed a higher likelihood in participating in an in-lab experiment, while experienced VR participants show a lower likelihood. Since the VR consumer market is rapidly growing [14], we recognize that the number of experienced VR users will increase and thus there will be an increased likelihood of participants preferring self-guided and remote GES via immersive VR.

However, among participants' given reasons for preferring in-lab experiments as well, 81.8% of non-experienced participants indicated that the standalone installation method was difficult to execute. This is further illustrated with an overall rating mean of 2.72/5 [from a scale from 1 (Very Dissatisfied) to 5 (Very Satisfied)] which shows some dissatisfaction among all non-experienced users. Overall, users' subjective evaluation of the installation experience is also rated at 3.34/5 on the same rating scale. We conducted unpaired two-samples Wilcoxon test (also known as Wilcoxon rank sum test or Mann-Whitney test) to determine whether satisfaction in non-experienced participants was different to experienced participant satisfaction when it comes to the installation setup, which was 3.77/5. We selected the non-parametric test as the non-experienced participants' satisfaction rating was not normally distributed, as assessed by Shapiro-Wilk's test (p = 0.0002643). The p-value of the test is 0.000052, which is less than the significance level alpha = 0.05. Thus we conclude that the average satisfaction rate of the non-experienced participants was highly significantly different from the experienced participant satisfaction rating. Nevertheless, out of all experienced VR users, less than half (only 41.9%) had experience in installing 3rd party VR applications. The likert scale analysis conducted to analyse the overall participant experience on application setup, as shown in Figure 3 with diverging stacked bar charts indicating that more than half of the participants (Strongly Agree - 49.0% and Agree - 17.0%) prefer to have the application to be downloadable from an app-store. Otherwise, for the standalone installation, a majority (58.5%) of participants preferred to have both video and written instructions to be presented in the information sheet.

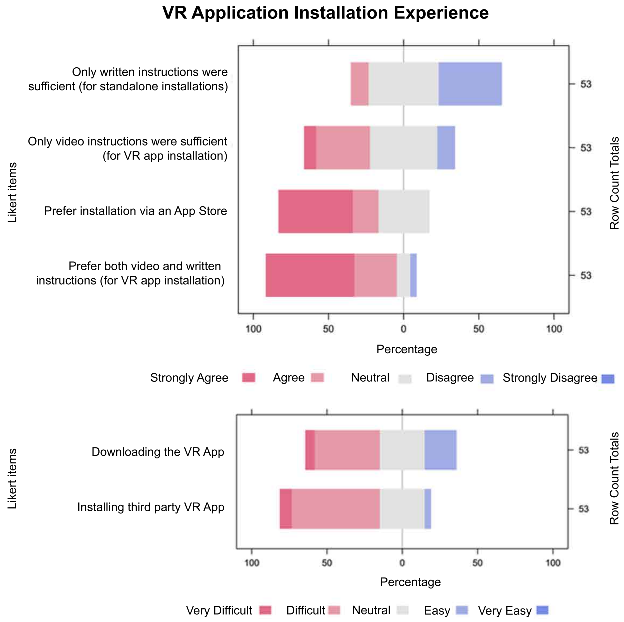


Fig. 3: Evaluation of standalone installation experience

Next, we evaluated the users' experience with the virtual hands. Firstly, we investigated both experienced and non-experienced participants' favorable or unfavorable perception of the virtual hand representation of their real hands with a 5-points likert scale from 1 (Very dissatisfied) to 5 (Very satisfied). The overall mean of 4.81/5 indicated a very satisfied experience by all participants. All participants reported that either always or often they were able to perform the gesture they wanted and the virtual hand correctly mapped their real hand movement. Figure 4 shows the summary of the user evaluation of the virtual hand performance. Two separate one-sample Wilcoxon signed rank tests were run to determine whether the overall participant perception score was significantly different between experienced and non-experienced participant perceptions. As assessed by Shapiro-Wilk's test ($p < .05$) each participant type's rating data cannot be assumed to be normally distributed, hence we conducted this non-parametric alternative to the one-sample t-test. The p-value of the non-experienced participants test is 0.126, which is greater than the significance level alpha = 0.05; similarly with experienced participants p-value = 0.109. Thus we conclude both non-experienced and experienced participants found the virtual hand performance equally very satisfying.

