# Using the Think Aloud Protocol in an Immersive Virtual Reality Evaluation of a Virtual Twin

Xuesong Zhang
xuesong.zhang@kuleuven.be
KU Leuven
Leuven, Belgium

Adalberto L. Simeone
adalberto.simeone@kuleuven.be
KU Leuven
Leuven, Belgium

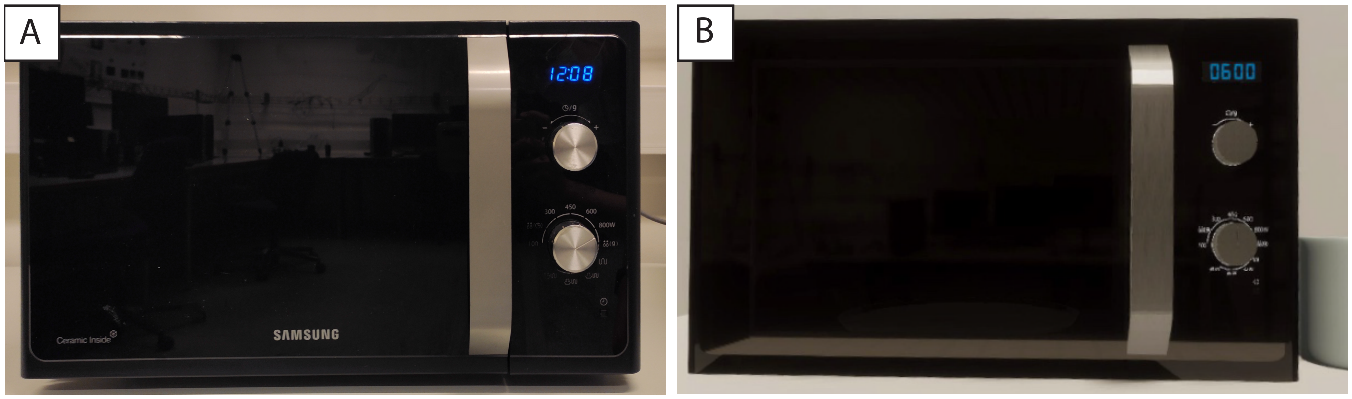

**Figure 1: The real Microwave (A) and its virtual twin rendered via Physically Based Rendering (B) are shown above. We used them for our user study based on the Think Aloud Protocol to compare the results of the usability inspection with a physical prototype and its virtual twin.**

## ABSTRACT

Employing virtual prototypes and immersive Virtual Reality (VR) in usability evaluation can save time and speed up the iteration process during the design process. However, it is still unclear whether we can use conventional usability evaluation methods in VR and obtain results comparable to performing the evaluation on a physical prototype. Hence, we conducted a user study with 24 participants, where we compared the results obtained by using the Think Aloud Protocol to inspect an everyday product and its virtual twin. Results show that more than 60% of the reported usability problems were shared by both the physical and virtual prototype, and the in-depth qualitative analysis further highlights the potential of immersive VR evaluations. We report on the lessons we learned for designing and implementing virtual prototypes in immersive VR evaluations.

## CCS CONCEPTS

• **Human-centered computing → User studies**; **Virtual reality**.

## KEYWORDS

VR, Usability Evaluation, Think Aloud Protocol, Virtual Twin

**ACM Reference Format:**
Xuesong Zhang and Adalberto L. Simeone. 2022. Using the Think Aloud Protocol in an Immersive Virtual Reality Evaluation of a Virtual Twin. In *Symposium on Spatial User Interaction (SUI '22), December 1–2, 2022, Online, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3565970.3567706

## 1 INTRODUCTION

The use of virtual reality (VR) as a medium in which to evaluate the usability of prototypes is gaining increased research attention [17, 23, 28]. We refer to this type of usability studies carried out in VR as "*Immersive Virtual Reality Evaluations*" (IVREs).

IVREs provide various advantages: 1) the potential of obtaining results with VR evaluations transferable to the real world [18]; 2) staging field studies in VR that could be difficult to replicate in real life [17]; 3) the capability to identify potential problems at an early stage, prior to building a physical prototype [23]; 4) the ability to simulate the interaction and appearance of physical devices at a lower cost, compared to building a physical prototype [18, 20].

In this paper, we designed a user study to investigate the effectiveness of performing usability evaluations in VR. We hypothesize that a significant share of the usability problems that will be uncovered in this manner, would also present themselves if the virtual prototype were to be built physically. Hence, in our study, we used the *Think Aloud Protocol* (TAP) to evaluate the usability of both a real-world appliance, a microwave oven, and its "virtual twin" (a virtual object that replicates the appearance and interactive affordances of the physical counterpart as close as possible).

The effectiveness of usability evaluation methods (UEMs) in identifying usability issues varies depending on the context, such as the product type [11], the evaluator [8], or the medium [3]. The 2010 study by Bruno and Muzzupappa is the closest to our work [6]. The authors compared the results of the usability evaluation

*SUI '22, December 1–2, 2022, Online, CA, USA*

of a real microwave oven with those resulting from the evaluation of its three-dimensional twin experienced via a semi-immersive, stereoscopic, projected screen without head-tracking, and with a fixed perspective. Users could only observe the front operation panel. Different from our study, the TAP was not applied: experimenters observed the participants who were interacting with the two microwave ovens. The development of HMD and input devices in recent times motivates a renewed interest in performing usability evaluations in VR [27].

