

Using Think-Aloud Protocol in Immersive VR Evaluations



Xuesong Zhang and Adalberto L. Simeone

Abstract There is a growing research interest concerning the use of VR as a medium in which to stage experimental evaluations. However, it is still unclear whether conventional usability evaluation methods can be directly applied to virtual reality evaluations and whether they will lead to similar insights when compared to the results of conventional usability real-world laboratory studies. Furthermore, the impact of graphic and interaction fidelity of the virtual prototype on the evaluation result is largely unexplored. Hence, we conducted two user studies with 44 participants. Results in the first study show that 61% of the reported usability problems were shared by both versions, highlighting the potential of Immersive Virtual Reality Evaluations (IVREs) as a method to evaluate early design concepts before committing to a physical prototype. In the second study, we found that usability issues related to the manual operation of the interactive device might not be identified if the interaction is implemented via supernatural techniques. In terms of graphical implications, designers should focus on those visual elements that communicate information to users. We discuss these implications and provide guidelines for deploying IVREs.

1 Introduction

The user-centered design (UCD) process starts with a concept based on the context of use, which is then followed by creating one or more prototypes at different fidelity levels (Lim et al. 2006). Prototypes are then iteratively refined through successive evaluations (Benyon 2019). However, building high-fidelity physical prototypes of a proposed concept or early idea is not always possible because of costs, time and practicality, or because the technology required could be immature or even unfeasible given current technological constraints.

X. Zhang (✉) · A. L. Simeone
Department of Computer Science, KU Leuven, Leuven, Belgium
e-mail: xuesong.zhang@kuleuven.be

A. L. Simeone
e-mail: adalberto.simeone@kuleuven.be

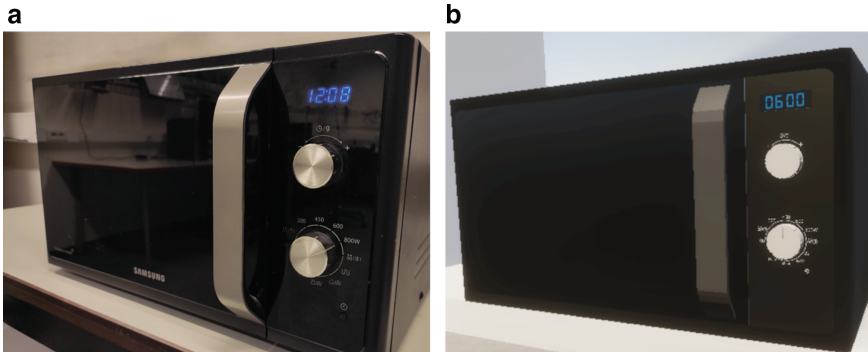


Fig. 1 Microwave (a) and its virtual twin (b) used in our user study

Virtual Reality (VR) can provide a way to address this challenge, since it allows us to simulate virtual counterparts of any device and technology, even non-existing ones. Voit et al. (2019) demonstrated the feasibility of inspecting the usability of an artifact in VR, as well as through online, laboratory, AR, and in situ studies. Further, Mäkelä et al. (2020) used VR as a medium to stage field studies in virtual environments (VEs) simulating the proposed context of use. We refer to this type of studies as “*Immersive Virtual Reality Evaluations*” (IVREs).

As a concept, IVREs propose the use of controllable VEs in which to stage the evaluation of a (virtual) prototype. Our hypothesis is 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 physically built. In this context, IVREs provide various advantages: (1) there is positive evidence that results obtained with VR evaluations can be transferred to the real world (Mathis et al. 2021); (2) we can stage field studies in VR that could be difficult to replicate in the real life (Mäkelä et al. 2020); (3) IVREs can be performed before a physical prototype is actually built, thus serving as a way to identify potential problems at an early stage; (4) since the evaluation happens fully in VR, IVRE can be performed from everyday environments instead of dedicated laboratories.

However, it is still unclear whether conventional usability evaluation methods can be as effective in VR as they are in reality in uncovering real usability issues from virtual prototypes. To answer this research question, we designed and conducted the first study, a comparative between-subjects user study: we evaluated the usability of a microwave oven model on both the real-world appliance and on an interactive “virtual twin”. We define *virtual twins* as the explicitly virtual interactive counterpart to the more popular “digital twin” term (Tao et al. 2019), which focus more in the direct data-based connection and do not always require the existence of a three-dimensional model that replicates its interactive functionalities in VR. We chose to focus our study on a microwave, as they represent household appliances that many people are familiar with.

Both prototypes were evaluated via the *Think-Aloud Protocol (TAP)* (Ericsson and Simon 1984) and questionnaires such as the *System Usability Scale (SUS)* (Brooke 1996), *Post-Study System Usability Questionnaire (PSSUQ)* (Lewis 1995, 2002) to rate the user perceived usability and the *NASA Task Load Index (NASA-TLX)* (Hart 2006) to assess the task workload. For more information on these methods, please see Sect. 3.1.

Results from the first user study indicate that 61% of the usability issues present on the virtual prototype were also identified on the physical appliance. Participants reported similar numbers of identified usability problems in terms of type, when evaluating both prototypes with the *TAP* method. The SUS, PSSUQ, NASA-TLX questionnaires also offered similar scores when performing inspection in both settings. However, some usability issues were only detected with the physical prototype, since the haptic feedback is missing in VR. Although most of the usability problems that were solely found in VR were attributable to limitations of 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.

Given the positive results in terms of overlapping problems, in the second study, we sought to understand whether the graphical appearance or the interactive fidelity of the virtual twin would affect the type and severity of the usability issues. Indeed, there is evidence in literature that the prototype representation itself or its interactive affordances can have an impact on the evaluator's behavior and perceived usability (Hoggenmüller et al. 2021; Simeone et al. 2017). Further, due to technical limitations of VR, it might be necessary to interact in a way that is different from its real-world analog. Interactions that we take for granted in the real world, such as pressing a button or turning a knob, are not so trivial in VR, especially when no haptic feedback is present (Nilsson et al. 2021) or the accuracy of the hand-tracking system is insufficient (Schneider et al. 2020).

In the second study, we created four different versions of the microwave oven that were modelled according to two levels of fidelity in both graphics (between *Simple* and *Physically Based Rendering* materials) and interaction (between *Natural* and *Supernatural* interaction modalities (McMahan et al. 2016; Lubos 2018; Yu and Bowman 2018), where simple 3D interaction metaphors replace complex real-world manipulations).

Similar to the first study, participants used the *TAP* method to evaluate the version of the microwave oven they were assigned to (see Sect. 6). We focused on the qualitative feedback of participants. Results indicate that supernatural interaction could make it difficult to identify problems which are related to manual manipulations, despite the fact that participants identified a comparable number of usability problems. In addition, more efforts should be made to achieve a less ambiguous graphical appearance during the implementation of virtual twin.

This chapter is an extended version of the paper entitled "Using the Think Aloud Protocol in an Immersive Virtual Reality Evaluation of a Virtual Twin," which was published in ACM SUI 2022. The contribution of this chapter is threefold: (1) We

report the identified usability problems and analyze them depending on whether they affected both prototypes or solely one of the two. (2) We analyze the impact of graphic and interaction fidelity during the usability evaluation for the virtual prototypes. (3) We discuss which factors may affect the identification of usability problems in IVREs and provide a set of guidelines for performing them in the future studies.

2 Related Work

In this section, we present user studies employing VR to stage user studies and research related to the effects of graphics and interaction in VR.

