COMPARATIVE MEASUREMENT OF TASK LOAD IN VR, AND THE REAL WORLD USING NASA-TLX

Ahmad CHOUEIB

JUNE 2022

COMPARATIVE MEASUREMENT OF TASK LOAD IN VR, AND THE REAL WORLD USING NASA-TLX

A THESIS SUBMITTED TO THE

GRADUATE SCHOOL

OF

BAHÇEŞEHİR UNIVERSITY

BY

AHMAD-CHOUEIB

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF MASTER OF GAME DESIGN

IN THE DEPARTMENT OF DIGITAL GAME DESIGN

JUNE 2022

T.C.

BAHÇEŞEHİR UNIVERSITY

GRADUATE SCHOOL

…./……../…..

MASTER THESIS APPROVAL FORM

Program Name:	GAME DESIGN
Student's Name and Surname:	AHMAD CHOUEIB
Name of The Thesis:	COMPARATIVE MEASUREMENT OF TASK LOAD IN VR, AND THE REAL WORLD USING NASA-TLX
Thesis Defense Date	June 7th, 2022

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Prof. Dr. Ahmet ÖNCÜ

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title, Name	**Signature**
Thesis Advisor:		
2. Member:		
3. Member:		

ETHICAL CONDUCT

I hereby declare that all information in this document has been obtained and presentedin accordance with academic rules and ethical conduct. I also declare that, as requiredby these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname :

Signature :

ABSTRACT

COMPARATIVE MEASUREMENT OF TASK LOAD IN VR, AND THE REAL WORLD USING NASA-TLX

Choueib, Ahmad

Game Design Masters Program

Supervisor: Dr. Instructor Member Mehmet İlker Berkman

June 2022, 53 pages

User performance and experience is compared in an HMD VR puzzle game and its real-world toy replica, with 28 participants in a within subjects experimental design. Mean comparisons on task time and success rate do not have a significant difference, as well as the subjective task load and task performance obtained through NASA-TLX. The subjective game user experience measures also did not reveal any significant difference between the toy and VR gameplay, except the Enjoyment measures via Game User Experience Satisfaction Scale. This difference can be explained through a novelty effect, since most of the participants had no or limited engagement with VR. Around one thirds of the participants responded the Spatial Presence Experience Scale for both VR and toy gameplays, suggesting that presence can be measured for real-world mediated experiences engaged with toys, and the presence measures also did not reveal any significant differences between conditions. Findings suggest that the VR gaming experiences can replace with the real-world game-play, within the context of a cognitive tasks in a limited representation of spatial surrounding. The interactions with the nearby objects in a HMD-based virtual environment and real-world lead to very similar measurements of user performance and experience.

Keywords: VR, HMD, NASA-TLX, virtual environment, real-word.

ACKNOWLEDGEMENT

I would like to thank my esteemed supervisor Dr. Mehmet İlker Berkman for his invaluable supervision and guidance during my Master thesis. I would also like to express my gratitude to my family and everyone who showed me support through my journey.

Name and Surname

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF PICTURES

PICTURES

LIST OF SYMBOLS AND ABBREVIATIONS

HMD	Head-Mounted Display
HMI	Human-Machine Interface
NASA	National Aeronautics and Space Administration
TLX	Task Load Index
PS	Play Station
2D	Two-Dimensional
UI	User Interface
UX	User Experience
VR	Virtual Reality

Chapter 1

Introduction

Technology is becoming more essential by the day, and industries are becoming dependent on solutions generated by mediums such as Virtual Reality (VR), as it became economically affordable in terms of hardware to individual users. Also, the fact that virtually all mainstream tech giants released their own VR product, even linking it to their basic products such as Samsung's phones, and Sony's PS console. Since VR technology is becoming mainstream (Jerald, J., 2015), entertainment is not the only purpose of its use anymore. Examples can be found in an array of studies from studies that used VR as a tool of training of medical staff (Huber et al., 2017), as a therapy method (Krijn et al., 2004), to simulation to assess performance (Stevens & Kincaid, 2015). Accessible tools for constructing VR applications allow an increasing source of developing tools to further expand the potential use of this technology in different areas.

Which opened up VR as a reliable tool for researchers to conduct experiments, that help to learn more about subjects such as presence (Sanchez-Vives & Slater, 2005), immersion (Bowman & McMahan, 2007), and performance (Lackey et al., 2016). Different studies have tackled the topics of Head-mounted display (HMD), and User Interface (UI)-based mediums such as 2D desktop, phones, or CAVE VR, and so on. However, the subject of comparing VR in HMD with real life is also a subject that was studied over the years since the early days of VR up until recent time (Moghimi et al., 2016; Didomenico & Nussbaum, 2008). A bulk of these studies focus on the subject of presence and immersion. There is another approach that compares VR to other mediums in terms of user experience (Kuliga et al., 2015). Furthermore, the study of Workload using VR versus Real life experience, which used subjective performance scales such as NASA-Task Load Index (NASA-TLX) has also been the subject of study for many researches (Didomenico & Nussbaum, 2008).

In this regard, research-based comparisons between with virtual reality experiences and real world had been made since the very early days of VR. With the advances in HMD hardware, some of the comparative studies focused on fidelity and acuity of the virtual environment. Best examples of visual fidelity can be found in field

of architectural lighting (e.g. Jin et al., 2021; Krupiński, 2020; Chen et al., 2019; Chamilothori et al., 2019) but not limited to (e.g. Maffei et al, 2016; Kuliga et al.,2015; Bishop & Rohrmann, 2003). Some studies focus on real-world learning outcomes of VR in terms of training and skill acquisition (e.g. Cooper et al., 2021; Hejtmanek et al., 2020; Kim et al., 2019; Michalski et al., 2019; Lackey et al., 2016; Lloyd et al., 2009; Rose et al., 2000). There is another area of research that focuses on human perception and behaviour in real and virtual environments. While some of those studies focus purely on perception (e.g. Abouelkhier et al., 2021; Bhargava el., 2020) or behavior (Li et al., 2019; Villani et al., 2012; Kort et al., 2003), many studies, which are given in detail in Related Studies section, report performance metrics and subjective assessments of participants regarding their performance and experience. These studies evaluate spatial issues like navigation or locomotion performance, or other interactions that require manipulating the objects in the virtual environment. The real-world tasks and their simulations in these studies mostly demand spatial skills for navigation in terms of mental effort, along with certain amount of physical motor skills. Some task scenarios also include time pressure.

This study is comparing the real-world task performance with VR environment task performance on a predominantly mental task, where users need to solve a puzzle by asking questions and following instructions. The game “Keep Talking and Nobody Explodes” (Steel Crate Games, 2015) is selected as a VR task, which is also available in PC and mobile platforms. Thus, further comparisons can be made between the HMD-based VR and other media, as well as real-world performance.

Unlike many other studies, we created the in-game digital artifacts and environments in the real-world, instead of creating a digital environment imitating the real world. A toy bomb with an electronic circuit board is manufactured which is a replica of the digital asset in the game. A within-subjects study is administered to compare subjective player performance as game achievement and duration, participants’ reflection on their performance, and their gameplay experience including presence. It also focuses on comparing task load in VR and real-life which was conducted in multiple works recently (Villani et al., 2012; Moghimi et al., 2016; Petukhov et al., 2020; Maffei et al., 2016), that aims to measure and compare the task load in the real world and VR. This was made possible by the usage of a multidimensional scale that endured the test of time and field of research that is NASA-

TLX which is a measurement tool that can differentiate the variable factors of demand, effort, and abilities in an assigned mission to be completed (Hart, 2006). By the usage of a self-reported multidimensional demand scale, this may prove to be useful to clarify the process of simulating a task.

Chapter 2

Literature Review

2.1. Virtual Environments (VE) and User Experience (UX) Evaluation

VR is widely used in a variety of fields. It has been used as a study tool for subjects such as navigation in airplanes and spatial cognition, surgical and medical training, and skill training (Cliburn et al., 2007; Loomis et al., 1999; Ruddle et al., 1997; Larsen et al., 2012). The studies that compare digital gaming and real-world toys focus on the social interaction (e.g. Ewin et al., 2021; Ho et al., 2018) or learning outcomes (Hinske et al., 2010) rather than the game user experience. Besides, there is not a study that compares the gaming experience on an electro-mechanical to with its digital counterpart. Direct comparisons between real and virtual environments were studied to find out how emotional and cognitive environmental assessment, as well as human movement patterns, operated in both (Suneson et al., 2006; Bishop & Rohrmann, 2003; Haq et al., n.d.; Witmer, 1998). In terms of understanding the behavioral aspect of an individual while undergoing a task, VR allows an accurate observation and recording of a participant's course of action. In some studies, the conclusion was that people act on cues and decide in virtual mediums similarly to the way they navigate their choices of action (Skorupka, 2009), while other works found that this match is not valid (Haq et al., 2005); the difference in the result can be linked to the level of realistic, visual of the virtually simulated environment examined in these studies, meaning that the level these experiences in the virtual environment renders a real-world experience (Kort et al., 2003).

The issue of simulating a realistic virtual environment and interaction opens up a challenge of replicating the real-world factors that can be defined as a user's experience, which is a negative factor when the aim is to isolate a task in a virtual experience. However, immersive technologies can now offer visual renders that are very similar to real-life visual experiences. Some tools, such as HMD CAVEs can reflect perceptual and physical responses that resemble real-world user performance, by providing the user with instant feedback on changes in viewing directions, head, and body rotations. Despite the fact that some of the aspects of a user's

experience in real and virtual environments can be transferable in immersive technology mediums, high levels of experiential realism, such as presence (the subjective, psychological state of being in one place while physically being in another can be attained by replicating a specific objective task in both mediums (Witmer & Kline, 1998).