Results indicate that more than 60% of the usability issues present on the virtual twin were also identified on the physical appliance. When evaluating both devices with the TAP method, participants reported similar amount of identified usability problems in terms of type and severity. Standard questionnaires provided similar scores in both settings as well.

The contribution of this paper is two-fold: (1) We report the usability problems identified with TAP and a qualitative in-depth analysis of whether they affected both devices or solely one of the two. (2) We discuss which factors may affect the identification of usability problems in IVREs, and report on lessons learned.

## 2 RELATED WORK

In this section, we present user studies employing VR to stage user studies and the UEM we focused.

### 2.1 User studies in Virtual Reality

In 2019, Voit et al. conducted a user study with 60 participants comparing the usability of smart artifacts inspected in VR, lab setup, online, augmented reality and in-situ with standard usability questionnaires [28]. They observed similar ratings in terms of the usability, attractiveness, qualitative feedback quality, pragmatic and hedonic quality in VR and in-situ settings. However, hand-object interaction in VR is mainly simulated through animation, which is different from the real world. Mäkelä et al. investigated the feasibility of using VR as test-bed to perform virtual field studies on public display [17]. The user behavior observed in the virtual environment (VE) was largely similar to that exhibited in the real-world environment. In 2021, Mathis et al. replicated a real-world authentication system in VR and evaluated its usability and security [18]. Compared to the real world setting, participants interacted with the virtual prototype with a similar entry accuracy and perceived similar workload. Paneva et al. simulated a levitation interface in VR and conducted two user studies [20]. Results show this virtual prototype offered similar performance, user engagement level as well as user experience compared to the real physical prototype. The authors stated the highly realistic interactions allow for good predictions of work performance and user experience. In 2022, Simeone et al. introduced the concept of "Immersive Speculative Enactments", where the usability of non-existing or unfeasible devices can be evaluated in VR [23]. Given the speculative nature, their work lacks a comparative evaluation of the extent to which usability issues overlap between the physical and virtual medium, which we report in this paper.

### 2.2 Think Aloud Protocol

In the following, we introduce the UEM we mainly focused on in this work: *Think Aloud Protocol (TAP)*. TAP asks participants to verbalize their thoughts while performing specific tasks [10]. The information collected provides an account of which usability problems were experienced and indications as to the source of these issues. A previous study [2] shows that the concurrent TAP method (users provide a report while interacting with the object of the evaluation [10]) detected more usability problems than the retrospective TAP (users provide a report after finishing interacting with the object of the evaluation [10]). No significant differences were found between concurrent and hybrid method (the combination of the those two types [12]) in terms of number of detected usability problems. The concurrent TAP needed the shortest amount of time in terms of execution and analysis among these two variants. For these reasons, we chose the concurrent TAP as the usability evaluation method to use in our user study.

## 3 USER STUDY: PHYSICAL VS. VIRTUAL PROTOTYPES

We designed a between-subjects user study, aiming to compare the results from a usability evaluation of a microwave oven under two different settings: one in a real-world lab, while the other is performed on its virtual twin with the evaluator being immersed in a VE. Both evaluations were performed using the *Think Aloud Protocol (TAP)*, combined with standard questionnaires i.e., *System Usability Scale (SUS)* [5], *Post-Study System Usability Questionnaire (PSSUQ)* [15], *NASA Task Load Index (NASA-TLX)* [14].

The independent variable in this study was the Study Modality with the two conditions of {Real Environment (RE), Virtual Environment (VE)}.

In the study, participants were asked to inspect a microwave oven and acted as the evaluators. A set of tasks were assigned to activate a desired function (such as defrosting food based on the time or the weight, the combination of grill and microwave function).

As a common kitchen appliance, microwave ovens share functions whose implementation varies between manufacturers. Although participants all shared to some extent a basic understanding of how a microwave oven works, it is expected that the difference in implementing the user interface across manufacturers could lead to various usability problems [13]. Further, we hypothesized that a share of these problems would have manifested themselves in the virtual twin as well. This study aims to investigate the extent of this overlap, and the severity of the identified problems.

### 3.1 Apparatus and Implementation

We created a virtual twin of a real microwave oven appliance (Figure 1-A) produced by Samsung (MG23F301E). This microwave oven was released in 2014[1] and has been superseded by an improved model released in 2021. It is fully functional and owned by one of the authors. One special feature of this microwave oven is that if there is no further change after setting the function or timer, the microwave oven will automatically start cooking within two

[1] Manual: https://www.manua.ls/samsung/mg23f301eas/manual

seconds. We choose to evaluate a microwave oven in part to link to previous work [6] and to have a link to an existing commercial product to use as "ground truth".