2.1 Evaluations Performed in Virtual Reality

In 2007, Ye et al. (2007) investigated the possibility of applying VR technologies to computer-aided product design. Users viewed the generated 3D graphics through lightweight stereoscopic LCD glasses and perceived haptic feedback with a SensAble PHANTOM Desktop device. Results showed that the proposed system offered a natural and intuitive interaction, which contributes to reducing the development time and design costs. The 2010 study by Bruno et al. (2010) is the closest to our first study. The authors compared the results of the usability evaluation of a real microwave, to those resulting from the evaluation of its three-dimensional twin experienced via a semi-immersive, stereoscopic, projected screen without head-tracking. Differently from our study, participants did not use the TAP: the experimenters observed them while they interacted with the two microwave ovens. Participants then filled in a usability satisfaction questionnaire. Results indicate that the number and type of reported problems were comparable across conditions.

In 2013, Falcão and Soares (2013) proposed the application of VR to the usability evaluation of consumer products. The authors stated that traditional evaluation with physical prototypes is expensive and difficult to rapidly iterate on. They suggested that employing VR technology could solve those limitations, but did not carry out any study.

More recently, with the increased affordability of VR HMD-based solutions (Bellalouna 2019), researchers began to explore the use of VR as a research method. In 2019, Voit et al. (2019) compared the evaluation of technological artifacts under different settings: online, VR, AR, laboratory setup, and in situ. Three standard questionnaires (AttrakDiff (Hassenzahl et al. 2003), ARI (Georgiou and Eleni 2017), and SUS) were employed for the evaluation. They found that the medium used to assess artifacts affected the results, with VR providing comparable results to the real world.

In 2020, Mäkelä et al. (2020) analyzed the feasibility of applying VR as a testbed for the evaluation of different implementations of public displays. A largely similar user behavior was observed in both the VE and the real-world setting. Paneva et

al. developed a “Levitation simulator” (2020), where a virtual twin allows users to prototype applications based on ultrasound hardware. They used the simulator to develop two levitation games, which were then implemented on the physical platform. Performance and engagement levels with the developed games were found to be comparable to those reported with the real apparatus.

In 2021, Mathis et al. (2021) replicated an authentication schema from a real-world laboratory setting into a VE. Results indicated that the virtual version has similar usability as the physical prototype in terms of entry accuracy, entry time, and perceived workload, as measured by task completion times and NASA-TLX scores (Hart 2006). Saffo et al. (2021) also leveraged a social VR platform (VRChat) and replicated two published user studies: a quantitative study on Fitt’s law and a qualitative study on tabletop collaboration. Those two social VR studies yielded analogous results as in the original study, which contributes positive evidence to the validity of using social VR to perform HCI evaluations.

With this work, we further contribute to the growing field of work on studies ran in VR, with a study comparing applying the TAP method to a conventional inspection and to one performed in immersive VR.

2.2 Effect of Prototype Representation in Terms of Graphic Appearance

There is evidence that using realistic lighting and graphics affects immersive users perceive the sense of presence in the VE. Previous work shows that when visual realism is improved, participants can perceive greater presence and respond with more realistic behavior (Slater et al. 2009; Yu et al. 2012). Furthermore, Simeone et al. found that the graphical appearance of a material affect user behavior in the VE (Simeone et al. 2017). In 2021, Hoggenmuller et al. presented a comparative user study of real-world VR (i.e., 360° video in VR), computer-generated VR, and real-world video (Hoggenmüller et al. 2021). By comparing the results gathered from these three prototype representations, authors found that in the real-world VR representation, despite users perceiving high presence, the user experience was similar to the other two representations. However, qualitative data showed participants focused on different experiential and perceptual aspects with different representations, highlighting that the choice of a proper graphic appearance is important in collecting user feedback.

2.3 Effect of the Interactivity Fidelity of Prototype

In this chapter, we define the virtual prototypes explicitly as the virtual interactive counterpart to the more popular term “digital twin” (Tao et al. 2019), which does not

always require the existence of a three-dimensional object that replicates its interactive functionalities in VR. The ability to interact is an important feature of the virtual twin. McMahan et al. pointed out the interaction fidelity along with display fidelity affects user performance, presence, engagement, and system usability (McMahan et al. 2012). Furthermore, the interaction fidelity is an “uncanny valley” in terms of user performance: the low fidelity interaction contributes to comparable results as the high fidelity (McMahan et al. 2016). User performance is better than with the mid-fidelity approach. Rogers et al. further suggested the impact of interaction fidelity varies between interaction tasks; however, users prefer the high interaction fidelity since it is more immersive and enjoyable (Rogers et al. 2019). To our knowledge, there is no other empirical work investigating the impact of interaction fidelity on IVREs.

3 User Study 1: Physical Versus Virtual Prototypes

We designed a between-subjects user study with the aim of comparing the results of a usability evaluation of a microwave oven in a real-world laboratory setting, with those resulting from the inspection of its virtual twin while the evaluator is immersed in a VE. Both evaluations were performed using the *Think-Aloud Protocol (TAP)* and standard questionnaires including *SUS*, *PSSUQ*, *NASA-TLX*. The independent variable in this study was the STUDY MODALITY: {Real Environment (RE), Virtual Environment (VE)}.

We recorded the entire evaluation process and transcribed the participants’ dialogs. We further analyzed the data resulting from the TAP to identify the problems, the scope each related to, and their severity. Participants were asked to inspect a microwave oven. As a common kitchen appliance, microwaves share functions whose implementation varies between manufacturers. In the study, participants acted as the evaluators. We assigned them a set of tasks that used the microwave’s advanced functions to achieve a desired effect (such as defrosting food based on the time or the weight, the combination of grill and microwave function).

While participants all shared to some extents a basic understanding of how a microwave works, we expected that the specific way in which the oven interfaces are implemented across manufacturers would have led to the surfacing of various usability problems. Further, we hypothesized that a share of these problems would have manifested themselves in the virtual twin as well. The study thus aims to investigate the extent of this overlap, and the nature and severity of the identified problems, through qualitative analysis. The results provide further insights on whether performing the IVRE alone on a concept of a product that does not yet exist physically could become a valid complementary tool to help refine and iterate faster on design activities, before committing to a physical prototype.

3.1 Usability Evaluation Methods

Think-Aloud Protocol. The Think-Aloud Protocol (TAP) asks participants to verbalize their thoughts while performing specific tasks (Ericsson and Simon 1984). The information collected provides an account of which usability problems were experienced and indications as to the source of these issues.

Alhadreti and Mayhew compared three TAP types to the evaluation of the user interface of a university library website (Alhadreti and Mayhew 2018). They were: the *concurrent* TAP method, where users verbalize their thoughts while interacting (Ericsson and Simon 1984); the *retrospective* TAP method, where users provide a report after finishing interacting with the object of the evaluation (Ericsson and Simon 1984); the *hybrid* method, which represents the combination of the first two types (Følstad and Hornbæk 2010). Results show that the concurrent TAP method detected more usability problems than the retrospective TAP. No significant differences were found between concurrent and hybrid method in terms of number of problems. The concurrent TAP needed the shortest amount of time in terms of conduction and analysis among those methods. For these reasons, we chose the concurrent TAP as the usability evaluation method to use in our user studies.

Questionnaires. Using questionnaires in usability evaluations is a common practice that allows experimenters to collect data in a quick and cost-effective way (Zaharias and Poylymenakou 2009). The *System Usability Scale (SUS)* (Brooke 1996) is a widely used validated questionnaire, since SUS is a valid and reliable tool to quickly measure the perceived usability of products or services (Brooke 2013). SUS is a cheap, effective, and robust tool and works even with a small sample (8–12 users) to get a valid assessment (Tullis and Stetson 2004; Bangor et al. 2009). The SUS questionnaire consists of ten different questions with a five-point scale ranging from 1: *strongly disagree* to 5: *strongly agree*. Researchers usually use a scoring system (from 0: *negative* to 100: *positive*) that enables the comparison of two versions of an application (Brooke 2013).