User experience in virtual environments can be challenging to engage a real-life like experience. In human–computer interaction, the term user experience is defined by Hassenzahl (n.d.) as an interplay of individual perception, emotion, cognition, motivation and action, acting in dialogue with the world (place, time, people, and objects). However, with the rising popularity of interactive technological products, the attention to user experience (UX) is an aspect of study that researchers are trying to get closer to a still missing universal definition for it (Hassenzahl, n.d.). This subject attracted the scientific community's attention to the subject of UX as well, which made this subject a motive human–computer interaction research to develop a deeper understanding than the usability and task-oriented values. Although both aspects are frequently correlated, while low usability levels can limit performance, high usability levels is not necessarily enough to create a good UX.

Taking into consideration that UX is an outcome of interaction between a user and a product within a physical, social, and cultural context, "researchers must be aware that, depending on the context at hand, users can have different experiences with the same product" (Rebelo et al., 2012). Therefore, the ability to find and create proper contexts remains a serious challenge. The evaluation of UX is a subjective task since these characteristics are based on a user's emotional state, which in turn is related to an emotional response that is consequently a subjective experience and behavior (Larsen & Prizmic-Larsen, 2006).

So, the question would be if VR is a reliable tool to evaluate UX in tasks despite the method of choice to assess the subjective performance. This would require us to first see what VR can offer, in very simplified terms, VR transports a user to a reality (virtual environment) in which they are not physically present but feels like they are there. By definition, VR is the use of computer modeling/simulation that enables a user to interact with an artificial three-dimensional sensory environment. This opens up the way to multidimensional experiences. The advantages of using VR for research on UX

can be split to three main topics: availability, safety, and data provision (Rebelo et al., 2012), meaning that VR can be a tool to use in UX evaluation for processes and products that can have the controlled conditions within the virtual environment (VE) that the VR system offers (Villani et al., 2012).

2.2. Workload Measurement in VR and Reality

Workload was defined as the cost an individual, considering their abilities, while accomplishing a certain level of performance on a given task with specific demands (Hart & Staveland, 1988). Workload evaluation is an important part of system design and analysis. Subjective workload measurements use numerical ratings to attempt to assess the energy invested during task performance. The natural world is reflected in an individual's subjective report of perceptions related to physical or mental demand on their mental and physical resources (Annett, 2002).

Workload being subjective in assessments since it is based on an individual's personal feelings and perceptions, needing the individual to provide their own judgments of efforts associated with performance in a assigned task (Eggemeier et al., 2020). The effect of physical demand is linked to the awareness of self-monitoring while undergoing the activity (Mihevic, 1981). Subjective feedback are influenced by the individual's objectives during a task, as well as motives, and plans (Annett, 2002), yet it depends on the person's ability to scale their sensations to a quantitative rating (Noble et al., 1983). Physical and mental demands vary in response to individual differences, which form a challenge in understanding the measure of workload levels (Kahneman, 1973).

As a medium to be used in order to compare real life, VR simulation training is a reliable option due to its common use in different fields ranging from medical (Gunn et al., 2018), astronautic studies (Everson et al., 2018), education (Greenwald et al., 2017), industrial (Berg & Vance, 2017), military, sports, architecture, and so on. This shows us that VR is a tool compatible with almost any task in terms of a user's practical activity monitoring and measurement. The unique aspects of VR that differentiate it from other user interface systems that are 2D is the nature of the setup that allows the

deep psychophysiological action (Bowman, 1998) which includes a multidimensional experience in terms of physical and mental demands. Moreover, the VR application provides a wider variety of interaction techniques other platforms lack (Mania & Chalmers, 2001), and the most distinctive trait of VR is the 3D interactivity and environments (Mine et al., 1997).

This makes the measurement of workload in a VR medium of a given task an objective to various professions from designers, manufacturers, managers, to operators, that are interested in performance assessment, and seek answers about the operator workload during different phases of system design and operation.

As shown above, the VR immersive technology is a reliable tool to embody the multidimensional task accomplishment factors.

2.3. Subjective Assessment of Workload and NASA-TLX

This introduces the tool developed by the National Aeronautics and Space Administration (NASA). The NASA-TLX was constructed to serve as a multi-dimensional scale designed to obtain workload estimates operators while they are performing a task or immediately afterwards (Hart & Staveland, 1988).

In a study that conducted an analysis of 556 papers on TLX (Hertzum, 2020) the criteria of selecting the papers insured all studies defined TLX in the same way.

The initial sum of papers collected was 2769, but only 556 got analyzed since they included all 6 subscales of the task load index. Four steps were used to analyze the data, across different domains, from technologies, regions, and real life/lab settings. Technologies got divided to five groups: handheld devices, desktop applications, environments, virtual reality, and other.

Handheld devices, desktop applications, environments, and virtual reality were investigated in 1024 conditions from 332 studies. The technologies were spread unevenly across domains, virtual reality in education got 28% of the technology category. This averaging served to make the analyses independent of how many conditions each study had.

The study found that the use of TLX for measuring workload increased over the 30-year period. Half of the included papers were published during the last four years of the period. The 556 studies reported TLX data from a total of 27616 participants

Originally measures of perceived usefulness and perceived ease of use of the VR during the task were used to evaluate the subjective experience, as well as performance metrics. In VR, the cognitive workload can be simulated of conditions, as it depends on the mental load demand, but collects the physical and temporal factors. This makes the NASA-TLX (Hart & Staveland, 1988; Hart, 2006) applicable in the assessment of the participants' subjective workload experience during VR sessions. NASA-TLX consists of six subscales that represent somewhat independent clusters of variables: Mental, Physical, and Temporal Demands, Frustration, Effort, and Performance. These scales hold the assumption that it can represent the "workload" experienced by most people performing most tasks (Hart, 2006). These scales are rated individually on a 0–100 scale, which allows the research to breakdown the task load into multiple dimensions and observe the way a task is conducted in a VR medium with respect to the physical world. The importance of this tool in the process of measurement of a workload in a given task can be seen in the adaptations it earned and the wide language range it was translated into (Hart, 2006). The common activities across controlled mediums and the real world allow such well-designed measurement tool. Other methods of task load and performance evaluation ranging from surveys to questionnaires were originally designed for 2D interface and setup (Lin & Hsu, 2017).

2.4. Related Work

2.4.1. Performance in VR vs real world.

There are several different foci of the studies that compare performance in real-world and virtual environments. Some studies focus on performance in spatial tasks such as wayfinding, which showed that performance in real world requires shorter time compared to virtual environment, while different paths are taken in each (Skorupka, 2009). Another aspect compared is locomotion, which shows that walking in VR is slower compared to the real world, but faster compared to the previous studies, which is probably due to the improvements in VR hardware (Agethen et al. 2018;), or

orientation (Kimura et al., 2017). Some compare task performances based on objective measures referring efficiency and effectiveness; in our experiment, participants in VR were no different from participants in the real world in terms of aperture passability judgments, accuracy of judgments, and certainty of judgments. Comparisons also showed that participants in real world improved in accuracy over time, while participants in VR did not. In addition, mixed results were observed with regard to some performance metrics, as some were better for simulator due to the design of the simulation (Bhargava et al., 2020; Ovaskainen, 2005; Kenyon & Afenya, 1995).

There are several studies that employ NASA-TLX for self-assessment of perceived performance. Stone et al. (2011) reported that Mental Demand measured via NASA-TLX was not significantly different for welding trainees using real-world welding equipment and VR training environment. While the training times in VR was significantly shorter, the average muscle activity was similar for both conditions. On a wheelchair driving task, Kamaraj et al. (2016) revealed significantly lower Mental Demand and Frustration scores for VR compared to real-world, which resulted with a lower adjusted overall NASA-TLX score for VR. Mouraviev et al. (2016) reported similar raw NASA-TLX scores for VR and porcine-model robotic surgery, except for Frustration, which is higher for VR. Narasimha et al. (2018) compared a collaborative card sorting task in desktop, HMD and real-world paper environments and did not observe a significant difference in time, match with master set (effectiveness), perceived presence and usability. However, the NASA-TLX score is significantly lower for Performance score in VR condition, resulting with a significantly better overall score. George et al. (2019) explored a smart home door authentication task in VR and real-world equipment, revealing that although the users performed significantly worser in VR in terms of time used on task, there is not a significant difference on NASA-TLX scores. Chang et al. (2019) findings for paediatric resuscitation is very interesting, suggesting that physicians scored higher significantly for their Temporal Demand, Frustration and Performance for VR simulation leading to higher overall score, compared to real-world emergency department resuscitations. However, heart rate and cortisol levels, which are stress indicators, were higher for real-world experiences. For a pointing task, Rizzuto et al. (2019) did not observe any difference on wrist, elbow or shoulder joint angle at the end state, while pointing accuracy is significantly higher in real world condition, with a shorter movement time.

Task Load Index scores were significantly higher in one virtual condition (VEA) compared to the real condition. For a throwing task, Zindulka et al. (2020) did not observe any significant difference on any of the 6 NASA-TLX variables, but throwing in real-world is almost twice as accurate and 2-3 times more precise than throwing in VR. For a golf put comparison, Harris et al. (2021) reported a significantly better putting performance in real world while the distance estimation is not significantly different for VR and real-world conditions. The SIM-TLX Index used in the study (Harris et al., 2020), which has similar measures with NASA-TLX, did not reveal any difference except for its Perceptual Strain item.

The above studies show that despite some minor differences depending on the task and participants, the VR experiences are comparable to their real-world counterparts in terms of perceived task load. However, the number of studies are very limited and there is a need for further exploration, to generalize the findings of the studies.

2.4.2. Presence in VR and real world.

Usoh et al. (2000) suggest that asking a person concerning "sense of being there" that person will come up with some interpretation that makes the question seem sensible, and then answer that question. They suggested that their findings of indifference between real-world and virtual reality presence are affected by this issue that it does not make sense for their participants to assess their real-world presence through a Likert scale, since the presence should always be the highest in the real world. So, the participants, who are asked to assess their presence in the real-world interpret the question differently from the VR condition, "as the sense of involvement, the lack of isolation, perhaps the degree of comfort"; unlike the virtual world presence that is understood as "as the sense of being in the environment that is depicted by the computer-generated displays, and the ability to act in that environment". Despite the subjective understanding of presence can be different for real-world and virtual environments, recent research provide evidence that brain activities of people are similar in HMD-based VR and real-world, compared to a desktop PC interaction scenario (Petukhov et al., 2020).