The virtual twin (Figure 1-B) was modeled in Blender[2]. It has the same dimensions as the physical microwave oven and simulates all the functionalities of the physical one. It also plays sounds and updates the information on the screen in the same way the real oven does when certain button combinations are pressed. Additionally, animations were created to mimic the defrost/heat/microwave/grill process inside the microwave oven. The interactive features were implemented in *Unity 2020.3.3*[3], and the rendering is done with the *High Definition Rendering Pipeline (HDRP)*[4]. Participants interacted with the virtual twin through a wired *HTC Vive Pro* HMD. Due to the insufficient reliability of the embedded hand-based detection [22], interaction with the dials and buttons was implemented via collision-based selection with a Vive controller. This differs from the use of a data glove and a joystick in the work of Bruno and Muzzupappa [6].

In the VE, the controller appears as a virtual hand with a small cube aligned to the index finger as a reference. The addition of this reference cube was necessary because during pilot testing, users without VR experience noted that it was difficult to determine whether the hand actually touched the button. To activate the desired function, the reference cube should collide with the corresponding component while pressing the trigger on the controller. If the cube collides with a dial, participants should press and hold the trigger while turning their wrist to rotate the dial.

The group that interacted with the physical microwave oven did so in our lab, where the oven had been temporarily placed. A cup filled with water is put inside the microwave oven to prevent it from running empty. No food or other drinks were actually heated during the experiment.

## 3.2 Demographics

We recruited 24 volunteer participants (13 male, 11 female) between the ages of 23 and 32 ($MEAN$ = 26.96, $SD$ = 2.63) for this lab-based user study. 12 participants were randomly assigned to each group.

They were recruited through internal mailing lists, word-of-mouth and social media. The user study was approved by the Ethical Review Board of our institution.

## 3.3 Procedure

After filling a consent and a demographics form, we introduced participants to the *TAP* evaluation method and the corresponding procedure. They were asked to sit in front of the (virtual) microwave oven to perform the evaluation which consisted of eight tasks in randomized order (see Table 1). The tasks required participants to press certain buttons and rotate the dials to defrost/heat/microwave/grill food with a specified power for a certain duration, with the purpose of prompting participants to pay attention to the icons, operate all the buttons and knobs, and experience all the functions of the microwave oven. We prepared an additional cup and asked participants to treat it as the "food" during the task. The operating instructions came from the manual of the microwave oven and were briefed to the participants.

To familiarize them with the VR interactivity, they underwent a training session where they could interact with the microwave's door, control knobs, and buttons, as well as grab and place a mug.

**Table 1: Task List for each microwave.**

| | |
|---|---|
| *Task 1* | Defrost food for 3 minutes |
| *Task 2* | Set clock to 15:34 |
| *Task 3* | Microwave 30 seconds on 600 W |
| *Task 4* | Keep the food warm for 1 minute 30 seconds |
| *Task 5* | Grill for 2 minutes |
| *Task 6* | Heat 4 minutes 30 seconds with the high microwave and grill function |
| *Task 7* | Heat 10 minutes with the low microwave and grill function |
| *Task 8* | Defrost 500 g food |

After participants confirmed they understood the purpose of the evaluation, the experimenter gave them a signal to start. Following the TAP, participants described their actions and thoughts while performing the tasks. They could either complete or abandon it after three unsuccessful attempts. There were 192 trials performed (8 $Tasks \times 24\ Participants$), three tasks were abandoned by two participants. We recorded the RE sessions with a smartphone camera, and VE sessions with OBS[5] to record the first-person view.

After evaluating the microwave, participants were asked to fill in four web-based questionnaires: SUS, PSSUQ, NASA-TLX, and a custom questionnaire with ten questions in 5-point scale (where 1 is strongly disagree and 5 is strongly agree): two of which aimed to understand participants' opinion on whether performing the TAP affected their task performance; five of which targeting at eliciting their view on the TAP method; the rest are intended to understand the impact of the experimenter's presence on the study. Next, participants proceeded to freely explore the microwave oven in either RE or VE without being required to complete any task. After this exploration phase, participants needed to fill in another custom questionnaire with three questions in 5-point scale, which asked participants to compare the physical and virtual models in terms of similarity of their perceived appearance and operation (1: completely inconsistent; 5: completely consistent). In addition, we also asked them to predict whether performing the task with the virtual prototype would take more time than with the physical prototype for both settings (1: strongly disagree; 5: strongly agree). At the end, we conducted semi-structured interviews to let participants walk us through their feedback on the use of the TAP in RE and in the VE. Each evaluation lasted about 60 minutes.

# 4 RESULT

In the following, we report the quantitative data collected during the study and the usability issues we detected.

