Using Heuristic Evaluation in Immersive Virtual Reality Evaluation

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

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

ABSTRACT

Previous works show that virtual reality itself can be used as a medium in which to stage an experimental evaluation. However, it is still unclear whether conventional usability evaluation methods can directly be applied to virtual reality evaluations and whether they will lead to similar insights when compared to equivalent real-world lab studies. Therefore, we conducted a user study with nine participants, comparing Heuristic Evaluation (HE) for the evaluations of a novel smart artefact. We asked participants to evaluate the physical prototype and their virtual counterparts in the real-world and the virtual environment, respectively. Results show the HE have similar performance when evaluating artefacts usability in VR and real-world in terms of identified usability problems. The VR implementation has an impact on the immersive VR evaluation result.

CCS CONCEPTS

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

KEYWORDS

VR, Usability Evaluation, Heuristic Evaluation

ACM Reference Format:

Xuesong Zhang and Adalberto L. Simeone. 2021. Using Heuristic Evaluation in Immersive Virtual Reality Evaluation. In *20th International Conference on Mobile and Ubiquitous Multimedia (MUM 2021), December 5–8, 2021, Leuven, Belgium.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3490632.3497863

1 INTRODUCTION AND STATE-OF-THE-ART

Using Virtual Reality (VR) as a proxy to perform usability testing has already gained researchers attention [1, 5–7, 12], as we can simulate the digital twin of any device, even before implementing a physical prototype with novel technology. However, it is still unclear whether conventional usability evaluation methods can directly be applied to VR evaluations and whether they will lead to similar insights when compared to equivalent real-world lab studies. To answer this research question, we designed and conducted a user study to test the performance of evaluating a prototype of an imagined future technology in the real life and an immersive virtual environment by using a *Heuristic evaluation (HE)*. HE is a popular usability inspection method [2], where an expert applies a set of heuristics principles as a usability inspection method to evaluate an existing design and identify any usability problems [8, 10]. A "health-box" is a smart device, which is assumed to be capable of detecting the correctness of the user's posture, as well as their hydration level and exercise frequency, since an unbalanced posture, dehydration, prolonged sitting can be harmful to our health [11].

The contribution of this work is two-fold: (1) We show the potential and discuss the results of using HE for performing usability evaluations in immersive VR. (2) We discuss which factors may affect the identification of usability problems in immersive VR evaluations.

MUM 2021, December 5–8, 2021, Leuven, Belgium

ACM ISBN 978-1-4503-8643-2/21/12.
https://doi.org/10.1145/3490632.3497863

2 USER STUDY

This user study follows a between-subjects design. For each group, the independent variable is Environment: {Real environment (RE), Virtual environment (VE) }.

2.1 Task

During the experiment, participants sit in front of a (virtual) desktop. They need to complete two English Tests during the experiment, once with responding to the health box and once without responding to the health box. Each test contains 25 single choice questions with the same difficulty level. The order of the two English tests and with or without responding the health box was counterbalanced.

In both settings, the health box is placed under the monitor and has three lights on it. When users deviate from the standard posture, the first light turns on. The second and third lights turns on at custom intervals to remind users to do some physical exercises or hydrate themselves. During the study, users acknowledge the exercise by tapping the health box. Likewise, users can touch the mug to indicate the drinking action. Once the user performs the action, the corresponding light will turn off. The functions are implemented via a Wizard of Oz (WoZ) [3] method.

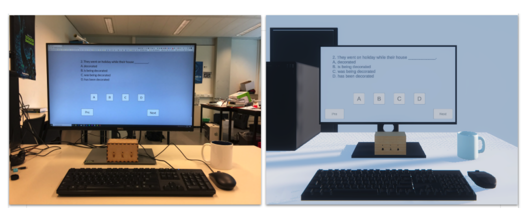

Figure 1: Participants view in RE and VE

2.2 Demographics and Procedure

Because of the Covid-19 pandemic, our options for recruiting participants were severely limited. We recruited nine participants (3 females, 6 males) aged from 21 to 28 ($MEAN = 24.89$, $SD = 2.37$) through the department mail list and word of mouth. Five participants evaluated the virtual health box remotely with HTC Vive Pro ($N = 3$) and Oculus Rift ($N = 2$), and four inspected the physical health box in the VR lab. All participants are fluent in English and had prior experience in VR and knowledge of fundamental HCI concepts, as HE should be carried out by usability experts [9].

After filling a consent and a demographics form, we introduced participants to the health box and the HE evaluation method. Then participants started the task Figure 1. After completing the English Test, participants inspected the health box with a set of heuristics proposed by Jakob Nielsen [8]. Finally, we conducted a semi-structured interview to elicit participants' additional insights on the health box's advantages and disadvantages, existing usability problems, suggestions for improvement, and opinions on the evaluation method and its virtual implementation.

3 RESULT

We asked users to inspect the health box using the HE and report any problems they found. We categorise usability issues according to the frequency with which they are mentioned, the results are in Table 1 in the appendix.