Although the visual fidelity of the representation of virtual objects to the real-world aspects affects the sense of presence (Slater, 2009;), it "seems to be far less important for presence than are other parameters, such as head tracking, frame rate, sound and interaction methods" (Sanchez-Vives & Slater, 2005) as depicted by Cummings and Bailenson (2016) through an analysis of empirical studies. That makes presence a phenomenon that can be observed in virtual reality, as it is based on the "real-world" consciousness transported into an alternative, virtual environment. Immersed in a virtual environment, "people respond to events in that place, feel their body in that place, and event transform their body ownership" to the virtual representation of a body that they see, resulting with "conscious and volitional behaviors" such as avoiding the heights and non-conscious psychophysiological responses, such as an elevated heart rate (e.g. Kisker et al., 2021). Presence "is a perceptual but not a cognitive illusion", as the perceived stimuli leads to a rapid automatic reaction of brain-body system while "the cognitive system relatively slowly catches up and concludes 'But I know that this isn't real'" (Slater, 2018).

On the other hand, Weber et al. (2021) suggest that presence should not be limited to spatial presence, i.e. being there, but also should consider the perceived realism. According to authors, spatial presence is affected by allocation of attentional resources including flow related measures along with place illusion. Perceived realism is the subjective judgement degree of the user on the fidelity of the stimulus.

Chapter 3

Methodology

3.1. Participants

14 male and 14 female participants, aged from 18 to 27 (M=23.11, SD=3.1) volunteered to participate in the study, who were recruited among the social circle of one of the researchers. A requirement was to have the ability to communicate verbally in English. Participants have been introduced to the HMD ahead of time, yet none knew about the subject of the experiment “Keep Talking and Nobody Explodes” on any medium the app is released on. Their mean daily average weighted gameplay exposure score was 3.22 out of 7 (SD=1.59) measured using the weekday/weekend weighted frequency scores adapted from LTE (Riddle, 2010), considering the participants’ weekly gaming exposure within the last 3 months. The responses for the weekday items have a weight of 71.4 and weekend responses are weighted as 28.6, considering the 5 weekdays to 2 weekends ratio in a week.

Male participants (M=3.76, SD=1.31) were gaming more frequently, compared to female (M=2.69, SD=1.7) participants. Distribution of the daily average weighted gameplay exposure scores in the group and comparison of male and female participants can be seen on Figure 1. Female and male participants equally assigned to two play sequence groups which engage with the toy version of the game at first and VR version of the game at first. None of the participants had played the game before in these platforms or any other platforms.

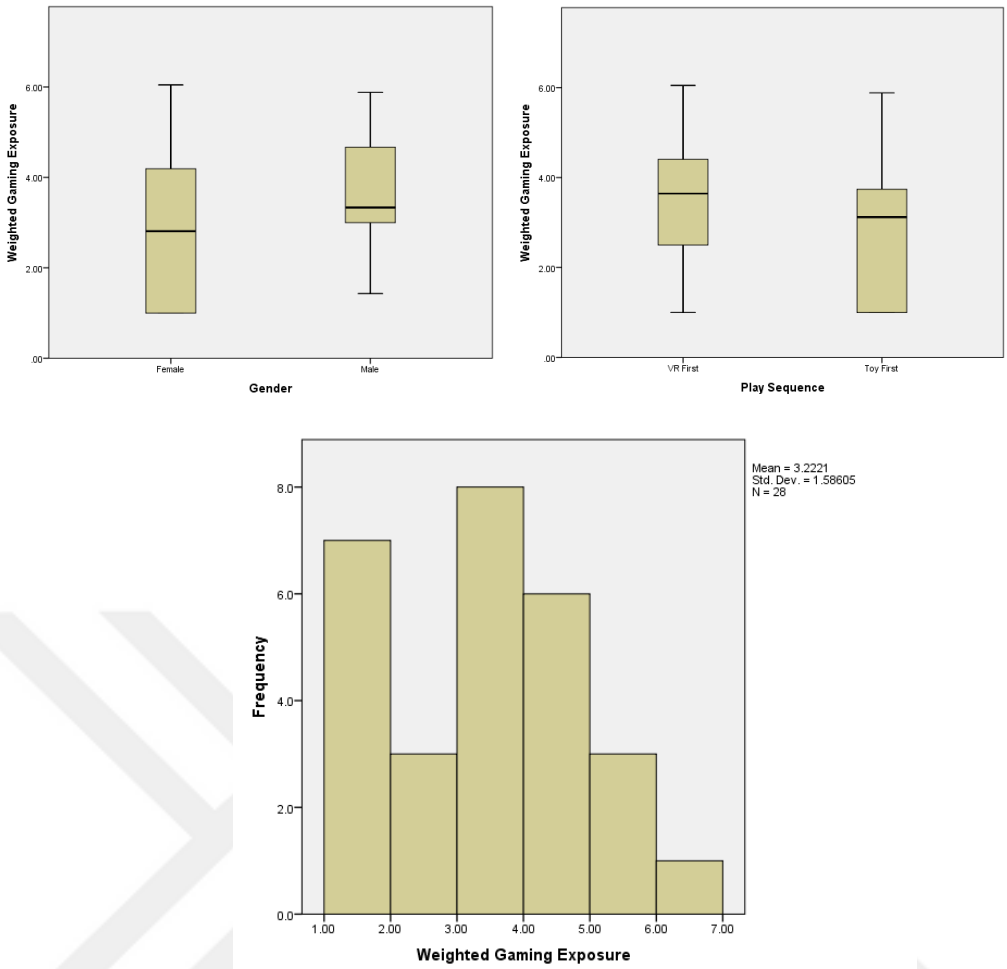


Figure 1. Weighted gaming exposure of participants by gender, play sequence groups, number of participants regarding their scores.

When the participants are compared according to play sequence, the group who played the game in VR at first had a slightly higher daily average weighted gameplay score (M=3.48, SD=1.66) compared to the players who engaged with the toy version first (M=2.97, SD=1.53).

The previous experiences with VR were not very high among the group, where 17 had no prior VR experience and six participants tried once, three participants had tried it a couple times, one has used more than 10 times and one reported himself as an occasional user, as given in Figure 2. Although it is not possible to assign the participants to the play sequence groups to have exactly the same level of prior VR experience for VR First and Toy First group, there is a balanced distribution of experienced and inexperienced participants to these groups. Most of the female

participants had no prior VR experience, while male participants were more experienced.

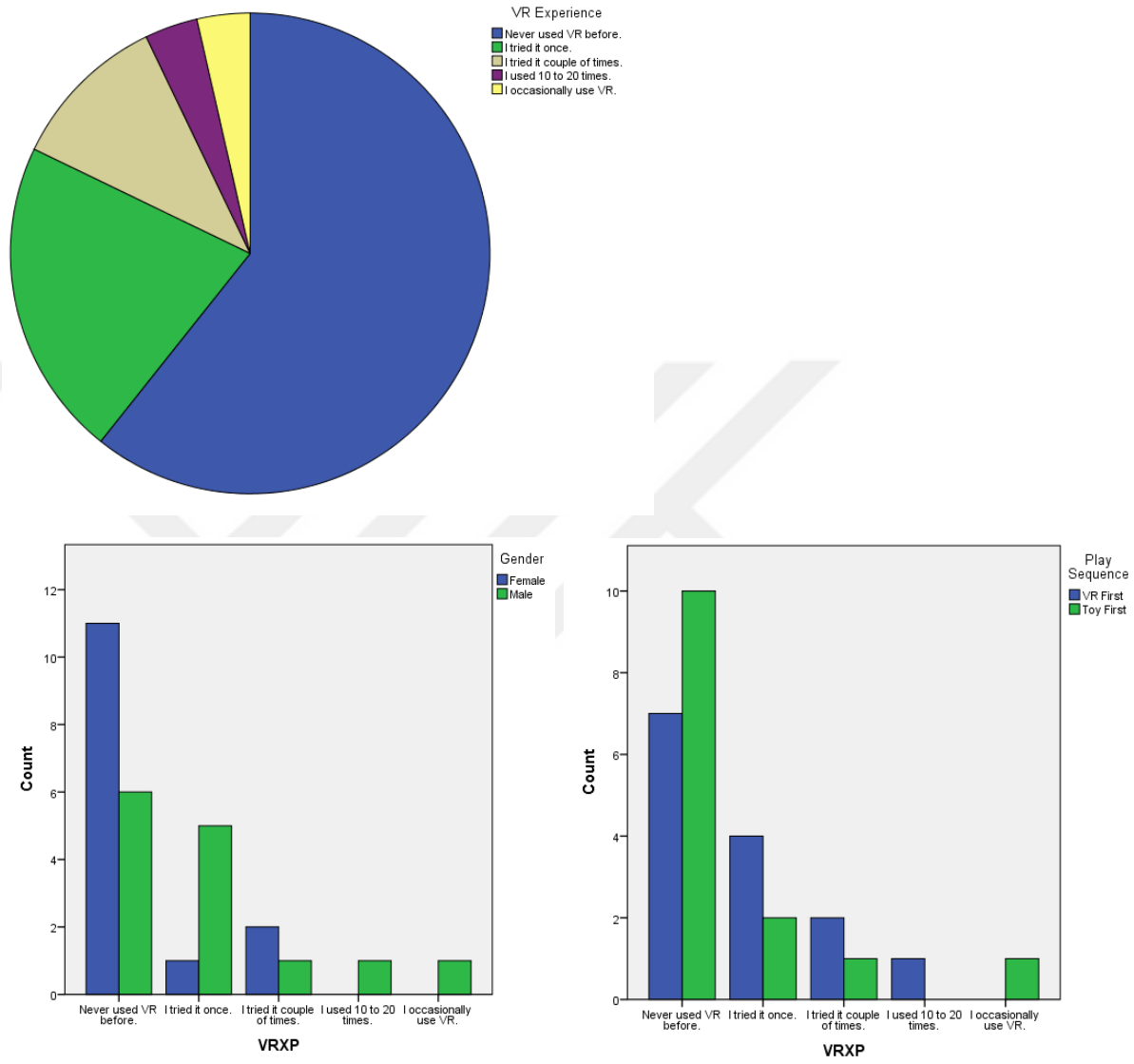
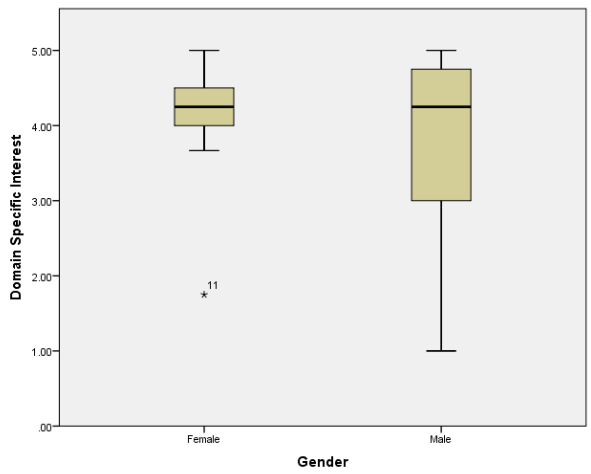
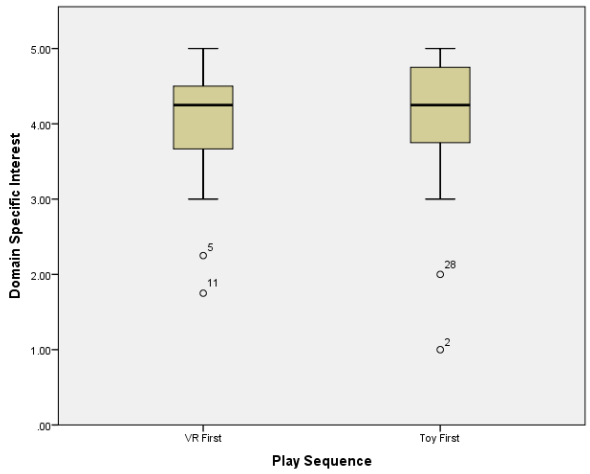