Finally, we investigated the users' perceptions of the fea-

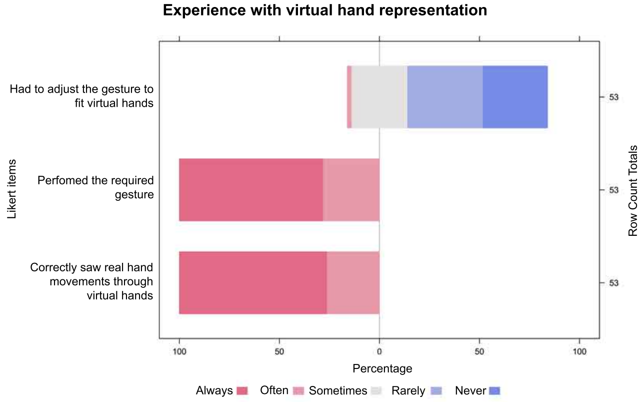


Fig. 4: Evaluation of virtual hand experience

tures that were added in the VR application aiming for self-guidance and remote conduct of GES. We evaluated users perception on the time given to think of a gesture (thinking time) and to perform a gesture (recording time), colour change of the virtual hand to indicate recording of the gesture, and the feedback mechanism we provided during the study. As shown in Figure 5, 90.6% of participants found the gesture thinking time to be sufficient, either always or often, whereas 86.8% participants reported that visualizing the thinking time was either very helpful or extremely helpful. However, 9.4% participants reported that the thinking time was sufficient only sometimes because sometimes they had to find where the device was in the virtual environment by looking at the given instructions. This has taken a few seconds from their thinking time. The feedback cube mechanism to mitigate this problem was rated 92.5% by participants as either very or extremely helpful. Recording time was rated as always sufficient by 88.7% participants while the rest mentioned that time is often sufficient. Finally, 81.1% found it very or extremely helpful to have a visualization to indicate the recording time. Therefore, the overall perception of users on the allocated time and visualization technique we incorporated was useful when designing self-guided GES studies.

Further, 96.2% of the participants had rated the virtual hand colour change during the recording time as an extremely helpful feature and that they have noticed the change always. While 94.3% recognised pre-training was extremely helpful, all participants found that the voice assistant was either very or extremely helpful. In all instances, at the expiration of the recording time, participants see the relevant feedback (an actuation or observation) from the device for the affordance. This method was rated 96.2% by participants as highly satisfying and 92.5% of the participants always recognised that the device responded to their gesture commands. Therefore, we conclude that these specific features we have added towards building self-guided and remote GES as very useful.

V. Discussion and Open Questions

As shown in section IV, there was a considerable number of users who had not installed third party VR applications before and preferred direct download from an app store. Even though uploading the application to an app store may increase

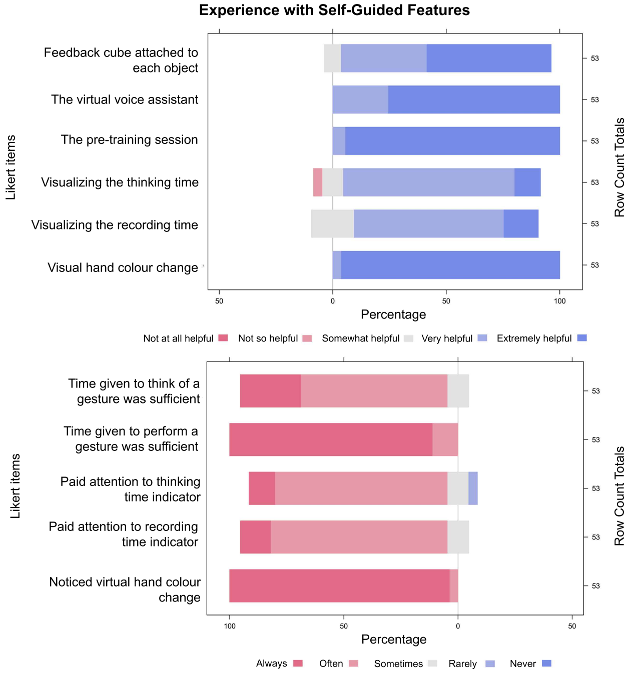


Fig. 5: Evaluation of user experience different features

the reach further and improve the installation experience of users, this may lead to unanticipated privacy concerns such as third parties' ability to track the usage of the app, hence this would require further investigation.