We also used the *Post-Study System Usability Questionnaire (PSSUQ)* (the shorter version) in our user study, which consists of 16 items and measures the system usability from *System Usefulness*, *Information Quality* to *Interface Quality* (Lewis 2002). In PSSUQ, unlike SUS, a seven-point scale was used, ranging from 1: *strongly agree* to 7: *strongly disagree*. Thus, a lower score means a better evaluation. This questionnaire is suitable for evaluating different types of products at different development stages based on its generalizability (Lewis 2002).

To measure the perceived workload while accomplishing tasks, we use the *NASA Task Load Index (NASA-TLX)*, which consists of six subscales, namely *Mental Demand*, *Physical Demand*, *Temporal Demands*, *Frustration*, *Effort*, and *Performance* (Hart 2006). A twenty-point scale (1: *Low*, 20: *High*) was used here. The summary of unweighted scores represents the overall workload experienced from 6 to 120.

Those questionnaires have been used in previous IVRE studies, e.g., (Voit et al. 2019; Mathis et al. 2021), thus we applied them in our study as well.

3.2 Apparatus

We created a virtual twin of a real microwave appliance (Fig. 1a) by Samsung (MG23F301E). This microwave was released in 2014 (according to its manual¹) and has been superseded by an improved model released in 2021. It is fully functional and owned by one of the authors. Based on this model, we created a three-dimensional replica using Blender (Fig. 1b). One of the features of this microwave oven is that if there is no further change after setting the function or timer, the microwave will automatically start cooking within two seconds. While participants could overall be assumed to be familiar with the basic functionality of a microwave, we expected to find various usability problems depending on how well its physical interface had been designed.

The virtual twin has the same dimensions as the physical microwave. Its interactive features were implemented in Unity 2020.3.3,² and it was rendered with the High Definition Rendering Pipeline.³ Participants interacted with the virtual twin through a HTC Vive Pro HMD.⁴ Interaction with the microwave dials and buttons was implemented via collision-based selection with a Vive controller, due to the insufficient accuracy of the embedded hand-based detection. 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.

When the reference cube collides with a button, participants can press the trigger to confirm the selection, which triggers the action associated with the collided button on the microwave. If the cube collides with a dial, participants need to press the trigger and hold while turning their wrist to rotate the dial. During the experiment, participants also needed to open the microwave door to put in or take out (virtual) food, if necessary. The door was likewise opened by holding the trigger on the controller in the proximity of the handle and pulling (respectively, pushing) it until it was fully opened (closed). The virtual twin simulated all the functionalities of the physical microwave. 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. An animation was created to simulate the defrost/heat/microwave/grill process inside the microwave.

The group that interacted with the physical appliance did so in our laboratory, where the oven had been temporarily placed. A separate cup containing water is secured inside the microwave oven to prevent it from running empty. No food or other drinks were actually heated in this condition. The user study was approved by the Ethical Review Board of our institution.

¹ <https://www.manua.ls/samsung/mg23f301eas/manual?p=1>.

² <https://unity3d.com/unity/whats-new/2020.3.3>.

³ <https://docs.unity3d.com/Packages/com.unity.render-pipelines.high-definition@11.0/manual/index.html>.

⁴ https://www.vive.com/eu/support/vive-pro-hmd/category_howto/about-the-headset.html.

3.3 Demographics

We recruited 24 participants (13 male, 11 female) between the ages of 23 and 32 (MEAN = 26.96, SD = 2.63) for this user study. Twelve participants were randomly assigned to each group. They were recruited through internal mailing lists, word-of-mouth and social media. There was no compensation for participation in this user study.

3.4 Procedure

After filling a consent and a demographics form, we introduced participants to the TAP evaluation method and the procedure to perform it. Then, participants were asked to sit in front of the (virtual) microwave to perform the evaluation. Participants had to complete eight tasks with the (virtual) microwave, 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. These tasks were designed to prompt participants to pay attention to the icons, operate all the buttons or knobs, and experience all the functions of the microwave oven. A cup is prepared and treated as the food described in the task. The brief instructions detailing how to perform them came from the manual and were relayed to the participants by the experimenter. Before the formal evaluation, participants went through a training session to get familiar with the interaction in the VE. The experimenter introduced the interaction with the HMD and its controller, where participants could practice the interaction action.

After participants confirmed they understood the purpose of the evaluation, the experimenter gave them a signal to start. While performing it, participants described their actions and thoughts. They finished the task by either completing or abandoning it. Participants were allowed to abandon the task after three unsuccessful attempts to complete it. Three tasks were abandoned by two participants. In total, 192 trials were performed (8 tasks \times 24 participants). We recorded the RE sessions with a smartphone camera, and VE sessions with OBS⁵ to record the first-person view.

After evaluating the microwave, we asked participants to fill in four web-based questionnaires: the SUS, the PSSUQ, the NASA-TLX, and a custom questionnaire with ten required five-point scale questions (where 1 is strongly disagree and 5 is strongly agree). In the custom questionnaire, two questions aimed to understand participants' opinion on whether performing the TAP affected their task performance; five questions aimed at eliciting their view on the TAP method. The remaining questions aimed to understand the impact of the experimenter's presence on the study. Then, the participants proceeded to freely explore the microwave in either the real or virtual environment without completing tasks. After this exploration phase, participants needed to fill in another custom questionnaire with three questions with

⁵ OBS- <https://obsproject.com/>.

Table 1 Task list for each microwave

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

a five-point scale. Those questions asked participants to compare the physical and virtual models in terms of the similarity of their appearance and their operational similarity (1: completely inconsistent; 5: completely consistent). We also asked them to predict whether performing the task with the virtual prototype would require more time than with the physical prototype (1: strongly disagree; 5: strongly agree). At the end of the study, we conducted semi-structured interviews to let participants walk us through their feedback on the use of the TAP in the RE and in the VE. Each evaluation lasted about 60 min.

4 Results of User Study 1

In this section, we report the quantitative data collected during the study. The identified usability problems were differentiated according to whether they affected one or both prototypes. We then categorized them into four common areas. Finally, we report the results of the questionnaires filled by the participants after the evaluation and the task completion times.

4.1 *Detected Usability Problems from Think-Aloud Protocol*

Distribution of the detected problems. We followed a two-stage extraction process to identify the usability problems, leading from individual problems to final problems, as proposed by Alhadreti and Mayhew (2018). After this process, a total of 46 distinct usability problems were identified; of these, 28 overlapping problems were detected for both the virtual and real microwave. Five problems were only found by the participants in the RE, and 13 problems are unique to the VE (see Fig. 2).

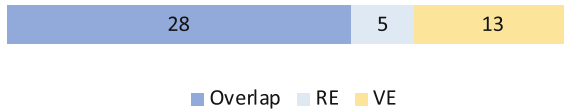


Fig. 2 Distribution of detected usability problems

We compared the number of detected problems by performing a Kruskal-Wallis H-test on the data, as in Alhadreti and Mayhew’s work (2018). 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 processes of the functions are not always same within the same microwave. However, if the user does not fully understand or remember the process correctly, they might then not know what the next step to perform 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 of the set time or function, the microwave oven will automatically activate. 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 this button again for the next task. However, the button did not work as expected, because pressing it enters 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. However, they cannot 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 digit after setting the hour digit. 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, and user 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.

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

	Overlap					RE					VR				
	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

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: *H*: Critical; *M*: Major; *L*: Low; and *E*: Enhancement (Dumas et al. 1999; Alhadreti and Mayhew 2017; Zhao et al. 2014; Alhadreti and Mayhew 2018). Their distribution is shown in Table 2.