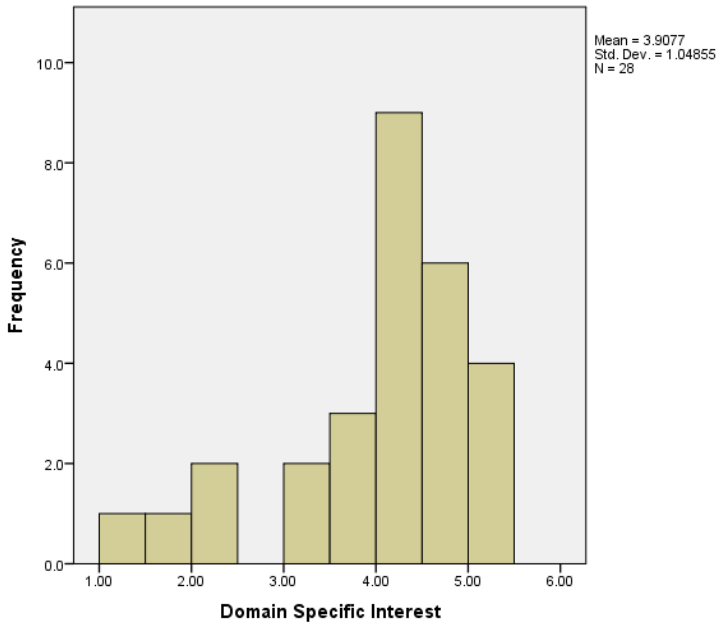
Figure 2. VR Experience by gender, play sequence and number of participants.

When the participants' MEC-SPQ domain specific interest scores were investigated, the mean scores for male and female participants were higher for female (M=4.08, SD=.76) than male (M=3.73, SD=1.28), where the mean score for all participants is 3.9 (SD=1.04). Many of the participants had a moderate interest in the subject of the game, which is suggested as an enduring personality factor that affects the attention allocation of participants. The mean score for the participants who played

the VR version at first (M=3.90, SD=.97) were similar to the participants who played the toy version at first (M=3.91, SD=1.16).







We also explored the Visual-Spatial Imagery score via MEC-SPQ, which is also suggested as a trait-like enduring personality factor that affects the experience of a player in terms of presence.

The scores for the VR-at-first (M=3.91, SD=.69) and toy-at-first (M=3.79, SD=.77) groups are also very close for Visual Spatial Imagery.

A series of independent samples t-test revealed that there is not significant difference between VR-at-first and toy-at-first in terms of prior VR experience $t(26)= .31$, $p > .05$., daily average weighted gameplay score $t(26)= .84$, $p > .05$, participants'

interest in the subject domain of the stimuli $t(26)=.015$, $p>.05$ and abilities of visual spatial imagery $t(26)=.45$, $p>.05$. Thus, it can be concluded that our participants are evenly represented in VR-at-first and toy-at-first groups based on their prior VR experience, gaming habits, domain interest and spatial abilities. On the other hand, the further comparisons would be made in a within-subjects manner, so that play sequence is alternated to prevent any possible bias that may emerge from the participants' familiarity with the goals and mechanics of the game.

3.2. Equipment and Stimulus

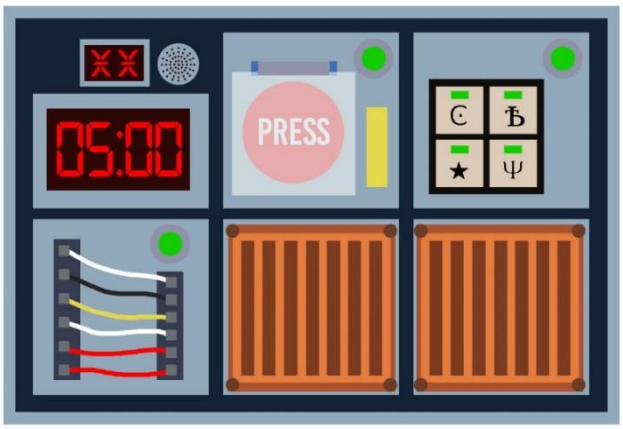
Participants used the same replica of the 'virtual bomb' for the real-life condition. An Oculus Rift HMD with a pair of Oculus Touch magic wand controllers were used for stereoscopic head-tracked virtual reality condition. In real life condition, all the interactions were directly applied on the 'physical bomb', while magic wands are the only method of input for VR condition.

All participants played the 'first bomb' which is the starter level of the game in which players try to diffuse a virtual bomb by following the instructions of another person – the moderator, who is not allowed to see the virtual bomb but reads the bomb-manual which is a printable booklet (bombmanual.com, n.d.) with the instructions to read and help the player navigate the process of diffusing each module of the puzzle. The player must describe what is seen on each module of the bomb. Although the medium of experimentation will change, the same person will be guiding the player, as a part of the replicated experience. In the VR version, the game will generate a different combination of variables every time, but the same level and modules will be at play. While in the real-world condition the same physical bomb is constructed, the player will be exposed to it once, which will make no difference in the experiment since in both mediums the participants will experience a different generated puzzle.

Looking at Picture 1, the first level contains three modules: a countdown of 5 minutes as a standard for this level, a module of wires with different colors, another module of four symbols on a keypad, and a button to deactivate the bomb. On the side of the bomb, there are elements that add to the visual only, and there is serial number, and battery count which decide certain variables in the “bomb-manual” read by the moderator. The bomb gets generated randomly, each time different variables in each module, the wires count and colors, for symbols its combination of sets, and the button has its own color, writing, and led color which determine the right diffusing action. There isn’t a specific order to diffuse these modules.

Picture 1. The "virtual bomb" in "Keep Talking and Nobody Explodes" game.

As seen in Picture 2, the game in VR offers an object of interaction; with specific user interface and user experience we replicated the object with its characteristics of interaction, interface, and visuals. This design took the form of laser cutting of plexiglass material to emulate the structure of the “virtual bomb”.



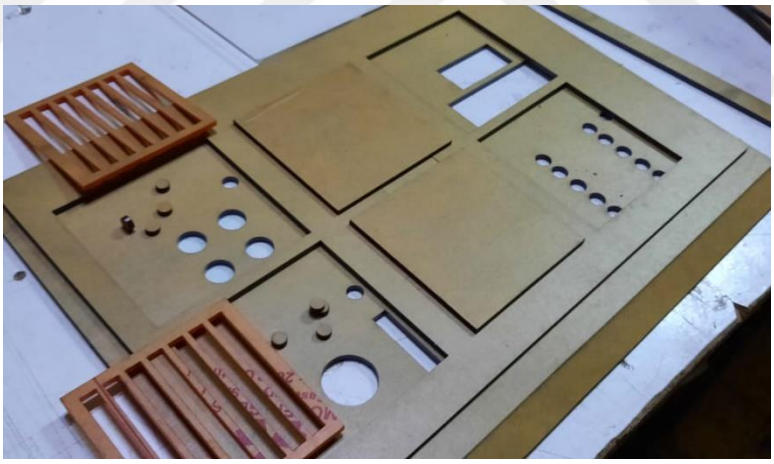
Picture 2. The concept design of the "Physical Bomb".

The result seen in Picture 3 to reach a fully functional device. Inside the box is an operational system of a programmable logic controller (PLC) or which is an industrial digital computer that has been ruggedized, and adapted for the control of manufacturing processes, such as assembly lines, robotic devices, or any activity that requires high reliability, ease of programming, and process fault diagnosis. This controller is connected to a Human-Machine Interface unit (HMI unit), which communicates with the PLC using serial communications or by way of an Ethernet connection. HMIs are programmed using free software and via a serial port (USB, or standard DB9 port). Other visual aspects of the “Virtual Bomb” are applied from the elements on the sides to the indicators of green led lights when a module is diffused. The process was meant to replicate a more identical experience by taking the more controllable condition of VR into the real-world condition, which is the opposite of what is usually done. As given in Picture 3, the physical toy has the same components with the virtual bomb, while their locations may not be identical, since the VR game is programmed to set their positions randomly in each gameplay.