We assessed user experience on 5-point Likert scales. A Kruskal-Wallis H test was run on the collected data. The level of enjoyment was similar ($p = 0.304$) across the VE ($MEAN = 3.58$) and the RE ($MEAN = 3.11$). There is no evidence ($p = 0.129$) that the level of attention was significantly different between the VE ($MEAN = 3.47$) and the RE ($MEAN = 2.77$). Participants rated the annoyance of using the health box in VE ($MEAN = 2.78$), and they feel more annoyed ($p = 0.385$) with health box in RE ($MEAN = 3.22$). Participants found the device to be more sensitive ($p = 0.662$) in VE ($MEAN = 3.31$) than in RE ($MEAN = 3.22$).

We calculated the difference in the number of incorrect answers for each participant on the two English tests.We compared this difference between using the health box in VE ($MEAN = 0.17$) and in RE ($MEAN = 1.37$) with a Kruskal-Wallis H test. No statistically difference ($p = 0.276$) was found.

4 DISCUSSION

Results of the user study show that eight out of nine usability issues present on the virtual prototype are also identified on the equivalent physical prototype. Four usability problems reported in the VE were not found in the RE (Problem 2, 7, 10, 11 in the Table 1), these were perhaps caused by the VR implementation, discussed successively. Only one problem was not revealed in the VE. Indeed, in the RE only one participant reported that the light was distracting. However, *"it's hard to notice the light"* was also reported twice. These two problems contrast with each other. We assume this is caused by the individual difference in perception of the environment.

Similar to the findings by Mäkelä [5], we observed that participants reported problems more actively when they interacted with virtual counterparts in the third user study. The VR system engaged participants more. Two participants actively looked around to observe the VE during the test and reported suggestions on how to improve the environment design.

We also noticed the impact of the VR implementation on the evaluation results. The hardware limitations resulted in usability problems associated with the HMD rather than with the evaluated artefact (in line with findings from [12]).

Four participants in the VE reported blurred icons as a usability problem, however, this is caused by the resolution of the HMD used (listed in the Table 2 in the appendix). They also confirmed those were false positives, as they were not applicable in the RE. The low resolution and narrow FoV of the headset did affect the use experience. Although the objects size and distance in the virtual scene are replicated from the real life measurements, two participants reported that they could not see the health box during the test. Indeed, human eyes have a horizontal FOV of about 200 degrees and a vertical FOV of over 120 degrees. And the HMDs FoV is narrower than that.

5 CONCLUSION

In this paper we investigated the performance of applying Heuristic Evaluation in immersive VR evaluations. We designed and conducted a user study, where nine participants evaluated a virtual and physical versions of a near-future imagined technology (a health box).

Using the heuristic evaluation, the most usability problems present on a virtual equivalent are also identified on the physical prototype. We noticed that the implementation of the virtual counterparts did affect the experts' evaluation results. The low resolution and limited field of view of VR headsets may lead inspectors to report false positives.

Thus we suggest that future work should focus on the impact of the representation fidelity in immersive virtual evaluations, in terms of both the graphical visual aspect.

ACKNOWLEDGMENTS

We thank all participants for taking part in the study and all anonymous reviewers for their helpful comments. The work was funded by the KU Leuven Internal Funds (3E18003).

REFERENCES

[1] 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.

[2] Jarinee Chattratichart and Gitte Lindgaard. 2008. A comparative evaluation of heuristic-based usability inspection methods. In *CHI'08 extended abstracts on Human factors in computing systems*. 2213–2220.

[3] John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)* 2, 1 (1984), 26–41.

[4] Oliver Kreylos. 2016. *Optical Properties of Current VR HMDs*. http://doc-ok.org/?p=1414

[5] 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*. 1–15.

[6] 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

[7] 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, CHI*, Vol. 21.

[8] Jakob Nielsen. 1992. Finding Usability Problems through Heuristic Evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 373–380. https://doi.org/10.1145/142750.142834

[9] J Nielsen. 1992. Usability problems through heuristic evaluation usability walkthroughs. In *Proceedings of ACM CHI 1992 Conference on Human Factors in Computing Systems*. 373–380.

[10] Jakob Nielsen and Rolf Molich. 1990. Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 249–256. https://doi.org/10.1145/97243.97281

[11] Mark Stephen Tremblay, Rachel Christine Colley, Travis John Saunders, Genevieve Nissa Healy, and Neville Owen. 2010. Physiological and health implications of a sedentary lifestyle. *Applied physiology, nutrition, and metabolism* 35, 6 (2010), 725–740.

[12] 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*. ACM, 507.

A APPENDIX

Nr	Problem description	VE	RE
1	The light size is too small	2	1
2	The light is not bright enough	2	-
3	It's hard to notice the light	5	2
4	The health box is out of view field	2	-
5	No idea how to correct the head position	2	3
6	Limited space/ posture detection is too sensitive	4	1
7	Blur icons	3	-
8	Icons are ambiguous to understand	2	2
9	Light notification is too dim	3	2
10	Icons are too small	1	-
11	The health box should be placed somewhere else	1	-
12	No handy documentation to refer	3	4
13	The light is distracting	-	1
14	Forgot how to react to the light during the test	1	2

Table 1: List of participants reported the usability problems in VE and RE and the the mentioned times

HMD	Resolution per eye	FoV
Oculus rift	1080×1200	94
HTC Vive	1080×1200	110
HTC Vive Pro	1440×1600	110

Table 2: Resolution and FoV of HMDs [4]