The list of detected usability problems. The following are the usability problems we identified through a two-stage extraction process (Alhadreti and Mayhew 2018), which were identified in both settings, or only identified in the RE or the VE. We will discuss these usability issues in more detail in the *Discussion* (Sect. 5).

Unique usability problems in the VE

- UP1** *Participants perceived the image as blurred. (C4, L, 9 times)*
- UP2** *Participants tried to start the microwave by pressing the knob. (C4, E, 4 times)*
- UP3** *The knob rotation is not intuitive and slow. (C4, L, 3 times)*
- UP4** *The knob rotation is tiresome. (C4, L, 3 times)*
- UP5** *Participants cannot open the door. (C4, L, 2 times)*

Unique usability problems in the RE

- UP6** *Confusion on how to set the timer. (C2, M, 10 times)*
- UP7** *Attempting to start the microwave by pressing the clock button. (C4, E, 8 times)*
- UP8** *Using an incorrect knob to adjust the minute setting. (C4, L, 1 time)*
- UP9** *The microwave door is hard to open and close. (C4, L, 2 times)*

Overlapping problems in both the RE and VE

- UP10** Participants are confused by similar icons (two defrost icons, three grill-microwave combination icons). (C4, L, RE: 4 times; VE: 7 times)
- UP11** Participants are confused by similar defrost functions. (C3, H, RE: 2 times; VE: 2 times)
- UP12** The user misselected another function adjacent to the position of the target function. (C4, M, RE: 2 times; VE: 3 times)
- UP13** Lack of a START button. (C2, L, RE: 6 times; VE: 8 times)
- UP14** Lack of a STOP button. (C2, H, RE: 3 times; VE: 3 times)
- UP15** Activating an empty microwave. (C1, H, RE: 1 time; VE: 2 times)
- UP16** The knob is not sensitive to small angle rotation. (C4, L, RE: 4 times; VE: 5 times)

4.2 Custom Questionnaire—TAP Experience

We ran Kruskal-Wallis H tests on the questionnaire results to detect if there were differences when conducting TAP in the RE and in the VE as in Alhadreti and Mayhew (2018). No significant differences were detected. Table 3 shows the results.

4.3 SUS

The SUS scores were first calculated (Brooke 2013), then Shapiro-Wilk tests were run to inspect the normality of the distribution. There is no evidence that the SUS scores of the physical microwave ($W(12) = 0.96$, $p = 0.785$) and its virtual twin ($W(12) = 0.92$, $p = 0.285$) deviate significantly from a normal distribution.

Table 3 Result of the TAP experience questionnaire, where 1 stands for “strongly disagree” and 5 is for “strongly agree”

	VE		RE		<i>p</i>
	MEAN	SD	MEAN	SD	
It felt tiring to perform the TAP	2.75	1.29	2.83	1.34	0.881
It felt unnatural to perform the TAP	2.50	1.24	3.50	1.51	0.098
It felt unpleasant to perform the TAP	2.00	0.95	2.50	1.31	0.384
It felt difficult to perform the TAP	1.92	0.79	2.83	1.47	0.114
It felt time consuming to perform the TAP	2.83	1.27	3.83	1.19	0.067
The presence of the evaluator was unnatural	1.83	0.72	2.58	1.38	0.167
The presence of the evaluator felt unpleasant	1.58	0.67	2.00	1.21	0.531
The presence of the evaluator felt disturbing	1.33	0.65	1.83	1.03	0.181

One-way ANOVA tests were run to determine if there were significant differences in terms of the SUS scores attribute to the microwave across the virtual and real conditions, as also used by Alhadreti and Mayhew (2018). Participants rated the virtual twin (MEAN = 61.25, SD = 11.75) as having a marginally higher usability than the physical microwave (MEAN = 55.63, SD = 16.00), where values between 50.9 between 71.4 correspond to a good level of usability (Bangor et al. 2009). However, the difference was not significant ($p = 0.339$).

4.4 PSSUQ

One-way ANOVA tests were run to determine if there were differences in terms of the PSSUQ scores of the physical microwave and its virtual counterpart. Overall, participants rated the physical microwave (MEAN = 2.12, SD = 1.38) as having lower performance and were less satisfied with it, compared to its virtual counterpart (MEAN = 1.90, SD = 1.01); the virtual microwave was also rated as having higher system usefulness, higher information quality, and higher interface quality; however, no significance was detected among the results ($p = 0.669$). Table 4 shows the partial results.

4.5 NASA-TLX

One-way ANOVA tests were run to determine if there were differences in terms of workload between interaction with the virtual twin or with the real appliance. Overall, participants perceived comparable levels of workload in both the RE (MEAN = 41.92, SD = 15.27) and in the VE (MEAN = 45.92, SD = 17.80) without statistical significance ($p = 0.561$) (Fig. 3).

Table 4 Partial result of PSSUQ questionnaire, where 1 stands for “strongly agree” and 7 is for “strongly disagree”

	VE		RE		
	MEAN	SD	MEAN	SD	p
Overall	1.90	1.01	2.12	1.38	0.669
System usefulness	1.74	1.01	1.76	1.50	0.958
Information quality	2.32	1.38	2.58	1.69	0.687
Interface quality	1.75	1.31	1.89	1.37	0.802

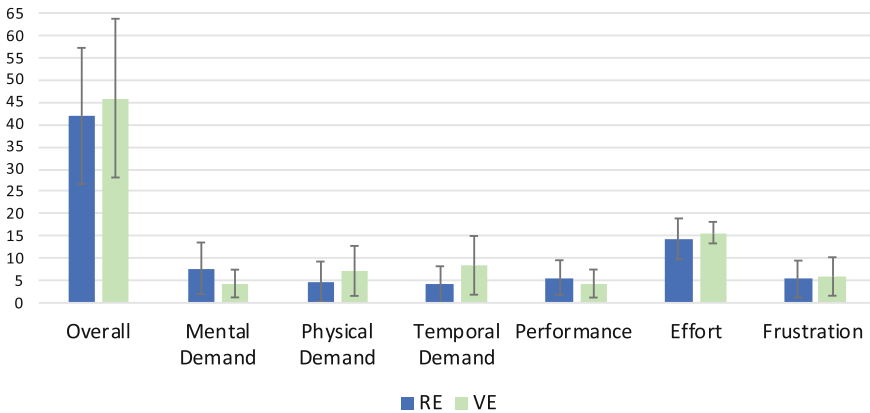


Fig. 3 NASA-TLX scores of the microwave and its virtual twin. The error bars indicate the standard deviation

4.6 Similarity of the Virtual Microwave

After the formal evaluation with TAP, participants were asked to freely interact with the other microwave. During this exploration session, they were not required to complete any task. To understand the perceived experience difference when using the two prototypes, we invited them to answer the following three questions afterward:

- *I think the virtual model is identical to the real microwave oven in terms of appearance. (1: strongly disagree; 5: strongly agree)*
- *I think the virtual model is identical to the real microwave oven in terms of working mechanism. (1: strongly disagree; 5: strongly agree)*
- *I feel that it would take more time when interacting with virtual microwaves than when interacting with physical microwaves. (1: interacting with virtual microwave takes less time; 3: there is no time difference; 5: interacting with virtual microwave takes more time).*

Participants were generally in agreement on the similarity of the implementation of the virtual twin based on the results from the custom questionnaire. They thought that the virtual twin represented a close approximation of the physical appliance in terms of both visual appearance (VE: MEAN = 4.08, SD = 1.08; RE: MEAN = 3.25, SD = 1.36) and operation (VE: MEAN = 3.58, SD = 1.56; RE: MEAN = 3.17, SD = 1.47). However, participants believed interacting with the virtual microwave takes more time than with the physical one (VE: MEAN = 4.50, SD = 0.67; RE: MEAN = 4.50, SD = 0.80). Scores here refer to which microwave type they interacted first.