Game wise, for the interaction part, each of the modules was based on the bomb-manual booklet which the moderator reads out loud for the participant. The color, and count of wires, the sequence of symbols, and the button are all designed and programmed based on the script provided by the game.

Picture 3. The physical toy bomb (on the left) replicated from the virtual bomb (on the right).

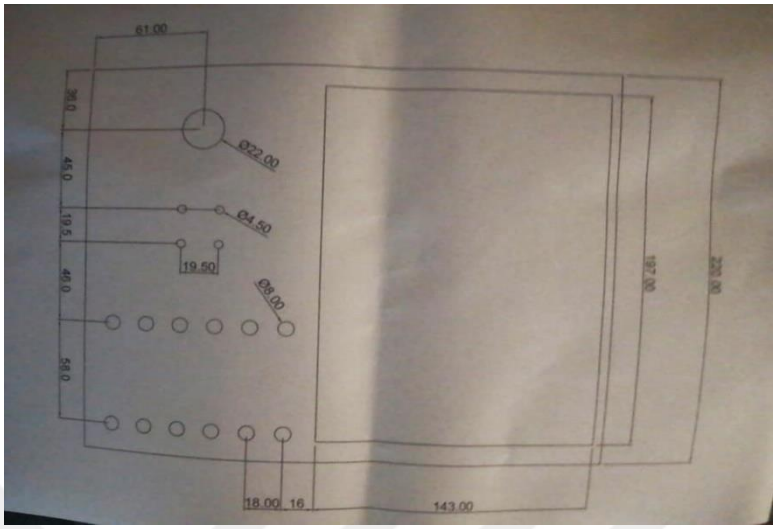
As seen in Picture 4, the layering present is just the last two plexy-glass cover which was used to divide and convey the same UI of the “virtual bomb” which was the 4th trial in remaking the box to fit the machines inside, and at the same time allow the distribution of elements on top to mirror that of the VR generated “virtual bomb”.



Picture 4. Part of the construction process of the "Physical Bomb".

In Picture 5 is another example of the earliest prototype which failed to complement the desired distribution of elements on the UI of the box. Noticing the details and accuracy of distribution in a mechanical point of view, this phase aimed to test the electronics and the functionality of the programmed software, since the first and most important challenge was to create a functional sequence of actions that exactly operate as the game manual dictates. So, when the player interacts with the

“physical bomb” it will be a copy of the experience to be within the immersive HMD VR medium.



Picture 5. The first design map of the "Physical Bomb".

3.3. Measures

The task performance is measured through the objective indicators of game success, i.e. successfully disarming the bomb in 5 minutes and time left in successful disarming; in the VR version it is fixed in the “Virtual Bomb”. As for the real-life condition, the timer was programmed to five minutes in the “Physical Bomb”, and the moderator served as the common factor of experience in both mediums.

For subjective assessment, the self- report measure of NASA-TLX (National Aeronautics and Space Administration – Task Load Index (Hart & Staveland, 1988; Hart, 2006) is used to measure the workload of the task, perceived subjectively, by the player diffusing the bomb. The six subscales, with items evaluated on a 20-point horizontal line scale and originally scored through a two-step process, as a weighted sum of paired comparisons of the six dimensions. Evaluation of participants’ performance on the task as they may have experienced, or perceived it to be, Physical Demand, Mental Demand, and Temporal Demand, adding Performance, Frustration, and Effort spent.

Each dimension is rated on a scale of 0 to 5. Based on these ratings, a weighted score is calculated for each dimension. This allows for the higher sensitivity of the NASA-TLX in determining the variables of the experiment. The NASA-TLX tool

provides a way to test different difficulty and complexity of tasks. In the game "Keep Talking and Nobody Explodes" NASA-TLX works well, since the temporal demand is the main aspect of success, and the mental demand is present by the puzzle solving, and frustration can be observed as the player and moderator try to diffuse each module, while the physical demand is engaged with the process of interacting with each medium of the bombs. As for effort and performance, those are subjectively reported experiences that are assessed and varied based on individual differences.

While SIM-TLX is a recent alternative (Harris et al., 2020) of NASA-TLX as a self-report measure for performance specific to VR, we preferred NASA-TLX for two reasons. First, it can be benchmarked (Hertzum, 2021) with a worldwide collection 556 studies. Second, the results can be compared with a previous study which assesses the same bomb disarming puzzle task between HMD VR and PC monitor conditions (Berkman et al., 2020). Third, since the mentioned study reports differences between adjusted and raw scores of NASA-TLX due to perceived importance of the task loads, we believe that this issue should be further explored through real-world to VR comparison.

Considering the research and conclusions about presence in the real-world vs. the virtual environments, which are given in the Related Studies, it may not seem to be conceivable to assess a real-world toy experience regarding the presence and comparing it to a virtual environment. On the other hand, the bomb used in our study is a "toy", i.e. a non-real artefact by its definition. The pretence in engaging with a toy is similar to a VR experience, while the subject is aware that the toy is not real, as he does in VR, at the cognitive level. Likewise, the pretended engagement with the toy at the conscious behavioural level is similar to behaviour in VR, where engagement with toys may also lead to unconscious psychophysiological responses (e.g. Hughes & Hutt, 1979). Since presence is a mediated experience, we claim that the "toy bomb" can also be considered as a medium that enables presence. For this reason, we let our participants to choose whether the presence related MEC-SPQ items were applicable to the toy condition or not. Although there are many questionnaires to assess presence (Grassini & Laumann, 2020), we decided to use SPES (Hartman et al., 2015) as it assesses spatial presence within Self-Location and Possible Actions dimension, while it evolved from MEC-SPQ, which also considers the realism and flow related measures as a component of presence. Furthermore, it is known that both are sensitive to different media and content (Yildirim et al., 2019; Berkman et al., 2020).

3.4. Data Analysis

The performance data and the survey data were analyzed through a series of paired samples t-tests, assessing the scores obtained with the VR compared to the toy condition for each participant where the sample size is enough and the mean differences between the pairs are assumed to be normally distributed, regarding the results of Shapiro-Wilk tests. Some of the participants did not respond to the SPES items for the toy condition, as they thought that they are not applicable for the real-world. For Spatial Presence dimension, 11 participants respondent for both toy and VR conditions, where there were 10 responses for Possible Actions dimension. Due to the small sample size, we executed non-parametric Wilcoxon signed-rank tests for these scores. In addition, the null hypothesis of normal distribution cannot be rejected for mean differences of MEC-SPQ Attention Allocation, Involvement, and raw NASA-TLX performance scores, which are also explored via non-parametric Wilcoxon signed-rank tests.

3.5. Procedure

Participants are asked to play either the VR or the toy version of the game, “Keep talking and Nobody Explodes”. The VR version is a commercially available product, not only for VR, but also for many other gaming platforms such as personal computers and mobile devices. The toy version is built based on the first level of the game, using an operational system of a programmable logic controller (PLC) or which is an industrial digital computer that has been ruggedized, and adapted for the control of manufacturing processes, such as assembly lines, robotic devices, or any activity that requires high reliability, ease of programming, and process fault diagnosis. This controller is connected to A Human-Machine Interface (HMI) unit, which communicate with the PLC using serial communications or by way of an Ethernet connection. HMIs are programmed using free software and via a serial port (USB, or standard DB9 port). Half of the participants played the game using Real life condition first. The other half was exposed to the VR version as the first experience with “Keep

Talking and Nobody Explodes". Each participant was left alone in a closed room, where the moderator would communicate via headset to give the guidance. After every session, participants are asked to fill out the NASA-TLX questionnaire that includes personal questions, along with the items that query on their gameplay session, in which they had to try and indicate the time remaining in order to state their duration of their gameplay within the questionnaire, which was recorded by experimenters and given to them after they completed the experiment. Participants are allowed to rest for 10 minutes after they finish the questionnaire and asked to play the game on the other platform again, which is also followed by a questionnaire. At this time, the questionnaire only included the gameplay related queries.

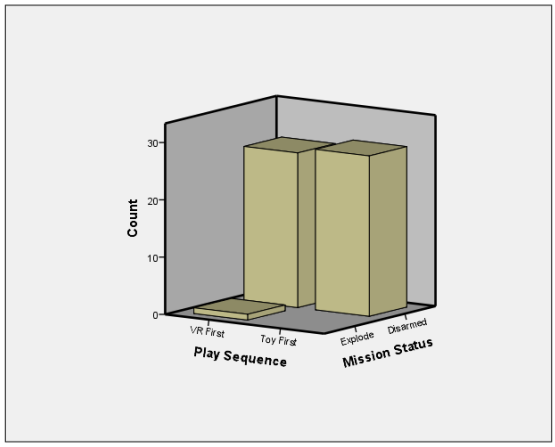
Picture 6. Female participant defusing the toy bomb (on the left) and male participant defusing the virtual bomb (on the right).

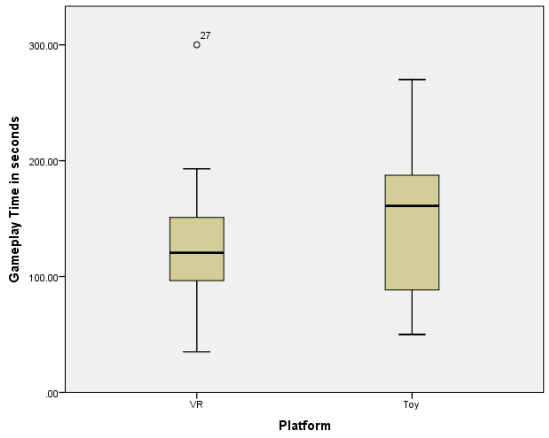
Chapter 4

Findings

4.1. Performance and Task Load

As an objective indicator of performance, the task completion success and task time was explored. All 28 participants were able to disarm the bomb an all their trials, except one participant who exploded on his first trial, which was in VR environment. The paired samples t-test did not reveal a significant difference on the time spent on task t(27)=-1.5, p>.05, where VR sessions took around 125.3 (SD=52.24) seconds and toy sessions took 147.4 (57.49) seconds.