We developed this application, with re-usbility and customizability in mind and share the code as open source project[2] such that HCI researchers could use the app as is or a customized version to conduct remote gesture elicitation studies. Additionally, there is an opportunity to add more modules by expanding the kinds of data that can be collected (e.g., user gaze pointer data, interaction logs and facility to embed research or consumer psychological signals measuring devices). Thus, researchers could not only use these tools to conduct GES but also for user behaviour analysis studies with remote users.

A challenge that we identified with GES was when investigators had to classify the elicited gestures, describe them and calculate the agreement scores during the analysis phase. Literature shows that recorded gestures are most often labelled by a human. Usually this gives elicited gestures a subjective description, which could result in duplicate descriptions for similar gestures making the comparison hard. In order to overcome this challenge the HDGI ontology [47] could be embeded for the gesture labelling process, such that the investigator can describe a hand gesture with a pre-defined set of labels and semantics. This could result in commonly accepted gesture labeling, in turn resulting in comparable qualitative gesture labels with their relevant contexts.

VI. Conclusion

We identified and elaborated on challenges that current GES methods face, and introduced a novel way of conducting remote and self-guided GES using immersive VR for

[2] https://madhawap.github.io/vr-ges/

identifying users' preferred gestures. This method especially addresses a lack of user participation, has the potential for an increased demographic representation, addressing the issue of not having ecologically valid setups in gesture elicitation studies, all while keeping the participant identities anonymous. We discussed the design and the development of our VR application, along with a sample study to evaluate the feasibility of conducting such studies in practice. We chose the context of smart homes and let participants interact with 42 affordances across 12 devices that were presented in a balanced order. We collected 2,226 gestures within three weeks from 53 participants without having to spend any time in the lab with participants. Our method shows high participation satisfaction levels from both experienced and non-experienced VR users. The majority indicated that they would perform the same gesture set in a similar real-world setup. Further, we presented participants' acceptance of the study and the important aspects to consider when conducting remote and self-guided GES. We discussed possible future directions for improving our method by integrating gesture extraction, visualization and bringing standardised gesture labelling into GES to improve the validity of elicited gesture vocabularies while making the elicitation process even more efficient compared to traditional methods. This study helped us to continue our work during the pandemic, thus, we share the reusable prototypes we developed with the HCI community to continue studies even in times when in-lab experiments are not possible.

References

[1] D. Schuler and A. Namioka, *Participatory design: Principles and practices*. CRC Press, 1993.

[2] C. Stephanidis, G. Salvendy, M. Antona, J. Y. Chen, J. Dong, V. G. Duffy, X. Fang, C. Fidopiastis, G. Fragomeni, L. P. Fu *et al.*, "Seven hci grand challenges," *International Journal of Human–Computer Interaction*, vol. 35, no. 14, 2019.

[3] M. R. Morris, A. Danielescu, S. Drucker, D. Fisher, B. Lee, M. Schraefel, and J. O. Wobbrock, "Reducing legacy bias in gesture elicitation studies," *interactions*, vol. 21, no. 3, 2014.

[4] S. Villarreal-Narvaez, J. Vanderdonckt, R.-D. Vatavu, and J. O. Wobbrock, "A systematic review of gesture elicitation studies: What can we learn from 216 studies?" in *ACM DIS'20*, 2020.

[5] J. O. Wobbrock, H. H. Aung, B. Rothrock, and B. A. Myers, "Maximizing the guessability of symbolic input," in *CHI EA '05*, 2005.

[6] J. O. Wobbrock, M. R. Morris, and A. D. Wilson, "User-defined gestures for surface computing," in *SIGCHI'09*, 2009.

[7] R.-D. Vatavu. and J. O. Wobbrock, "Formalizing agreement analysis for elicitation studies: new measures, significance test, and toolkit," in *SIGCHI'15*, 2015.