Table 5 Table reports the full results of the analysis of the task completion times

	VE			RE		
	MEAN	SD	MEAN	SD	$H(2)$	p
Task 1	45.88	15.43	29.67	18.53	4.204	0.040
Task 2	59.83	29.80	73.40	53.30	0.004	0.951
Task 3	34.87	17.23	40.45	20.83	0.641	0.423
Task 4	57.25	41.44	27.71	14.19	5.079	0.024
Task 5	57.02	32.01	32.50	22.10	4.449	0.035
Task 6	54.13	18.00	41.36	37.59	3.649	0.056
Task 7	56.99	19.60	45.83	34.22	2.903	0.088
Task 8	45.01	19.37	52.00	41.02	0.041	0.840
Total time	405.98	88.88	336.12	184.16	3.205	0.073

4.7 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 Sect. 3.4. In total, there was one participant who abandoned two tasks in the RE, while one task was abandoned by one participant in the VE.

Since a Shapiro-Wilk test determined that the TCTs of the microwave significantly deviated from a normal distribution, we used non-parametric methods. Kruskal-Wallis H tests were run on the data 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: $p = 0.035$). The full results of all TCTs are shown in Table 5.

5 Discussion of User Study 1

In this section, we discuss the usability issues that were found as a result of both the IVRE and the conventional laboratory-based evaluation.

Table 6 Resolution used in contemporary retail HMDs (Kreylos 2016), we used a HTC Vive Pro in our user study

HMD	Resolution per eye (px)
Oculus Rift S	1440 × 1280
Oculus Quest 2	1982 × 1920
HTC Vive	1080 × 1200
HTC Vive Pro	1440 × 1600

5.1 Unique Problems in the VE

Through the IVRE, participants reported 13 unique problems that were not found in the inspection of the physical appliance. We classified these into four categories according to the underlying causes (see Sect. 4.1). The most commonly reported usability problems (UP_n) in the IVRE were of type C4: eleven such problems were issues related to the “physical” interface of the virtual twin (see Table 2).

The resolution of the HMDs affects the user’s perception of the information in the VE. Usability problems related to blur image, such as $UP1$, was caused by the low resolution of the used HMD (see Table 6). After the evaluation process, we asked participants in the RE whether they perceived icons to be blurry. They confirmed those were false positives, as it was not the case in the RE. 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.

The different interaction modality compared to the real world also resulted in other unique VE usability problems (i.e., $UP3$, $UP4$).

$P7$, $P8$ reported that the knob turning action in 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 (Simeone et al. 2015).

Similarly, seven participants in the VE reported that they felt tired when rotating the knob, since the interaction is performed mid-air without arm support in a non-ergonomic position. This is similar to the gorilla arm syndrome (Boring et al. 2009). We expect that in the near future, with improved hand-tracking accuracy, natural hand-based interaction 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 Oculus Quest 2 controllers.

We also noticed that the insufficient prior VR experience would lead to VE unique usability problems (e.g., $UP5$). Two participants had no prior experience with manip-

ulating objects in VR and encountered problems while opening the microwave door. They forgot to press the controller trigger when they tried to open the door. Thus, the system did not detect the collision, and the door did not turn to follow the user's hand movements. Our training session lasted for two minutes. A longer session with activities to complete in order to progress could reduce these problems.

5.2 *Unique Problems in the RE*

When evaluating the microwave with the TAP, five problems were only detected in the RE. These problems are largely dependent on the participant's individual experience and understanding of the system (e.g., UP6, UP7, UP8). These usability issues reveal that the system state is not clear enough to the user. Users' previous experience with similar products can also influence them. P16 reported that timer button is placed on the bottom right, which was the "same position as the start button of my own microwave". This led to the misoperation.

The lack of haptic feedback of the virtual model is also partly responsible for the unique usability issues in the VE. UP9 was only found affecting the real microwave, due to the haptic feedback of the physical model, which was absent in the VE. P1 mentioned that the door was difficult to open and to close properly, as it required more strength than expected. In the VE, participants only needed to press the controller trigger to manipulate the door. This usability problem related to the amount of force necessary to operate the door could only be identified in the RE.

5.3 *Overlapping Problems in both the RE and VE*

When evaluating the microwave with TAP, 61% of the usability problems of the microwave were identified in both the RE and VE conditions.

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 microwave. The high number of overlapping problems is in line with findings by Bruno and Muzzupappa (2010), where participants experienced similar difficulties in understanding the microwave features in both the real-life laboratory and its virtual replica settings.

We also noticed participants exhibited different behavior when they encountered the same usability problem. In the RE, participants received different haptic sensations when they touch the buttons on the microwave or grab the cup. This feedback varies depending on the material. However, there is no haptic feedback after the controller collides with different virtual objects in the VE. This leads to differences in the way users perceive system state and perform operations, as implied in UP16.

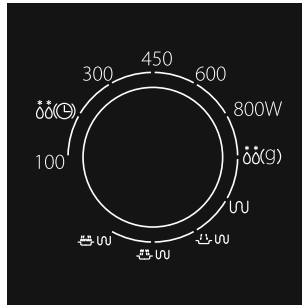


Fig. 4 Icons around the bottom right knob, as seen on the virtual twin

Despite the HMD resolution further complicating participants' perception, each icon on the physical microwave is smaller than $0.5\text{ cm} \times 0.5\text{ cm}$, which limits the space available for the icon design. In both conditions, participants pointed out that they encountered problems in understanding the meaning of the icons. Such as *P14* assuming that the “(g)” label is a special icon; eleven participants had problems distinguishing between the three grill-microwave icons because their differences are minimal. Those icons would need to be redesigned in order to become more easily recognizable.

Usability problems related to the product design were identified in both VE and RE settings, such as [UP11](#), [UP13](#), [UP14](#), [UP15](#). Users can correctly interpret the working mechanism of the prototype in the VE, because they can interact with the same virtual buttons of RE in shape, size, and position, and the virtual microwaves have similar sound and animation responses to mimic the working process. These design faults were present in the virtual prototype too, and participants could identify them. Likewise, the usability problems related to inappropriate placement can be identified in both setting, such as [UP12](#). The icons for weight-based defrost and grill functions are located at the 3 and 4 o'clock positions around the bottom right knob (see [Fig. 4](#)). Since participants were asked to sit in front of the (virtual) microwave, both icons may have been (partially) obscured by the knobs in view. Indeed, we observed *P20 (RE)* and *P22 (VE)* both misselect one of the functions, since both are adjacent to each other.

However, we noticed participants acted differently in the RE and VE when they encountered the same problem. For example, [U16](#), in the time/weight setting, when users turn the upper knob less than 30°C , the time/weight information on the screen will not change. We believe this represents a feature *working as intended* to prevent misoperation. *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*”.

5.4 Task Completion Time

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 s to 180 s. Participants were asked to continuously rotate the knob, and within this time range, the timer increased by 10 s 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.

5.5 Experience with Think-Aloud Protocol

Results show that that participants experienced the TAP in the VE and in the RE in a similar fashion. During the user study, the experimenter guides the TAP process by giving the participant essential instructions (e.g., “*could you describe your current action?*”) to carry on the study. These interruptions on behalf of the experimenter have often been associated with “Breaks-in-Presence” (Slater and Steed 2000) experienced by the participant. However, in the case of the TAP, the user is explicitly asked to describe their thinking and actions from the start. Thus, the connection between the VE and real world was always there.

However, participants reacted differently to the knowledge of the presence of the evaluator when conducting the TAP. Some, like *P6* were surprised by this “Cross-Reality” (Cools et al. 2021) 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 evaluator; this because I felt that they knew the next steps very well*”. Some participants did not notice the evaluator: “*I did not even notice the evaluator during the VR session. The headset blocks out the physical environment*”. (*P4*). Indeed, the evaluator 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.