The task load is also explored through the subjective NASA-TLX assessments, using raw and adjusted scores.

When the raw results on VR condition is benchmarked with the reference values obtained from 72 VR studies (Hertzum, 2021), we see that all raw scores are above the mean reference values. Physical Demand, Temporal Demand and Frustration is also very high above the upper bounds of benchmarked studies, showing that the VR task has a high workload. On the other hand, it should be considered that the benchmarked VR applications are for education and healthcare purposes. It is not unexpected that a puzzle game, which is deliberately designed to be challenging, would lead to higher workload scores. Hertzum (2021) did not report a category of applications that can be compared with our toy condition results. Both toy and VR scores are compared with the reference values gathered from 93 studies in Asia and the scores of our participants are higher than benchmark means except the Effort score for toy condition. It should also be noticed that the Mental Demand score for toy condition is also very similar to scores obtained in Asia. Considering the 556 studies, all conditions for all dimensions are still above the average NASA-TLX scores except for the Effort score for toy condition, which corresponds to the 50th percentile of benchmark scores. These comparisons suggest that the system evaluated in this study lead to a task load that are above the average, considering the percentile correspondence of the scores.

When adjusted TLX scores are assessed along with the percentile ranks, the mental demand is higher for the VR condition than the toy condition, as well as Temporal Demand, Effort and Frustration scores. However, paired sample t-tests t(27)= see Table 1, p>.05) did not reveal a significant difference on any of the NASA-TLX scores, neither for the raw nor for the adjusted scores. The raw Performance score is assessed through a Wilcoxon signed rank test, which did not reveal a significant difference between the VR (Mdn=62.5) and toy (Mdn=80) conditions; z=-2, p=.045.

Table 1

NASA TLX Scores, Benchmark Values and Mean Comparisons

		Mean		SD		t (27)		Benchmark (Raw)			
NASA TLX -	Cond.	Raw	Adj.	Raw	Adj.	Raw	Adj.	VR	Asia	General	Percentile
Mental Demand	VR	63.57	198.21	29.02	130.91	1.57	1.72	47 ± 20	55 ± 19	49 ± 17	70th
	Toy	56.07	156.43	30.98	128.46						60th

Physical Demand	VR	51.25	85.54	26.09	95.87	0.36	1.11	26 ± 13 42 ± 18 32 ± 16	80th
	Toy	48.93	117.14	31.57	127.54				80th
Temporal Demand	VR	66.25	217.32	25.73	145.38	1.73	1.28	40 ± 17 47 ± 19 42 ± 16	90th
	Toy	56.43	178.57	28.89	145.02				70th
Performance	VR	56.61 62.5*	153.21	28.77	115.17	.94 -2*	.95	47 ± 24 49 ± 21 45 ± 19	70th
	Toy	63.04 80*	181.25	33.26	133.94				80th
Effort	VR	64.64	138.75	25.05	110.53	1.97	0.50	48 ± 19 55 ± 19 50 ± 16	80th
	Toy	53.04	127.86	31.01	111.25				50th
Frustration	VR	59.11	157.86	22.77	131.13	1.52	1.37	34 ± 18 40 ± 18 36 ± 14	90th
	Toy	50.36	117.14	28.83	118.39				80th
Overall TLX	VR	60.24	63.39	15.98	16.53	1.54	1.26	41 ± 15 48 ± 16 42 ± 13	90th
	Toy	54.64	58.56	19.65	19.95				80th

* Median values since the mean differences cannot be assumed as normally distributed
^ z score obtained for Wilcoxon signed rant test

Our results suggest that the workload is not different when the bomb puzzle task is executed either in a real world setting or a VR setting. The result is not expected, as many of the previous studies report similar findings that does reveal a difference between real-world and VR, or some resulted on behalf of VR condition leading to a lower task load. Especially, the studies conducted in 2015 reported workload for VR that comparable or lower than real-world settings, which might be due to the advancements in displays quality, motion tracking and controllers.

4.2. Game User Experience

From the game user experience (UX) perspective, the Enjoyment dimension of GUESS revealed a significant difference, where the mean score was 5.55 (SD=.65) for

VR condition compared to the 5.19 on toy condition, t(25)=3.11, p<.05). The playability/usability score obtained via GUESS items for both conditions are very similar. The Personal Gratification score obtained via GUESS is not also significantly different, as given in Table 2.

The Attention Allocation score of MEC-SPQ, which were explored through a Wilcoxon signed rank due to the distribution of mean differences, did not also reveal a significant difference between the VR (Mdn=5) and Toy (Mdn=4.75) conditions, z=1.73, p=.083; along with the MEC-SPQ Involvement scores, z=-1.5, p=.148.

Table 2

Mean Comparison of GUESS Scores

UX metrics	Condition	Mean Score	SD	T-Test
Playability / Usability (GUESS)	VR	5.61	0.86	t(26)=-.07, p=0.95
	Toy	5.62	0.87	
Enjoyment (GUESS)	VR	5.55	0.65	t(25)=3.11, p=.005
	Toy	5.19	0.67	
Personal Gratification (GUESS)	VR	6.23	0.67	t(25)=2.05, p=.05
	Toy	6.02	0.60	

The similar score on Playability/Usability dimension supports our suggestion that the VR equipment is not an obstacle anymore that creates extra workload for the user. As we aimed, we arranged the real-world environment very similar to the VR condition. Attention/Allocation and Involvement scores suggest that both conditions are alike. Since the task on both conditions are equivalent, the focus of participants' attention was on the puzzle, regardless of the bomb being a digital artefact in VR or an electro-mechanical toy in the real world. Both objects had evoked a similar level of reflections upon them, i.e. the involvement of participants. Players had a similar sense of achievement regardless of the gameplay medium, according to the Personal Gratification scores. The mean difference between the enjoyment scores is significantly different but not very large. Since most of our participants have very limited or no VR experience, the higher VR score might be due to a novelty effect. It

could be concluded that using VR did not create a richer game user experience within the context of a bomb puzzle game.

4.3. Presence

Presence related metrics of MEC-SPQ were also led to very similar scores for both conditions, as given in Table 3. The Spatial Situation Model score, where the participants are queried for their mental image of the medium regarding to its arrangements, size and surroundings were almost equal for VR and toy experiences. The four Suspension of Disbelief items where the participants reflect on the inconsistences, errors, contradictions, and their attention on these issues also led to a very similar score for both conditions.

Table 3

Comparison of Mean Scores for Presence-related Measures

UX metrics	Condition	Mean Score	SD	T-Test
Spatial Situation Model (MEC-SPQ)	VR	4.04	0.90	t(27)=-0.16, p=.87
	Toy	4.06	0.84	
Suspension of Disbelief (MEC-SPQ)	VR	2.93	0.87	t(27)=0.14, p=.89
	Toy	2.91	0.76	

Regarding the Spatial Situation Model score, we think that participants were able to perceive and understand the object of their interest, the bomb, clearly in both conditions. Considering the Suspension of Disbelief score, the design of the bomb, both the "real-world" object and the VR representation, were found equally plausible. We think that the puzzle-like design of the bomb might have looked irrelevant to the participants, leading to medium level score, i.e. around 3 over 5 for both conditions. However, it did not have any flaws or glitches for its functionality and its look.

The SPES items were not found to be applicable to the toy condition by many participants. On the other hand, more than one-thirds of the participants though that they can respond to these questions within the context of their experience with the toy.

The Wilcoxon Signed Rank Test analysis run on this small sample size revealed that there is not a significant difference between the toy (Mdn=4.38) and the VR (Mdn=4.5) condition, z=-2,49, p=.013 for the Self-Location dimension of SPES, which is querying the participants on "actually and physically being there and taking part in the presentation".
For the Possible Actions score which is about acting on the objects, there was not a significant difference between the toy condition (Mdn=4.5) and the VR condition (Mdn=4.19), z=-.71, p=.48.

These results are consistent with (Usoh et al., 2000), where the real-world office space is compared with its VR representation through (Witmer & Singer, 1998) and (Slater at al.,1994) presence questionnaires. While the authors hypothesized that presence questionnaires would result with higher scores for the real-world experience, the difference was marginally significant or similar.

It should be remembered that the environment presented in the VR condition is a dark room where no details of the surrounding is available to the user. Likewise, the toy condition is experienced in a dark room where the only lit area is the table that the bomb resides, which is quite similar to VR. Although the median score for "Self-location" is very high at 4.5, suggesting that users are "being there", we think that "there" is not well defined in our VE, but mostly left to the user's imagination. While not significantly different, the Possible Actions score is slightly higher for toy condition. Actually, there are more possible actions in the toy condition. Users can break apart the toy bomb, throw it outside the room or do whatever they want that we cannot imagine right now, within the limitations of the physical world. Although their actions in VR are limited to rotating the bomb and engaging with the controls, it is interesting that they scored the Possible Actions items in a similar manner. We think that our users restricted themselves with the rules of "the game", although they are not forced by any design limitations.

Chapter 5

Discussion and Conclusions

The results presented and discussed above contributed to our understanding of simulated experiences across an immersive medium, and the physical replication of it. They reveal that a gaming task that highly depends on cognitive processes such as solving a puzzle do not lead to significant differences in user performance, workload, user experience or presence either it is executed through a physical toy, or an HMD based VR environment. That is in line with the findings of studies conducted after 2015, which report that workload for VR is comparable or lower than real-world settings; due to the advancements in displays quality, motion tracking and controllers after 2015. The scores of participants revealed that a well simulated and controlled environment can eliminate the drastic variables of user experience in the given mediums; however the limitation of simulated elements in the dark room, and the focus on the object as the main subject of interaction allows the imagination of the user to simulate a similar level of capability with the VE as in the physical version, while we observed only a minor difference just for the Enjoyment rating by the participants, that can be explained through a novelty effect since our participants had some very limited prior experience with HMDs.