[2] Blender: https://www.blender.org/
[3] Unity: https://unity3d.com/unity/whats-new/2020.3.3
[4] HDRP: https://docs.unity3d.com/Packages/com.unity.render-pipelines.high-definition@11.0/manual/index.html
[5] OBS: https://obsproject.com/

**Table 2: Allocation of the number of reported usability problems between VE and RE in terms of problem type and severity.**

| | Overlap | | | | | RE | | | | | VE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *SUM* | *H* | *M* | *L* | *E* | *SUM* | *H* | *M* | *L* | *E* | *SUM* | *H* | *M* | *L* | *E* |
| *C1* | 9 | 4 | 3 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 1 |
| *C2* | 5 | 1 | 2 | 1 | 1 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| *C3* | 5 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *C4* | 9 | 1 | 2 | 6 | 0 | 2 | 0 | 0 | 1 | 1 | 11 | 0 | 1 | 9 | 1 |

## 4.1 Task Completion Times

Task completion times (TCTs) were recorded from the moment when the experimenter gave the signal to start and until the participant communicated they were finished with the task. Participants were free to abandon the task as specified in subsection 3.3. In total, there was one participant who abandoned two tasks in the RE, while one task was abandoned by one participant in the VE.

Due to the non-normal distribution, we used Kruskal-Wallis H tests to determine if there were differences in terms of TCTs between the data measured in the VE condition and those measured in the RE. Overall, participants took more time in the VE condition (*MEAN* = 405.98, *SD* = 88.88) to complete all eight tasks than in the RE (*MEAN* = 336.12, *SD* = 184.16). Participants performed tasks 1, 4, 5 significantly quicker in the RE than in the VE (*T1:* $p = 0.04$; *T4:* $p = 0.024$; *T5:* $p = 0.035$).

## 4.2 Questionnaires

For data collected from SUS, PSSUQ, NASA-TLX we applied the Shapiro-Wilk test to check the normality of the distribution. Unless otherwise specified, it is assumed that data is normally distributed. No significant difference was detected in terms of the SUS, PSSUQ, NASA-TLX scores attribute to the microwave across the virtual and real conditions with one-way ANOVA tests.

Participants experienced no significant difference when performing the TAP in both settings according to results with Kruskal-Wallis H tests. They agreed to the subjective statement that the virtual twin is identical in terms of appearance and function to the physical microwave oven.

## 4.3 Detected usability problem from the TAP

### Distribution of the detected problems

We recorded the entire evaluation process and transcribed the participants' dialogues and then followed a two-stage extraction process to identify the usability problems, leading from individual problems to final problems, as proposed by Alhadreti and Mayhew [2]. After this process a total of 46 distinct usability problems were identified; of these, 28 overlapping problems were detected in both settings. Five problems were only found by the participants in the RE, and 13 problems are unique to the VE. (see Figure 2).

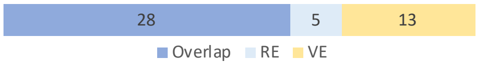

**Figure 2: Distribution of detected usability problems**

We compared the number of detected problems by performing a Kruskal-Wallis H-test on the data, as in Alhadreti and Mayhew's work [2]. There was no significant difference in terms of Study Modality ($p = 0.975$): the number of reported usability problems was comparable across both conditions (VE: *MEAN*=8.91, *SD*=4.30; RE: *MEAN*=8.92, *SD*=4.21).

### Categorization of the detected usability problems

We grouped them into four categories according to the motivations behind their occurrences (from the users' perspective). Their distribution is shown in Table 2.

**C1** *Misoperation of the appliance due to misunderstanding the process.*

The setting-start process of the functions is not always the same in this microwave. However, if the user does not fully understand or remember the process correctly, they might then not know what the next step is. Hence, the user might press the wrong button or get stuck in the process. For example, if there is no other operation by the user after two seconds after the last button press, the microwave oven will automatically start running. During the user study, five participants pressed the clock button after setting the function and timer, and the microwave started running at the same time by accident. They assumed the clock button represents the *"start"* function and pressed it again for the next task. However, the button did not work as expected, because pressing it is only used to enter the time setting mode.

**C2** *Misoperation due to not being able to find the desired button/dial/functions.*

Participants know what the next step is but are unable to find the desired function or button. For example, participants need to set the clock to 15:34 in Task 2. They need to change the minute digits after setting the hour digits. However, seven participants did not know which button they should have pressed to change the mode from hours to minutes.

**C3** *Confusion caused by similar functions.*

The appliance provides two or more similar functions under a different menu and the participant could not distinguish them. Hence, the participant chose the wrong function and cannot then reach their goal. For example, there are two defrosting functions based on either the time or the weight, respectively.

**C4** *Confusion caused by the text, icon, position, shape of button/knob.*

The icon and text on the device surface is ambiguous, users might misunderstand the function intended by the designers. The description of the text or the button shape misleads users to operate them incorrectly. For example, participants pressed the knob, which can only be rotated.

**Distribution of the detected problem in terms of the severity** According to the problem's impact on the performance (Task Completion Time), each problem is assigned with one of four severities [1, 2, 9, 29]:

- **H** *Critical, the usability problem prevented the completion of a task*;
- **M** *Major, the usability problem caused significant delay or frustration*;
- **L** *Low, the usability problem had minor effect on usability, several seconds of delay and slight frustration*;
- **E** *Enhancement, participants made suggestions or indicated a preference, but the issue did not cause impact on performance.*

Their distribution is shown in Table 2.