[8] M. Nielsen, M. Störring, T. B. Moeslund, and E. Granum, "A procedure for developing intuitive and ergonomic gesture interfaces for hci," in *International gesture workshop*. Springer, 2003.

[9] N. Madapana and J. Wachs, "Database of gesture attributes: Zero shot learning for gesture recognition," in *IEEE FG*, 2019.

[10] H. A. Landsberger, "Hawthorne revisited: Management and the worker, its critics, and developments in human relations in industry." 1958.

[11] L. Koeman, "Hci/ux research: What methods do we use?" Jun 2020. [Online]. Available: https://lisakoeman.nl/blog/hci-ux-research-what-methods-do-we-use/

[12] L. Koeman., "How many participants do researchers recruit? a look at 678 ux/hci studies," May 2020. [Online]. Available: https://lisakoeman.nl/blog/how-many-participants-do-researchers-recruit-a-look-at-678-ux-hci-studies/

[13] A. Mottelson and K. Hornbæk, "Virtual reality studies outside the laboratory," in *ACM VRST '17*, 2017.

[14] A. Steed, F. R. Ortega, A. S. Williams, E. Kruijff, W. Stuerzlinger, A. U. Batmaz, A. S. Won, E. S. Rosenberg, A. L. Simeone, and A. Hayes, "Evaluating immersive experiences during covid-19 and beyond," *Interactions*, vol. 27, no. 4, 2020.

[15] P. Vogiatzidakis and P. Koutsabasis, "Gesture elicitation studies for mid-air interaction: A review," *Multimodal Technologies and Interaction*, vol. 2, no. 4, p. 65, 2018.

[16] G. A. Rovelo Ruiz, D. Vanacken, K. Luyten, F. Abad, and E. Camahort, "Multi-viewer gesture-based interaction for omni-directional video," in *SIGCHI'14*, 2014, pp. 4077–4086.

[17] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of oz studies—why and how," *Knowledge-based systems*, vol. 6, no. 4, 1993.

[18] M. Henschke, T. Gedeon, and R. Jones, "Touchless gestural interaction with wizard-of-oz: Analysing user behaviour," in *OzCHI '15*, 2015.

[19] S. Connell, P.-Y. Kuo, L. Liu, and A. M. Piper, "A wizard-of-oz elicitation study examining child-defined gestures with a whole-body interface," in *IDC '13*, 2013.

[20] A. S. Williams, J. Garcia, F. De Zayas, F. Hernandez, J. Sharp, and F. R. Ortega, "The cost of production in elicitation studies and the legacy bias-consensus trade off," *Multimodal Technologies and Interaction*, vol. 4, no. 4, 2020.

[21] D. Akers, "Wizard of oz for participatory design: inventing a gestural interface for 3d selection of neural pathway estimates," in *CHI EA '06*, 2006.

[22] M. Henschke, "User behaviour with unguided touchless gestural interfaces," Ph.D. dissertation, ANU (Australia), 2020.

[23] R. Jääskeläinen, "Think-aloud protocol," *Handbook of translation studies*, vol. 1, 2010.

[24] A. Schieben, M. Heesen, J. Schindler, J. Kelsch, and F. Flemisch, "The theater-system technique: Agile designing and testing of system behavior and interaction, applied to highly automated vehicles," in *AutomotiveUI '09*, 2009.

[25] A. Mahr, C. Endres, C. Müller, and T. Schneeberger, "Determining human-centered parameters of ergonomic micro-gesture interaction for drivers using the theater approach," in *AutomotiveUI '11*, 2011.

[26] K. R. May, T. M. Gable, and B. N. Walker, "Designing an in-vehicle air gesture set using elicitation methods," in *AutomotiveUI '17*, 2017.

[27] O. Nonnarit, N. Ratchatanantakit, S. Tangnimitchok, F. Ortega, A. Barreto *et al.*, "Hand tracking interface for virtual reality interaction based on marg sensors," in *IEEE VR*, 2019.

[28] J. Cremer, J. Kearney, and Y. Papelis, "Driving simulation: challenges for vr technology," *IEEE Computer Graphics and Applications*, vol. 16, no. 5, 1996.