Unlike in other experiences where the environment itself would require interaction, or the object of interaction would be more demanding to simulate, playing a puzzle game through the different mediums did not indicate a significant difference specifically in terms of effort or performance. The subjective workload is higher than NASA-TLX benchmark results of other VR studies, studies executed in Asia and worldwide. Since the evaluated task is a gameplay that is deliberately designed to be challenging, these results are not unexpected.

While the results obtained regarding the user experience and presence are very similar in both real-world and virtual environment tasks as well as performance metrics, our findings suggest that the VR gaming experiences can replace with the real-world game-play. However, it should be noted again that this study is conducted through a puzzle game, in which the task has a mental and temporal demand rather than being physical, exposing the user to the pressure of making right decision choices

to pass a level before the countdown ends. In our experiment, the tasks were applied by a handful of mechanics that manipulate the basic object of a box, ranging from wire cutting, pressing the buttons, and rotating the bomb, which is a different system all together than the type of experiences such as racing, or jet flying games where the user becomes surrounded by the interactive environment that is constantly changing and also interacting with the user's choices of speed, steering, and mobility. The application evaluated in our study has a minimal spatial surrounding which does not affect the gameplay. On the other hand, our study provided evidence that the interactions with the nearby objects are very similar in an HMD-based virtual environment and real-world, within the context of a cognitive task.

In "Keep Talking and Nobody Explodes", the in-game system is also different from other games. Having such mechanics can define the metrics of demand, mental, temporal, and physical. While the balance of such interactions causes a similar amount of effort, and reported level of frustration, and quality of performance, a defining characteristic of the experiment was the element of narrative in both mediums (the moderator); while both the player and the moderator join efforts to challenge the temporal limit, each of them would express different renders of what the other cannot see in the puzzle's main parts. This creates a tension that forms a sense of solidarity and enjoyment when the puzzle is solved (bomb is diffused). This dynamic of interaction places the simulated elements of a racing game, or mountain climbing out of the essential factor of the experiment replication.

The primary goal of this study was not exploring the concept of presence beyond the virtual environments. On the other hand, our results imply that real-world experiences such as engaging with a toy can also invoke sense of presence as a consequence of mediated interaction. The findings of this experiment allowed us to conclude through replicating a virtual experience to the real physical world, with such simulated familiarity as a tabletop puzzle, the results can show that a task can be measured by TLX in VR. As the NASA-TLX was conducted in physical world simulations of an experience, and this is what we had in this experiment, the results' similarity of performance and effort indicate that a task, training, or skill can be measured and taught in a carefully simulated VR task, and environment. However, if the simulated task would have been more demanding in the elements to replicate, the

results would have shown that VR requires more variables to offer a similar level of involvement in the assigned task.

In order to provide an opportunity for a more detailed comparison of findings with future studies, we decided to publish our dataset, as we are expecting the other researchers to share their data publicly.

REFERENCES

Abouelkhier, N., Shawky, D., & Marzouk, M. (2021). Evaluating distance perception for architecture design alternatives in immersive virtual environment: a comparative study. Construction Innovation.

Agethen, P., Sekar, V. S., Gaisbauer, F., Pfeiffer, T., Otto, M., & Rukzio, E. (2018). Behavior analysis of human locomotion in the real world and virtual reality for the manufacturing industry. ACM Transactions on Applied Perception (TAP), 15(3), 1-19.

Annett, J. (2002). Subjective rating scales: science or art?. *Ergonomics*, *45*(14), 966-987. DOI: 10.1080/00140130210166951

Berg, L. P., & Vance, J. M. (2017). Industry use of virtual reality in product design and manufacturing: a survey. *Virtual reality*, *21*(1), 1-17

Berkman, M. I., Çatak, G., & Eremektar, M. C. (2020) "Comparison of VR and Desktop Game User Experience in a Puzzle Game:“Keep Talking and Nobody Explodes”." *AJIT-e: Bilişim Teknolojileri Online Dergisi* 11(42), 180-204.

Bhargava, A., Lucaites, K. M., Hartman, L. S., Solini, H., Bertrand, J. W., Robb, A. C., ... & Babu, S. V. (2020). Revisiting affordance perception in contemporary virtual reality. Virtual Reality, 24(4), 713-724.

Bishop, I. D., & Rohrmann, B. (2003). Subjective responses to simulated and real environments: a comparison. *Landscape and urban planning*, *65*(4), 261-277. DOI: 10.1016/S0169-2046(03)00070-7

Bombmanual.com (n.d.). Bomb Defusal Manual for the game Keep Talking and Nobody Explodes. Archived at https://web.archive.org/web/20220321124706/https://www.bombmanual.com/web/index.html

Bowman, D. A. (1998). Interaction techniques for immersive virtual environments: Design, evaluation, and application. *Journal of Visual Languages and Computing*, *10*, 37-53.

Bowman, D. A., & McMahan, R. P. (2007). Virtual reality: how much immersion is enough?. *Computer*, *40*(7), 36-43. DOI: 10.1109/MC.2007.257

Chamilothori, K., Wienold, J., & Andersen, M. (2019). Adequacy of immersive virtual reality for the perception of daylit spaces: comparison of real and virtual environments. *Leukos*, *15*(2-3), 203-226.

Chang, T. P., Beshay, Y., Hollinger, T., & Sherman, J. M. (2019). Comparisons of Stress Physiology of Providers in Real-Life Resuscitations and Virtual Reality–Simulated Resuscitations. Simulation in Healthcare, 14(2), 104-112.

Chen, Y., Cui, Z., & Hao, L. (2019). Virtual reality in lighting research: Comparing physical and virtual lighting environments. Lighting Research & Technology, 51(6), 820-837.

Cliburn, D., Winlock, T., Rilea, S., & Van Donsel, M. (2007, November). Dynamic landmark placement as a navigation aid in virtual worlds. In *Proceedings of the 2007 ACM symposium on Virtual reality software and technology* (pp. 211-214).

Cooper, N., Millela, F., Cant, I., White, M. D., & Meyer, G. (2021). Transfer of training—Virtual reality training with augmented multisensory cues improves user experience during training and task performance in the real world. PloS one, 16(3), e0248225.

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media psychology*, *19*(2), 272-309.

DiDomenico, A., & Nussbaum, M. A. (2008). Interactive effects of physical and mental workload on subjective workload assessment. *International journal of industrial ergonomics*, *38*(11-12), 977-983. DOI: 10.1016/j.ergon.2008.01.012

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (2020). Workload assessment in multi-task environments. In *Multiple-task performance* (pp. 207-216). CRC Press.

Everson, T., McDermott, C., Kain, A., Fernandez, C., & Horan, B. (2018). Astronaut training using virtual reality in a neutrally buoyant environment. 319–327.

Ewin, C. A., Reupert, A., McLean, L. A., & Ewin, C. J. (2021). Mobile devices compared to non-digital toy play: The impact of activity type on the quality and quantity of parent language. *Computers in Human Behavior*, *118*, 106669.

George, C., Khamis, M., Buschek, D., & Hussmann, H. (2019, March). Investigating the third dimension for authentication in immersive virtual reality and in the real world. In 2019 ieee conference on virtual reality and 3d user interfaces (vr) (pp. 277-285). IEEE.

Grassini, S., & Laumann, K. (2020). Questionnaire measures and physiological correlates of presence: A systematic review. *Frontiers in psychology*, *11*, 349.

Greenwald, S. W., Kulik, A., Kunert, A., Beck, S., Fröhlich, B., Cobb, S., & Maes, P. (2017). Technology and applications for collaborative learning in virtual reality. Philadelphia, PA: International Society of the Learning Sciences. 719–726.

Gunn, T., Jones, L., Bridge, P., Rowntree, P., & Nissen, L. (2018). The use of virtual reality simulation to improve technical skill in the undergraduate medical imaging student. *Interactive Learning Environments*, *26*(5), 613-620. DOI: 10.1080/10494820.2017.1374981

Haq, S., Hill, G., & Pramanik, A. (2005, June). Comparison of configurational, wayfinding and cognitive correlates in real and virtual settings. In *Proceedings of the 5th International Space Syntax Symposium* (Vol. 2, pp. 387-405).

Harris, D. J., Buckingham, G., Wilson, M. R., Brookes, J., Mushtaq, F., Mon-Williams, M., & Vine, S. J. (2021). Exploring sensorimotor performance and user experience within a virtual reality golf putting simulator. Virtual Reality, 25(3), 647-654.

Harris, D., Wilson, M. & Vine, S. Development and validation of a simulation workload measure: the simulation task load index (SIM-TLX). *Virtual Reality* **24,** 557–566 (2020). DOI: 10.1007/s10055-019-00422-9

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage publications. DOI: 10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Hartmann, T., Wirth, W., Schramm, H., Klimmt, C., Vorderer, P., Gysbers, A., Böcking, S., Ravaja, N., Laarni, J., Saari, T., Gouveia, F., Sacau, A.M. (2015). The spatial presence experience scale (SPES): A short self-report measure for diverse media settings. Journal of Media Psychology, 28(1), 1–15. DOI: 10.1027/1864-1105/a000137

Hassenzahl, M. (2010). Experience design: Technology for all the right reasons. *Synthesis lectures on human-centered informatics*, *3*(1), 1-95. DOI: 10.2200/S00261ED1V01Y201003HCI008

Hejtmanek, L., Starrett, M., Ferrer, E., & Ekstrom, A. D. (2020). How Much of What We Learn in Virtual Reality Transfers to Real-World Navigation? Multisensory Research, 33(4-5), 479–503. doi:10.1163/22134808-20201445

Hertzum, M. (2021). Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. *Ergonomics*, *64*(7), 869-878.

Hinske, S., Lampe, M., Price, S., Yuill, N., & Langheinrich, M. (2010, March). Let the play set come alive: supporting playful learning through the digital augmentation of a traditional toy environment. In *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)* (pp. 280-285). IEEE.

Ho, A., Lee, J., Wood, E., Kassies, S., & Heinbuck, C. (2018). Tap, swipe, and build: Parental spatial input during i P ad® and toy play. *Infant and Child Development*, *27*(1), e2061.