In the following, we report the usability problems (UPs) that were reported at least more than once, which were identified in both settings, or only identified in the RE or the VE.

**Overlapping problems in both the RE and VE**

- **UP1** *Lack of a Start button. (C2, L, RE: 6 times; VE: 8 times)*
- **UP2** *Lack of a Stop button. (C2, H, RE: 3 times; VE: 3 times)*
- **UP3** *Participants are confused by similar icons (two defrost functions, three grill-microwave combination functions). (C4, L, RE: 4 times; VE: 7 times)*
- **UP4** *Participants are confused by similar defrost functions. (C3, H, RE: 2 times; VE: 2 times)*
- **UP5** *The user misselected another function adjacent to the position of the target function. (C4, M, RE: 2 times; VE: 3 times)*
- **UP6** *Activating an empty microwave. (C1, H, RE: 1 time; VE: 2 times)*
- **UP7** *The knob is not sensitive to small angle rotation. (C4, L, RE: 4 times; VE: 5 times)*

**Unique problems in the VE**

- **UP8** *Participants perceived the image as blurred. (C4, L, 9 times)*
- **UP9** *The knob rotation is not intuitive and slow. (C4, L, 3 times)*
- **UP10** *The knob rotation is tiresome. (C4, L, 3 times)*
- **UP11** *Participants cannot open the door. (C4, L, 2 times)*
- **UP12** *Participants tried to start the microwave by pressing the knob. (C4, E, 4 times)*

**Unique problems in the RE**

- **UP13** *Confusion on how to set the timer. (C2, M, 10 times)*
- **UP14** *Attempting to start the microwave by pressing the clock button. (C4, E, 8 times)*
- **UP15** *The microwave door is hard to open and close. (C4, L, 2 times)*
- **UP16** *Using an incorrect knob to adjust the minute setting. (C4, L, 2 time)*

## 5 DISCUSSION

In this section, we discuss the usability issues that were found as a result of both the IVRE and the conventional lab-based evaluation, the difference of applying the TAP in the RE and VE.

### 5.1 Participants' Task Performance

Three out of the eight tasks took significantly longer in the VE than in the real-world setting. These three tasks required participants to select a function and then set a certain time interval from 90 seconds to 3 minutes. Participants were asked to continuously rotate the knob, and within this time range, the timer increased by 10 seconds for every 30° rotation movement, which required participants to precisely control the knob. When the time interval was not within this range, there was no significant difference in TCTs. Thus, this difference was solely attributable to the interaction technique used to rotate the knob in the VE. There is no evidence to support the notion that participants' task performance is affected whether or not the TAP is performed immersively.

### 5.2 Experience with the TAP across the RE and VE

Results show that participants' experience of the TAP were similar in both settings. During the user study, when the participants stopped to report for more than ten seconds, the experimenter guided the TAP process by giving the participant essential instructions (e.g., *"could you describe your current action?"*). The interruption from the experimenter is often associated with "Breaks-in-Presence" [26] experienced by the participant. However, in the case of the TAP, the evaluator is asked explicitly to describe their thinking and actions from the start. Thus the connection between the VE and the real-world always exists.

However, participants reacted differently to the presence of the experimenter when conducting the TAP. Some of them, for instance *P6*, were surprised by the "Cross-Reality"[7, 21] co-presence: *"It's strange, I know there is another person in the room, but I can't see them."* Conversely, *P10* had the opposite reaction: *"I felt extra comfortable with the presence of the experimenter, because I felt that they knew the next steps very well."* Some participants did not notice the experimenter: *"I did not even notice the experimenter during the VR session. The headset blocks out the physical environment."* (*P4*). Indeed, the experimenter did not have an avatar in the VE, because they did not interact with the user directly neither in the RE nor in the VE conditions.

The experimenter had a different experience in both environments. Unlike conventional lab-based user studies, it is difficult in IVREs to observe both the interaction inside the VE and the user's behavior in the real world at the same time. Due to the physical space required for interaction, the display which monitors the view of the VR user cannot be placed within close range. Choosing between observing the movement of the user in the real world or their interaction in VE would provide the experimenter with limited information about what is happening. Thus, the additional verbal information from the TAP is beneficial for the experimenter to understand the intention and the behavior of the user while performing an IVRE.

### 5.3 Overlapping problems found in both the RE and VE

When evaluating the microwave with TAP, more than 60% of the usability problems of the microwave were identified in both the RE and VE conditions, including 89% of the problems categorized as critical, 82% as major, 48% as low and 20% as enhancements.

In both conditions, participants exhibited similar behavior. We followed up the study with a semi-structured interview where we inquired about their experience with this microwave model, and there were only three participants who had prior experience with operating this same microwave oven. The high number of overlapping problems is in line with findings by Bruno and Muzzupappa [6], where participants experienced similar difficulties in understanding the microwave features in both the real-life lab and its virtual twin settings. Based on the verbal information from the TAP, participants took similar steps to complete the task, which means they had the same understanding of the workflow. The usability issues of critical severity caused by design flaws (such as missing button, inconsistent workflow) were also detected in the VE.