6 User Study 2: Graphics × Interactivity Fidelity

In our previous study, we found that both the interaction modalities and the rendering quality of a virtual twin had an impact on the results of an usability evaluation. The focus of this study was these two aspects.

We compared the perceived usability, the task workload, the presence experience, as well as the task completion time when users use the TAP in an immersive VE to inspect a virtual twin of a microwave oven. The TAP was applied here again, so that verbal expressions can be used to understand users’ intentions and behaviors. By

replicating the functionalities of the microwave in VR in the exact same way as the real appliance operates in the real world, we exclude confounding factors that might be introduced by the design of an ex novo interactive device.

This user study followed a 2×2 factorial design with two between-subjects variables, i.e., GRAPHICS FIDELITY, INTERACTIVE FIDELITY. The levels in each independent variables were:

- GRAPHICS FIDELITY: {Color Based Rendering (CBR), Physically Based Rendering (PBR) }
- INTERACTIVE FIDELITY: {Natural Interaction(NI), Supernatural Interaction (SI)}.

This user study also followed a split-plot design. We assigned five participants to each condition to find out the usability issues, since four or five users could identified 80% of usability problems when performing the TAP (Virzi 1992; Hwang and Salvendy 2010).

6.1 Interactivity Fidelity

The *interactivity fidelity* indicates the extents of the similarity between an action performed in VR and its analogous action in the real world (McMahan et al. 2016). In this study, only the interaction with the knob and door was implemented with different interaction fidelities.

In the SI condition, a user needs to touch the knob with their virtual hand and press either the left or the right area on the trackpad to turn the knob counter-clockwise or clockwise (Fig. 5a). Each click is equivalent to 30° rotation, which is the minimum increment with the real microwave. In the NI condition, a user needs to physically rotate their wrist in correspondence of the knob while pressing the trigger button. The knob rotates along with the user's wrist movement (Fig. 7b).

Likewise, the door of the microwave oven can be controlled with two approaches. In the SI, pressing the trigger of the controller while touching the virtual door starts an animation that rotates the door to the fully open or closed position, depending on its initial state. In contrast, the user needs to keep holding the trigger and simultaneously move the door along with the controller in the case of NI.

6.2 Graphics Fidelity

The *Graphics Fidelity* represents an indication of the similarity between materials used to render the microwave oven and those of the physical appliance. Basic Unity *HDRP/Lit* shaders were created in the CBR style (6a), to match the diffuse (Albedo in the Unity terminology) colors from various parts of the virtual microwave oven (enclosure, glass door, handle, buttons, etc.) to those from the physical appliance,

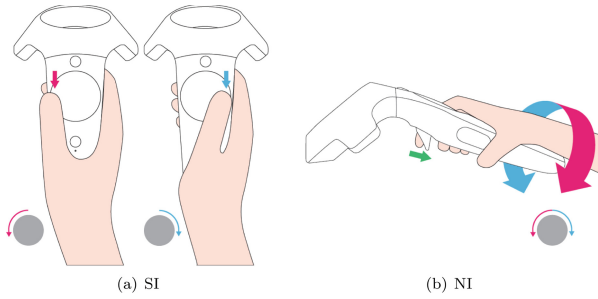


Fig. 5 Rotating the knob using the controller with different interactivity fidelities



Fig. 6 Graphics fidelity

whereas PBR simulates the physical properties of light (Pharr et al. 2016), for example, absorption and reflection. We used several PBR materials from the *Measured Material Library for High Definition Render Pipeline*⁶ to render the plastic, glass, and metal parts of the virtual twin (see Fig. 6b).

6.3 Apparatus and Implementation

We used the same process to implement the virtual twins and apparatus as in USER STUDY 1 (Sect. 3.2).

6.4 Demographics

Twenty participants (four females) were recruited for the laboratory-based user study aged from 25 to 38 (MEAN = 30.10, SD = 4.32). Eight participants had no experience with VR before. We assigned five participants to each of the four groups. Each

⁶ Library: <https://github.com/Unity-Technologies/MeasuredMaterialLibraryHDRP>.

participant was compensated with a ten € shopping voucher. The user study was approved by the Ethical Review Board of our institution.

6.5 Procedure

The procedure was similar to USER STUDY 1 (Sect. 3.4). Participants filled in three different questionnaires after completing all tasks with the assigned virtual microwave oven: SUS, NASA-TLX, Igroup Presence Questionnaire (IPQ) (Schubert et al. 2001; Regenbrecht and Schubert 2002). We used IPQ to rate the participants' perceived presence experience in VR. Then, a semi-structured interview was conducted. At the end, we asked them to experience the virtual twin in another combination of interactivity and graphics fidelity and express their preference.

7 Result of User Study 2

In this section, we report the qualitative data collected during the study, the identified usability issues and the user preferences for interactive fidelity.

7.1 SUS, NASA-TLX, IPQ

No significant difference was found between the SUS, NASA-TLX, and IPQ scores when performing evaluation neither with different levels of graphics or interactivity fidelity, nor with the different combinations of these fidelities.

7.2 Preference for Graphics and Interaction Fidelity

Participants showed different preferences for SI and NI in the door and knob manipulation. About 62.5% participants reported they preferred the button click instead of rotating the knob. Half of the participants stated they liked the pulling action for opening the door of the microwave oven.

Only *P10* noticed the graphics difference of the virtual replicas and commented “*this one (with PBR) looks like it has a better quality and is more expensive*”. The other participants reported they were “*focusing on the task*”, and the graphics difference is “*too slight to notice*”.

7.3 *Detected Usability Problems*

Only new unique usability problems, which were identified in each condition, are listed below.

UP17 *The microwave oven's door was not fully closed (SI only)*

UP18 *Mismatch between the actions in VR and real life (NI only)*

UP19 *Click left button to turn right (SI only)*

UP20 *Cumbersome rotation (NI only)*

UP21 *Inconsistency spatial presence experience (CBR only)*

UP22 *The indicator bar is hard to observe (PBR only).*

8 Discussion of User Study 2

8.1 *Interactive Fidelity*

According to the semi-structured interview, the continuous drag action in NI was rated by participants as “natural” while the click action in SI as “quick” by the participants. Our participants were split evenly on which of the two approaches they preferred.

The absence of haptic feedback induces certain misoperations in the NI, despite a sound being played by the virtual twin as feedback to indicate that the door has been closed. However, as on the real model, there is no other indication if the door is left open. Participants did not push the door in the closed position properly and incorrectly believed the door was closed from their visual perspective, because of the missing haptic feedback compared to the real-life equivalent, and resulted in the device fail to work. While in the SI, the whole open/close process is unambiguously completed through one click. So participants do not need to check the door status carefully. This, however, suggests that SI could cause UPs related to the manual operation to not be detected, as these actions are performed by the system instead of the user.

Participants reported that they preferred clicking the button in the SI over rotating the controller in the NI for turning the knob. Users need to operate the knob more frequently than opening the door in each task. This caused a user preference to appear. *P18* stated “*when clicking the button, it sounds like the clicking sound of gears biting together when rotating the knob in the real life*”. This insight suggests potential benefits in introducing multi-sensory (Obrist et al. 2017), pseudo-haptic (Pusch and Lécuyer 2011) in the future IVRE designs. *P9* also prefers the click action, which is “*easy and quick*”, while *P17* described the wrist-based rotation action as more “*natural*”.

In SI, we observed that when participants reached the knob rotation limit, they then wanted to rotate the knob counter-clockwise to reach the desired function on the

right side of the indicator bar (see Fig. 4). Participants clicked the right button that is actually used for clockwise turning. They were then stuck and complained that the system did not work as expected. After several tries, they figured out they needed to press the left button to reach the desired position on the right side. They rated this interaction as “*needing high mental demand*”.