Huber, T., Paschold, M., Hansen, C., Wunderling, T., Lang, H., & Kneist, W. (2017). New dimensions in surgical training: immersive virtual reality laparoscopic

simulation exhilarates surgical staff. *Surgical endoscopy*, *31*(11), 4472-4477. DOI: 10.1007/s00464-017-5500-6

Hughes, M., & Hutt, C. (1979). Heart-rate correlates of childhood activities: play, exploration, problem-solving and day-dreaming. *Biological Psychology*, *8*(4), 253-263.

Jin, X., Meneely, J., & Park, N. K. (2021). Virtual Reality Versus Real-World Space: Comparing Perceptions of Brightness, Glare, Spaciousness, and Visual Acuity. Journal of Interior Design.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063, pp. 218-226). Englewood Cliffs, NJ: Prentice-Hall.

Kamaraj, D. C., Dicianno, B. E., Mahajan, H. P., Buhari, A. M., & Cooper, R. A. (2016). Stability and Workload of the Virtual Reality–Based Simulator-2. Archives of Physical Medicine and Rehabilitation, 97(7), 1085–1092.e1. doi:10.1016/j.apmr.2016.01.032

Kenyon, R. V., & Afenya, M. B. (1995). Training in virtual and real environments. Annals of Biomedical Engineering, 23(4), 445-455.

Kim, A., Schweighofer, N., & Finley, J. M. (2019). Locomotor skill acquisition in virtual reality shows sustained transfer to the real world. Journal of neuroengineering and rehabilitation, 16(1), 1-10.

Kimura, K., Reichert, J. F., Olson, A., Pouya, O. R., Wang, X., Moussavi, Z., & Kelly, D. M. (2017). Orientation in virtual reality does not fully measure up to the real-world. Scientific reports, 7(1), 1-8.

Kisker, J., Gruber, T., & Schöne, B. (2021). Behavioral realism and lifelike psychophysiological responses in virtual reality by the example of a height exposure. *Psychological research*, *85*(1), 68-81.

Kort, Y. A. D., Ijsselsteijn, W. A., Kooijman, J., & Schuurmans, Y. (2003). Virtual laboratories: Comparability of real and virtual environments for environmental psychology. *Presence: Teleoperators & Virtual Environments*, *12*(4), 360-373.

Krijn, M., Emmelkamp, P. M., Olafsson, R. P., & Biemond, R. (2004). Virtual reality

exposure therapy of anxiety disorders: A review. *Clinical psychology review*, *24*(3), 259-281. DOI: 10.1016/j.cpr.2004.04.001

Krupiński, R. (2020). Virtual reality system and scientific visualisation for smart designing and evaluating of lighting. Energies, 13(20), 5518.

Kuliga, S. F., Thrash, T., Dalton, R. C., & Hölscher, C. (2015). Virtual reality as an empirical research tool—Exploring user experience in a real building and a corresponding virtual model. *Computers, environment and urban systems*, *54*, 363-375. DOI: 10.1016/j.compenvurbsys.2015.09.006

Lackey, S. J., Salcedo, J. N., Szalma, J. L., & Hancock, P. A. (2016). The stress and workload of virtual reality training: the effects of presence, immersion and flow. *Ergonomics*, *59*(8), 1060-1072. DOI: 10.1080/00140139.2015.1122234

Larsen, C. R., Oestergaard, J., Ottesen, B. S., & Soerensen, J. L. (2012). The efficacy of virtual reality simulation training in laparoscopy: a systematic review of randomized trials. *Acta obstetricia et gynecologica Scandinavica*, *91*(9), 1015-1028. DOI: 10.1111/j.1600-0412.2012.01482.x

Larsen, R. J., & Prizmic-Larsen, Z. (2006). Measuring Emotions: Implications of a Multimethod Perspective.

Li, R., van Almkerk, M., van Waveren, S., Carter, E., & Leite, I. (2019, March). Comparing human-robot proxemics between virtual reality and the real world. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 431-439). IEEE.

Lin, C. H., & Hsu, P. H. (2017). Integrating procedural modelling process and immersive VR environment for architectural design education. In *MATEC Web of Conferences* (Vol. 104, p. 03007). EDP Sciences.

Lloyd, J., Persaud, N. V., & Powell, T. E. (2009). Equivalence of real-world and virtual-reality route learning: A pilot study. Cyberpsychology & Behavior, 12(4), 423-427.

Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior*

research methods, instruments, & computers, *31*(4), 557-564.

Maffei, L., Masullo, M., Pascale, A., Ruggiero, G., & Romero, V. P. (2016). Immersive virtual reality in community planning: Acoustic and visual congruence of simulated vs real world. *Sustainable Cities and Society*, *27*, 338-345. DOI: 10.1016/j.scs.2016.06.022

Mania, K., & Chalmers, A. (2001). The effects of levels of immersion on memory and presence in virtual environments: A reality centered approach. *CyberPsychology & Behavior*, *4*(2), 247-264.

Michalski, S. C., Szpak, A., Saredakis, D., Ross, T. J., Billinghurst, M., & Loetscher, T. (2019). Getting your game on: Using virtual reality to improve real table tennis skills. PloS one, 14(9), e0222351.

Mihevic, P. M. (1981). Sensory cues for perceived exertion: a review. *Medicine and science in sports and exercise*, *13*(3), 150-163

Mine, M. R., Brooks Jr, F. P., & Sequin, C. H. (1997, August). Moving objects in space: exploiting proprioception in virtual-environment interaction. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (pp. 19-26).

Moghimi, M., Stone, R., Rotshtein, P., & Cooke, N. (2016). The Sense of embodiment in Virtual Reality. *Presence: Teleoperators & Virtual Environments*, *25*(2), 81-107. DOI: 10.1162/PRES

Mouraviev, V., Klein, M., Schommer, E., Thiel, D. D., Samavedi, S., Kumar, A., ... & Patel, V. (2016). Urology residents experience comparable workload profiles when performing live porcine nephrectomies and robotic surgery virtual reality training modules. Journal of robotic surgery, 10(1), 49-56.

Narasimha, S., Scharett, E., Madathil, K. C., & Bertrand, J. (2018, September). WeRSort: preliminary results from a new method of remote collaboration facilitated by fully immersive virtual reality. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 62, No. 1, pp. 2084-2088). Sage CA: Los Angeles, CA: SAGE Publications.

Noble, B. J., Borg, G. A., Jacobs, I. R. A., Ceci, R., & Kaiser, P. (1983). A category-ratio perceived exertion scale: relationship to blood and muscle lactates and heart rate. *Medicine and science in sports and exercise*, *15*(6), 523-528.

Ovaskainen, H. (2005). Comparison of harvester work in forest and simulator environments. Silva fennica, 39(1), 89-101.

Petukhov, I. V., Glazyrin, A. E., Gorokhov, A. V., Steshina, L. A., & Tanryverdiev, I. O. (2020). Being present in a real or virtual world: A EEG study. *International journal of medical informatics*, *136*, 103977. DOI: 10.1016/j.ijmedinf.2019.103977

Rebelo, F., Noriega, P., Duarte, E., & Soares, M. (2012). Using virtual reality to assess user experience. *Human Factors*, *54*(6), 964-982. DOI: 10.1177/0018720812465006

Riddle, K. (2010). Remembering past media use: Toward the development of a lifetime television exposure scale. *Communication Methods and Measures*, *4*(3), 241-255. DOI: 10.1080/19312458.2010.505500

Rizzuto, M. A., Sonne, M. W., Vignais, N., & Keir, P. J. (2019). Evaluation of a virtual reality head mounted display as a tool for posture assessment in digital human modelling software. Applied ergonomics, 79, 1-8.

Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., & Penn, P. R. (2000). Training in virtual environments: transfer to real world tasks and equivalence to real task training. Ergonomics, 43(4), 494-511.

Ruddle, R. A., Payne, S. J., & Jones, D. M. (1997). Navigating buildings in" desk-top" virtual environments: Experimental investigations using extended navigational experience. *Journal of Experimental Psychology: Applied*, *3*(2), 143–159. DOI: 10.1037/1076-898X.3.2.143

Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, *6*(4), 332-339. doi:10.1038/nrn1651

Skorupka, A. (2009, June). Comparing human wayfinding behavior in real and

virtual environment. In *Proceedings of the 7th International Space Syntax Symposium* (Vol. 104, pp. 1-7). Stockholm: KTH Royal Institute of Technology.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3549-3557.

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British Journal of Psychology*, *109*(3), 431-433.

Slater, M., Usoh, M., & Steed, A. (1994). Depth of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *3*(2), 130-144.

Steel Crate Games (2015) Keep Talking and Nobody Explodes. Video Game (VR). https://www.igdb.com/games/keep-talking-and-nobody-explodes.

Stevens, J. A., & Kincaid, J. P. (2015). The relationship between presence and performance in virtual simulation training. *Open Journal of Modelling and Simulation*, *3*(02), 41–48.

Stone, R. T., Watts, K. P., Zhong, P., & Wei, C. S. (2011). Physical and cognitive effects of virtual reality integrated training. Human factors, 53(5), 558-572.

Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using presence questionnaires in reality. Presence, 9(5), 497-503.

Villani, D., Repetto, C., Cipresso, P., & Riva, G. (2012). May I experience more presence in doing the same thing in virtual reality than in reality? An answer from a simulated job interview. *Interacting with Computers*, *24*(4), 265-272. DOI: 10.1016/j.intcom.2012.04.008

Weber, S., Weibel, D., & Mast, F. W. (2021). How to get there when you are there already? Defining presence in virtual reality and the importance of perceived realism. Frontiers in psychology, 12.

Witmer, B. G., & Kline, P. B. (1998). Judging perceived and traversed distance in virtual environments. *Presence*, *7*(2), 144-167.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments:

A presence questionnaire. *Presence*, *7*(3), 225-240.

Yildirim, Ç., Bostan, B., & Berkman, M. I. (2019). Impact of different immersive techniques on the perceived sense of presence measured via subjective scales. *Entertainment Computing*, *31*, 100308.

Zindulka, T., Bachynskyi, M., & Müller, J. (2020, April). Performance and Experience of Throwing in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-8).