Interestingly, we also noticed that participants across both environments exhibited different behavior when they encountered the same usability problem. For example, in **UP7**: when users turn the upper knob less than 30°, the time/weight information on the screen will not change. We believe this represents a feature *working as intended* to prevent misoperation. In the evaluation process, participants acted differently to this design in both settings: *P17 (RE)* was confused and commented *"I think it is not the right knob"*, since they were certain that the rotation action did occur. While *P3, P9* (VE) continued to try to rotate the knob with the controller and commented *"It's hard to rotate in VR."* In the RE, participants receive different haptic sensations when they touch the buttons on the microwave or grab the cup, and the haptic feedback varies depending on the material. However, there is no haptic feedback after the controller collides with different virtual objects in the VE. An additional vibration feedback could help user to confirm the collision.

## 5.4 Unique problems found in the VE

Through the IVRE, participants reported 13 unique problems that were not found in the inspection of the physical appliance. The most commonly reported usability problems (UP*n*) in the IVRE were of type *C4 (confusing text or icons)*: eleven such problems were issues related to the "physical" interface of the virtual twin (see Table 2).

Nine problems (*severity: M: 1; L: 6; E: 2*) were real usability problems which were not detected in the RE. Those problems (*C1: 2; C4: 7*) were only reported once except **UP12**. We believe the detection of these problems may depend largely on the participants' own experiences. The other four usability issues (i.e. **UP8**, **UP9**, **UP10**, **UP11**) were caused by the implementation or hardware limitations.

There is no negative consequence when interacting with the digital twin (as commented by *P3, P9, P11, P12*), which prompts participants to be more open to exploration in the VE [17]. During the evaluation, *P9* tried to press the dial on the virtual twin, because it has a smooth appearance and has a height of 1 cm (**UP12**). Participants behaved intuitively and more directly. In this case, the IVRE helped detect usability problems related to the ambiguous shape of the button, which could have been beneficial in the early design stage.

The results also indicate that we need to pay attention to the hardware used in the VE. Problems related to blurred images in **UP8** constituted false positives, as it was not the case in the RE. This issue is attributable to the resolution of the HMDs used. To fully replicate the physical microwave oven, the icons on the virtual twin are the same size as the physical ones. The resolution of the headset led to difficulties in interpreting the information as intended by the designers, which affected the user experience. We anticipate that as VR headset technology matures, this will become less of a problem for IVREs in the near future.

During the IVRE, all tasks required users to turn the knob for the setting process. *P7, P8* reported that the knob turning action in the VE did not match with their experience. People typically use their fingers to turn knobs, however, wrist rotation is necessary when using a controller in the VE. The interaction with the virtual twin did not reproduce the natural interaction style that is possible in the RE, and was reported as slower than expected. Introducing a haptic proxy for the most common interactable controls could mitigate the occurrence of this problem [19].

The knob rotation was found to be tiring because the interaction is performed in mid-air without arm support in a non-ergonomic position. This is similar to the gorilla arm syndrome [4]. We expect that in the near future, hand-based interaction metaphors will alleviate this problem, and reduce the effect of fatigue resulting from holding a controller with a non-negligible weight. Alternatively, using smaller form-factor controllers could provide an interim alternative, as the Vive wands weigh 307 g compared to the 137 g weight of the Meta Quest 2 controllers.

We also observed two participants who encountered problems because of controller misoperation. In **UP11**, when participants forgot to press the trigger, the system did not detect the collision and the virtual door did not turn to follow the user's hand movements. We then inquired and found that those two participants had no prior experience with manipulating objects in VR. Since our training session lasted for two minutes, a longer session with activities to complete in order to progress could reduce these problems.

## 5.5 Unique problems found in the RE

When evaluating the microwave with the TAP, five problems were only detected in the RE. These problems are largely due to the design not being aligned with participants' experience (e.g., **UP16**).

Thanks to a wider field of view (FoV) as well as the higher resolution of human eyes, the additional visual information from the RE helps uncovering FoV and rendering quality related problems. Certain components of the microwave oven were accidentally ignored in the VE, whereas it did not happen in the RE. For example, the button which sets the clock on the lower right corner of the control panel was hardly noticeable during the VE tasks. According to the participants, the main reasons were the limited FoV and the low-quality of the rendering. In contrast, the presence of this button in the RE leads to misoperation, i.e., **UP14**.

Evaluation in the RE also allows users to interact with the prototype and get multisensory feedback from the physical device itself, such as temperature, force, roughness. Users reported the difficulty in operating the door, as more strength than expected was needed, i.e., **UP15**. Without having a physical proxy or weight simulation in the VE, this kind of usability problems, which are related to force feedback, could only be identified in the RE.

## 5.6 Lessons learned for IVREs

Based on the above results and qualitative feedback, we think that performing an IVRE can represent an efficient method to uncover usability problems in VR, and use the insights gained to further refine the design, before finalizing it into a physical prototype. In our setup, the IVRE allowed us to find 89% of the usability problems categorized as critical, 82% as major, 48% as low and 20% as enhancement, that were eventually identified in the RE as well. In addition, in VR, the other 9% major, 48% low and 40% enhancements of the usability issues were identified exclusively via the VE.