Surprisingly, the SI has caused users to misunderstand the knob mechanism. Since the rotation was completed by clicking the trackpad, *P17* stated “*these should have been buttons, not rotary knobs*”. Furthermore, participants misunderstood how to operate the knob. Two participants attempted to press the rotary button to achieve their goal. *P15* explained that this action was based primarily on his/her everyday experience that rotary knobs of this shape are sometimes pressable. While the rotation action in NI is still different from itself in the real world, the NI does not affect how users understand the interacting object. The interaction fidelity has different impact on user experience, and preference varies from the tasks, in line with the work by Rogers et al. (2019).

8.2 Graphics Fidelity

In our study, the graphics fidelity did not affect participants’ opinion on the microwave oven’s usability. About 18 participants (8 PBR, 10 CBR) reported that the blurred icons on the microwave oven were difficult to observe, which forced them to bend down close to the microwave oven; this problem was caused by the low resolution of the VR headset and occurred in both graphics fidelities as in USER STUDY 1. In both graphics settings, participants also complained that icons were hard to recognize. This was caused by the design of the real microwave oven itself, as the icons did not express their meaning with sufficient clarity.

Despite the use of PBR to render the microwave oven, we used pictures of our laboratory as textures for the room used in the VE, instead of replicating a 3D version of all the furniture and equipment (see Fig. 7), since a VE that looked different from the physical room could have constituted a “Break-in-Presence” (Chertoff et al. 2008).

The initial purpose of setting up the VE was to replicate our real-world laboratory to give the glass material surroundings that could be reflected. During the study, *P6* was distracted by the reflection on the glass and started to focus on the VE. Then stated “*it’s strange, the wall is flat*”. Being unlit textures, the walls were not impacted by the lighting. The contrast between the high graphical fidelity of the PBR version contrasted with the lower fidelity of the environment, which was not noticed by the participants in the CBR. Future works should explore how the fidelity of the surrounding environment affects the focus where the interaction lies.

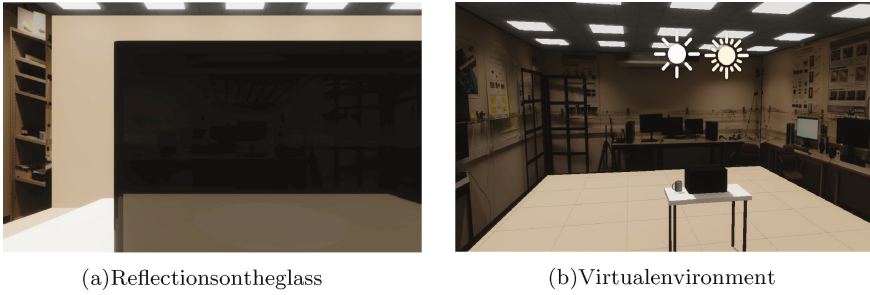


Fig. 7 Setting of the VE and its reflections on the glass

8.3 IVRE for Human-Centered Usability Evaluation

After comparing the usability problems found in both versions of the evaluations, we found that 61% of these affected both real and virtual prototypes. Crucially, this study was performed with precisely the intention of assessing the extent of this overlap. However, we ideally envision the IVRE to be performed *before* a physical prototype is implemented. In that case, if a physical equivalent of the virtual microwave had not yet been built, the IVRE would have uncovered numerous problems that could also have affected the physical version. Thus, 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.

We recommend product designers, researchers, and other stakeholders to consider the following guidelines 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. However, in line with previous findings from Voit et al. (2019), these are attributable to the VR interaction techniques, rather than the product itself. Evaluators should thus identify and categorize these problems accordingly and reflect on the likelihood of these interactivity issues affecting a physical prototype. Moreover, although both interactive variants of the fidelity led to the identification of a comparable numbers of usability problems, it is worth noting that those related to the manual operation of a device might not have been identified if it had been implemented through supernatural interaction.
- *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.
- *Use haptic proxies to uncover related problems.*
A problem that was uniquely identified in the RE (UP9) 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 (Bruno and Muzzupappa 2010; Nilsson et al. 2021; Simeone et al. 2015) to further enhance the fidelity of the interaction and uncover related problems in VR, future work should explore the inclusion of 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 (Simeone et al. 2017), 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., see UPI). Future work should also explore multi-sensory VR experiences, if relevant (Mahalil et al. 2020).
- *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 break*". 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.

9 Conclusion

In this chapter, we compared the results obtained after performing an evaluation based on the Think-Aloud Protocol on a real kitchen appliance, to those resulting from inspecting a virtual twin of the same everyday appliance with 24 participants. Furthermore, we analyzed the impact of graphic and interaction fidelity on evaluating the usability of virtual replicas based on a user study with 20 participants.

Results showed that there is a 61% overlap in terms of identified usability problems observed in both the real and virtual evaluations. The results highlight the potential of using Immersive Virtual Reality Evaluations (IVREs) to assess the usability of early VR concepts of physical artifacts and still obtain insightful results that can inform and improve the design of the physical version, which can also reduce the need to build expensive physical prototype. No significant effect of the graphics and interaction fidelity was found. However, participants detected unique usability problems under different settings. The different levels of interactive fidelity did not affect their understanding of the prototype operation in real life. The class of usability

problems related to manual operations of the interactive device could be missed if those actions are replaced by supernatural analogs.

However, IVREs are mediated by the VR interfaces used. Some usability issues uniquely found on the virtual twin are attributable to limitations of the VR technology itself, such as the low resolution which affected the interpretation of text and icons. The lack of haptic feedback in the VR condition did not allow to identify related problems that were only identified after inspecting the physical appliance. Thus, we suggest that future work should focus on studying the impact of the graphical and haptic fidelity on IVREs.

Acknowledgements This research is supported by Internal Funds KU Leuven (HFGD8312-C14/20/078).

References

- Alhadreti O, Mayhew P (2017) To intervene or not to intervene: an investigation of three think-aloud protocols in usability testing. *J Usability Stud* 12(3):111–132
- Alhadreti O, Mayhew P (2018) Rethinking thinking aloud: a comparison of three think-aloud protocols. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp 1–12
- Bangor A, Kortum P, Miller J (2009) Determining what individual SUS scores mean: adding an adjective rating scale. *J Usability Stud* 4(3):114–123
- Bellalouna F (2019) Vr-based design process of industrial products. In: *Proceedings of international conference on competitive manufacturing (COMA-19)*, South Africa, Stellenbosch, pp 239–245
- Benyon D (2019) *Designing user experience*, Pearson Educación. ISBN: 9781292155531
- Boring S, Jurmu M, Butz A (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*, pp 161–168
- Brooke J (1996) Sus: a 'quick and dirty' usability scale. *Usability evaluation in industry* 189
- Brooke J (2013) Sus: a retrospective. *J Usability Stud* 8(2):29–40
- Bruno F, Muzzupappa M (2010) Product interface design: a participatory approach based on virtual reality. *Int J Human-Comput Studi* 68(5):254–269
- Chertoff DB, Schatz SL, McDaniel R, Bowers CA (2008) Improving presence theory through experiential design. *presence: teleoperators and virtual environments* 17(4):405–413
- Cools R, Han J, Adalberto SL (2021) Selectvisar: selective visualisation of virtual environments in augmented reality. In: *Designing interactive systems conference (New York, NY, USA), DIS '21*, Association for Computing Machinery, pp 275–282
- Dumas JS, Dumas JS, Redish J (1999) *A practical guide to usability testing*. Intellect Books
- Ericsson KA, Simon HA (1984) *Protocol analysis: verbal reports as data*. The MIT Press
- Falcão CS, Soares MM (2013) Application of virtual reality technologies in consumer product usability. In: Marcus A (ed.) *Design, User Experience, and usability. Web, mobile, and product design* (Berlin, Heidelberg). Springer Berlin Heidelberg, pp 342–351
- Følstad A, Hornbæk K (2010) Work-domain knowledge in usability evaluation: experiences with cooperative usability testing. *J Syst Softw* 83(11):2019–2030. Interplay between Usability Evaluation and Software Development
- Georgiou Y, Eleni KA (2017) The development and validation of the ARI questionnaire: an instrument for measuring immersion in location-based augmented reality settings. *Int J Human-Comput Stud* 98:24–37