[29] M. T. Schultheis, J. Rebimbas, R. Mourant, and S. R. Millis, "Examining the usability of a virtual reality driving simulator," *Assistive Technology*, vol. 19, no. 1, 2007.

[30] F. Weidner, A. Hoesch, S. Poeschl, and W. Broll, "Comparing vr and non-vr driving simulations: An experimental user study," in *IEEE VR*, 2017.

[31] C. Holzner, C. Guger, G. Edlinger, C. Gronegress, and M. Slater, "Virtual smart home controlled by thoughts," in *WETICE '09*, 2009.

[32] D. Spoladore, S. Arlati, and M. Sacco, "Semantic and virtual reality-enhanced configuration of domestic environments: the smart home simulator," *Mobile Information Systems*, vol. 2017, 2017.

[33] T. Bates, K. Ramirez-Amaro, T. Inamura, and G. Cheng, "On-line simultaneous learning and recognition of everyday activities from virtual reality performances," in *IEEE/RSJ IROS*, 2017.

[34] B. Brooks and F. Rose, "The use of virtual reality in memory rehabilitation: current findings and future directions," *NeuroRehabilitation*, vol. 18, no. 2, 2003.

[35] P. Kourtesis, S. Collina, L. A. Doumas, and S. E. MacPherson, "An ecologically valid examination of event-based and time-based prospective memory using immersive virtual reality: the effects of delay and task type on everyday prospective memory," *arXiv preprint arXiv:2102.10448*, 2021.

[36] P. Kourtesis., S. Collina, L. A. Doumas, and S. E. MacPherson, "Validation of the virtual reality everyday assessment lab (vr-eal): an immersive virtual reality neuropsychological battery with enhanced ecological validity," *Journal of the International Neuropsychological Society*, vol. 27, no. 2, 2021.

[37] H. Huygelier, B. Schraepen, R. van Ee, V. V. Abeele, and C. R. Gillebert, "Acceptance of immersive head-mounted virtual reality in older adults," *Scientific reports*, vol. 9, no. 1, 2019.

[38] A. Borrego, J. Latorre, M. Alcañiz, and R. Llorens, "Comparison of oculus rift and htc vive: feasibility for virtual reality-based exploration, navigation, exergaming, and rehabilitation," *Games for health journal*, vol. 7, no. 3, 2018.

[39] R. Rivu, V. Mäkelä, S. Prange, S. D. Rodriguez, R. Piening, Y. Zhou, K. Köhle, K. Pfeuffer, Y. Abdelrahman, M. Hoppe *et al.*, "Remote vr studies–a framework for running virtual reality studies remotely via participant-owned hmds," *arXiv preprint arXiv:2102.11207*, 2021.

[40] J. Ratcliffe, F. Soave, N. Bryan-Kinns, L. Tokarchuk, and I. Farkhatdinov, "Extended reality (xr) remote research: a survey of drawbacks and opportunities," *arXiv preprint arXiv:2101.08046*, 2021.

[41] T. O. Team, "Walmart expands vr training with oculus go," Oct 2018. [Online]. Available: https://www.oculus.com/blog/walmart-expands-vr-training-with-oculus-go/

[42] T. Alsop, "Vr headset unit sales by device worldwide 2020," Feb 2021. [Online]. Available: https://www.statista.com/statistics/987701/vr-unit-sales-brand/

[43] T. Alsop., "Steam user vr headset share worldwide by device 2021," Apr 2021. [Online]. Available: https://www.statista.com/statistics/265018/proportion-of-directx-versions-on-the-platform-steam/

[44] J. Lazar, J. H. Feng, and H. Hochheiser, *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[45] J. Nielsen, "Time budgets for usability sessions," *Useit. com: Jakob Nielsen's web site*, vol. 12, 2005.

[46] R.-D. Vatavu and I.-A. Zaiti, "Leap gestures for tv: insights from an elicitation study," in *ACM TVX '14*, 2014.

[47] M. Perera, A. Haller, S. J. R. Méndez, and M. Adcock, "Hdgi: A human device gesture interaction ontology for the internet of things," in *ISWC*. Springer, 2020.