We recommend product designers, researchers, and other stakeholders to consider the following lessons we learned when performing an IVRE.

- *Implement natural interaction techniques that approximate as closely as possible the way the product will be interacted with in the real world.*

As our results suggest, differences in the interaction modality will be likely flagged as usability problems (i.e. **UP9**, **UP10**). Due to technical limitations of VR, it might be necessary to interact in a way that is different from its real-world analogue. In line with previous findings from Voit et al. [28], these are attributable to the VR interaction techniques, rather than the physical device itself. Evaluators should thus identify and categorize these problems accordingly and reflect on the likelihood of these interactivity issues affecting a physical prototype.

- *Use haptic proxies to uncover related problems.*

A problem that was uniquely identified in the RE (**UP15**) was not identified via the IVRE due to the lack of a physical proxy. Due to the positive effects of incorporating haptic feedback in VR experiences on the believability of the experience [6, 19, 25] to further enhance the fidelity of the interaction and uncover related problems in VR, future work should explore how different types of haptic proxies in IVREs affect the results (e.g., from passive and completely static proxies to proxies with working but faked buttons or actuators).

- *Emphasize the visual accuracy of the virtual twin.*

According to our results, we found that visual cues did affect users when evaluating the virtual twin. Making sure that text, icons, buttons, labels are replicated to the same degree of accuracy can provide beneficial cues on their affordances to users. In line with previous research suggesting that the graphical realism of the scene can affect user behavior [24], we also think that by improving the physical accuracy of the materials properties, shadows and lighting used in the scene and on the virtual twin can minimize the occurrence of related problems (e.g., **UP8**). Future work should also explore multisensory VR experiences, if relevant [16].

- *IVREs can be especially suited for performing tasks that could be difficult to replicate in the RE.*

Participants (*P3, P6, P11, P12*) commented that they felt more free to explore the virtual twin's function since *"It won't be broken."* Analogously, hazardous scenarios (e.g., the microwave catching fire) could be tested in VR without repercussions. In the future, an IVRE coupled with a high-fidelity physics system could also be used to "stress test" devices, and simulate conditions that might lead to structural integrity problems.

- *Providing enough training sessions before performing an IVRE.*

If the interaction in IVRE does not match the real world's and involves additional devices, such as controllers, designers should introduce users to the VR interaction via a training session. Completing a quick "tutorial" before proceeding to the actual can help to rule out simple issues related to inexperience with the VR interface. Participants should be introduced to all the functions and given the opportunity to explore the interface on their own.

## 6 CONCLUSION

In this work we investigated the feasibility of applying a conventional usability method such as the Think Aloud Protocol and standard questionnaires in an Immersive Virtual Reality Evaluation with a case study.

More than 60% of the usability issues found were shared in both the virtual and physical prototype. Although most of the usability problems that were solely found in VR were attributable to limitations of hardware and the interactive modality, VR participants behaved more actively and felt more free to interact with the virtual prototype because of the perceived lack of consequences from any wrongdoings.

In future work, we will investigate the impact that the inclusion of the context of use, as a three-dimensional VE, and of the haptic fidelity of the manipulable components on the results of immersive VR evaluations, as well as evaluate the usability of other classes of devices and prototypes using IVREs.

## ACKNOWLEDGMENTS

This research is supported by *Internal Funds KU Leuven* (HFG-D8312-C14/20/078).

## REFERENCES

[1] Obead Alhadreti and Pam Mayhew. 2017. To Intervene or Not to Intervene: An Investigation of Three Think-Aloud Protocols in Usability Testing. *J. Usability Studies* 12, 3 (may 2017), 111–132.

[2] Obead Alhadreti and Pam Mayhew. 2018. Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173618

[3] Chase Boothe, Lesley Strawderman, and Ethan Hosea. 2013. The effects of prototype medium on usability testing. *Applied Ergonomics* 44, 6 (2013), 1033–1038. https://doi.org/10.1016/j.apergo.2013.04.014

[4] Sebastian Boring, Marko Jurmu, and Andreas Butz. 2009. Scroll, Tilt or Move It: Using Mobile Phones to Continuously Control Pointers on Large Public Displays. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7* (Melbourne, Australia) *(OZCHI '09)*. Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/1738826.1738853

[5] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189 (1996), 4–7.

[6] Fabio Bruno and Maurizio Muzzupappa. 2010. Product interface design: A participatory approach based on virtual reality. *International Journal of Human-Computer Studies* 68, 5 (2010), 254–269. https://doi.org/10.1016/j.ijhcs.2009.12.004

[7] Robbe Cools, Jihae Han, and Adalberto L. Simeone. 2021. SelectVisAR: Selective Visualisation of Virtual Environments in Augmented Reality. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 275–282. https://doi.org/10.1145/3461778.3462096

[8] Afke Donker and Panos Markopoulos. 2001. Assessing the effectiveness of usability evaluation methods for children. *PC-HCI2001, Patras, Greece* (2001), 409–410.