- Hart SG (2006) NASA-task load index (NASA-TLX); 20 years later. *Proc Human Fact Ergonom Soc Ann Meet* 50(9):904–908
- Hassenzahl M, Burmester M, Koller F (2003) AttrakDiff: a questionnaire to measure perceived hedonic and pragmatic quality. *Men Comput* 57:187–196
- Hoggenmüller M, Tomitsch M, Hespagnol L, Tran TTM, Worrall S, Nebot E (2021) Context-based interface prototyping: understanding the effect of prototype representation on user feedback. In: *Proceedings of the 2021 CHI conference on human factors in computing systems* (New York, NY, USA), CHI '21, Association for Computing Machinery
- Hwang W, Salvendy G (2010) Number of people required for usability evaluation: the 102 rule. *Commun. ACM* 53(5):130–133
- Kreylos O (2016) Optical properties of current vr hmds
- Laugwitz B, Held T, Schrepp M (2008) Construction and evaluation of a user experience questionnaire. In: *Holzinger A (ed) HCI and usability for education and work* (Berlin, Heidelberg). Springer, Berlin Heidelberg, pp 63–76
- Lewis JR (1995) IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int J Human-Comput Interact* 7(1):57–78
- Lewis JR (2002) Psychometric evaluation of the PSSUQ using data from five years of usability studies. *Int J Human-Comput Interact* 14(3–4):463–488
- Lim Y, Pangam A, Periyasami S, Aneja S (2006) Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices. In: *Proceedings of the 4th nordic conference on human-computer interaction: changing roles* (New York, NY, USA), NordiCHI '06, Association for Computing Machinery, pp 291–300
- Lubos PB (2018) Supernatural and comfortable user interfaces for basic 3d interaction tasks. Ph.D. thesis, State and University Library Hamburg Carl von Ossietzky
- Mahalil I, Yusof AM, Ibrahim N (2020) A literature review on the effects of 6-dimensional virtual reality's sport applications toward higher presence. In: *2020 8th International conference on information technology and multimedia (ICIMU)*, pp 277–282
- Mäkelä V, Radiah R, Alsharif S, Khamis M, Xiao C, Borchert L, Schmidt A, Alt F (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* (New York, NY, USA), CHI '20, Association for Computing Machinery, pp 1–15
- Mathis F, Vaniea K, Khamis M (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* (New York, NY, USA), CHI '21, ACM
- McMahan Ryan P, Bowman Doug A, Zielinski David J, Brady Rachael B (2012) Evaluating display fidelity and interaction fidelity in a virtual reality game. *IEEE Trans Visual Comput Graph* 18(4):626–633
- McMahan RP, Lai C, Pal SK (2016) Interaction fidelity: the uncanny valley of virtual reality interactions, virtual, augmented and mixed reality (Cham) (Stephanie Lackey and Randall Shumaker, eds.), Springer International Publishing, pp 59–70
- Nilsson NC, Zenner A, Simeone AL (2021) Propping up virtual reality with haptic proxies. *IEEE Comput Graph Appl* (in press), to appear
- Obrist M, Gatti E, Maggioni E, Thanh Vi C, Velasco C (2017) Multisensory experiences in HCI. *IEEE MultiMedia* 24(2):9–13
- Paneva V, Bachynskyi M, Müller J (2020) Levitation simulator: prototyping ultrasonic levitation interfaces in virtual reality. In: *Proceedings of the 2020 CHI conference on human factors in computing systems* (New York, NY, USA), CHI '20, Association for Computing Machinery, pp 1–12
- Pharr M, Jakob W, Humphreys G (2016) From theory to implementation, Morgan Kaufmann, Phys Based Rendering
- Pusch A, Lécuyer A (2011) Pseudo-haptics: from the theoretical foundations to practical system design guidelines. In: *Proceedings of the 13th international conference on multimodal interfaces* (New York, NY, USA), ICMI '11, Association for Computing Machinery, pp 57–64

- Regenbrecht H, Schubert T (2002) Real and illusory interactions enhance presence in virtual environments. *Presence* 11(4):425–434
- Rogers K, Funke J, Frommel J, Stamm S, Weber M (2019) Exploring interaction fidelity in virtual reality: object manipulation and whole-body movements. In: *Proceedings of the 2019 CHI conference on human factors in computing systems* (New York, NY, USA), CHI '19, Association for Computing Machinery, pp 1–14
- Saffo D, Di Bartolomeo S, Yildirim C, Dunne C (2021) Remote and collaborative virtual reality experiments via social vr platforms. In: *Proceedings of the 2021 CHI conference on human factors in computing systems* (New York, NY, USA), CHI '21, Association for Computing Machinery
- Schneider D, Otte A, Kublin AS, Martschenko A, Kristensson PO, Ofek E, Pahud M, Grubert J (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)*, pp 804–805
- Schubert T, Friedmann F, Regenbrecht H (2001) I group presence questionnaire. *Teleoperators Virtual Environ* 41:115–124
- Simeone AL, Mavridou I, Powell W (2017) Altering user movement behaviour in virtual environments. *IEEE Trans Vis Comput Graph* 23(4):1312–1321
- Simeone AL, Velloso E, Gellersen H (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* (New York, NY, USA), CHI '15, Association for Computing Machinery, pp 3307–3316
- Slater M, Khanna P, Mortensen J, Yu I (2009) Visual realism enhances realistic response in an immersive virtual environment. *IEEE Comput Graph Appl* 29(3):76–84
- Slater M, Steed A (2000) A virtual presence counter. *Presence: Teleoperators Virt Environ* 9(5):413–434
- Tao F, Zhang H, Liu A, Nee AYC (2019) Digital twin in industry: state-of-the-art. *IEEE Trans Ind Inf* 15(4):2405–2415
- Tullis TS, Stetson JN (2004) A comparison of questionnaires for assessing website usability. *Usability professional association conference*, vol. 1, Minneapolis, USA, pp 1–12
- Virzi RA (1992) Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors* 34(4):457–468
- Voit A, Mayer S, Schwind V, Henze N (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* (New York, NY, USA), CHI '19, Association for Computing Machinery, pp 1–12
- Ye J, Badiyani S, Raja V, Schlegel T (2007) Applications of virtual reality in product design evaluation. In: Jacko JA (ed) *Human-computer interaction. HCI applications and services* (Berlin, Heidelberg), Springer, Berlin Heidelberg, pp 1190–1199
- Yu I, Mortensen J, Khanna P, Spanlang B, Slater M (2012) Visual realism enhances realistic response in an immersive virtual environment—part 2. *IEEE Comput Graph Appl* 32(6):36–45
- Yu R, Bowman DA (2018) Force push: exploring expressive gesture-to-force mappings for remote object manipulation in virtual reality. *Frontiers in ICT* 25
- Zaharias P, Poylymenakou A (2009) Developing a usability evaluation method for e-learning applications: beyond functional usability. *Int. J Human-Comput Int* 25(1):75–98
- Zhao T, McDonald S, Edwards HM (2014) The impact of two different think-aloud instructions in a usability test: a case of just following orders? *Behav Inf Tech* 33(2):163–183