[9] Joseph S Dumas and Janice Redish. 1999. *A practical guide to usability testing*. Intellect books.

[10] K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data.* the MIT Press.

[11] Adrian Fernandez, Silvia Abrahão, and Emilio Insfran. 2012. A systematic review on the effectiveness of web usability evaluation methods. (2012), 52–56. https://doi.org/10.1049/ic.2012.0007

[12] Asbjørn Følstad and Kasper Hornbæk. 2010. Work-domain knowledge in usability evaluation: Experiences with Cooperative Usability Testing. *Journal of Systems and Software* 83, 11 (2010), 2019–2030. https://doi.org/10.1016/j.jss.2010.02.026 Interplay between Usability Evaluation and Software Development.

[13] ROGER R HALL. 2001. Prototyping for Usability of New Technology. *International Journal of Human-Computer Studies* 55, 4 (oct 2001), 485–501. https://doi.org/10.1006/ijhc.2001.0478

[14] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[15] James R Lewis. 2002. Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction* 14, 3-4 (2002), 463–488. https://doi.org/10.1080/10447318.2002.9669130

[16] Imran Mahalil, Azmi Mohd Yusof, and Nazrita Ibrahim. 2020. A literature review on the effects of 6-Dimensional virtual reality's sport applications toward higher presense. In *2020 8th International Conference on Information Technology and Multimedia (ICIMU)*. 277–282. https://doi.org/10.1109/ICIMU49871.2020.9243570

[17] Ville Mäkelä, Rivu Radiah, Saleh Alsherif, Mohamed Khamis, Chong Xiao, Lisa Borchert, Albrecht Schmidt, and Florian Alt. 2020. Virtual Field Studies: Conducting Studies on Public Displays in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376796

[18] Florian Mathis, Kami Vaniea, and Mohamed Khamis. 2021. RepliCueAuth: Validating the Use of a lab-based Virtual Reality Setup for Evaluating Authentication Systems. In *Proceedings of the 39th Annual ACM Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3411764.3445478

[19] Niels C. Nilsson, André Zenner, and Adalberto L. Simeone. in press. Propping up Virtual Reality with Haptic Proxies. *IEEE Computer Graphics and Applications* (sep in press), to appear. https://doi.org/10.1109/MCG.2021.3097671

[20] Viktorija Paneva, Myroslav Bachynskyi, and Jörg Müller. 2020. Levitation Simulator: Prototyping Ultrasonic Levitation Interfaces in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376409

[21] Priyanka Pazhayedath, Pedro Belchior, Rafael Prates, Filipe Silveira, Daniel Simões Lopes, Robbe Cools, Augusto Esteves, and Adalberto L. Simeone. 2021. Exploring Bi-Directional Pinpointing Techniques for Cross-Reality Collaboration. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 264–270. https://doi.org/10.1109/VRW52623.22021.00055

[22] Daniel Schneider, Alexander Otte, Axel Simon Kublin, Alexander Martschenko, Per Ola Kristensson, Eyal Ofek, Michel Pahud, and Jens Grubert. 2020. Accuracy of Commodity Finger Tracking Systems for Virtual Reality Head-Mounted Displays. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 804–805. https://doi.org/10.1109/VRW50115.2020.00253

[23] Adalberto L. Simeone, Robbe Cools, Stan Depuydt, João Maria Gomes, Piet Goris, Joseph Grocott, Augusto Esteves, and Kathrin Gerling. 2022. Immersive Speculative Enactments: Bringing Future Scenarios and Technology to Life Using Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 17, 20 pages. https://doi.org/10.1145/3491102.3517492

[24] Adalberto L Simeone, Ifigeneia Mavridou, and Wendy Powell. 2017. Altering user movement behaviour in virtual environments. *IEEE transactions on visualization and computer graphics* 23, 4 (2017), 1312–1321. https://doi.org/10.1109/TVCG.2017.2657038

[25] Adalberto L. Simeone, Eduardo Velloso, and Hans Gellersen. 2015. Substitutional Reality: Using the Physical Environment to Design Virtual Reality Experiences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3307–3316. https://doi.org/10.1145/2702123.2702389

[26] Mel Slater, Andrea Brogni, and Anthony Steed. 2003. Physiological Responses to Breaks in Presence: A Pilot Study. In *Presence 2003: The 6th annual international workshop on presence*.

[27] Anthony Steed, Tuukka M. Takala, Daniel Archer, Wallace Lages, and Robert W. Lindeman. 2021. Directions for 3D User Interface Research from Consumer VR Games. *IEEE Transactions on Visualization and Computer Graphics* 27, 11 (2021), 4171–4182. https://doi.org/10.1109/TVCG.2021.3106431

[28] Alexandra Voit, Sven Mayer, Valentin Schwind, and Niels Henze. 2019. Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300737

[29] Tingting Zhao, Sharon McDonald, and Helen M Edwards. 2014. The impact of two different think-aloud instructions in a usability test: a case of just following orders? *Behaviour & Information Technology* 33, 2 (2014), 163–183. https://doi.org/10.1080/0144929X.2012.708786