UNIVERSITY OF OULU
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Aleksi Seppänen
Markus Hirsimäki
Patrik Pyykkönen

EXPERT EVALUATION OF ASPECTS RELATED TO VIRTUAL REALITY SYSTEMS AND SUGGESTIONS FOR FUTURE STUDIES

Bachelor's Thesis
Degree Programme in Computer Science and Engineering
June 2020

Seppänen A., Hirsimäki M., Pyykkönen P. (2020) Expert Evaluation of Aspects Related to Virtual Reality Systems and Suggestions for Future Studies. University of Oulu, Degree Programme in Computer Science and Engineering, 75 p.

ABSTRACT

In this bachelor's thesis, we review existing quantitative and qualitative research on virtual reality systems. We then present suggestions for performing a future study to combine the objective and subjective measurements of virtual reality experience. Additionally, we adapted an existing heuristics-based expert evaluation method to suit evaluating virtual reality systems. Using our method, we performed the expert evaluation on a selection of five modern consumer virtual reality systems to understand the connections between the subjective experience and the physical variables related to the virtual reality system.

From this evaluation, we present findings that are used to construct discussion and to draw conclusions on these said connections. We found the most prominent conclusion to be that the experience of virtual reality is highly subjective and dependent on the content being viewed in virtual reality. Additionally, we concluded that some of the most important aspects in need of improvement are display resolution, lens design, user ergonomics, and lack of wirelessness. Finally, we state that two optimization problems are present; the first one being the optimization required to design a virtual reality system and the second one being the act of choosing a system to match a consumer's preferred content.

Keywords: Virtual reality, head-mounted display, user experience, subjective, consumer, suggestion

Seppänen A., Hirsimäki M., Pyykkönen P. (2020) Asiantuntija-arviointi virtuaalitodellisuusjärjestelmiin liittyvistä tekijöistä ja ehdotuksia tuleville tutkimuksille. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 75 s.

TIIVISTELMÄ

Tässä kandidaatin tutkielmassa käymme läpi aiempaa kvantitatiivista ja kvalitatiivista tutkimusta virtuaalitodellisuusjärjestelmistä. Esitämme myös ehdotuksia myöhempää tutkimusta varten virtuaalitodellisuuteen liittyvien objektiivisten ja subjektiivisten mittausten yhdistämiseksi. Tämän lisäksi adaptoimme aiemman heuristiikkapohjaisen asiantuntija-arvioinnin sopimaan virtuaalitodellisuusjärjestelmien arviointiin. Käyttäen metodiamme toteutimme asiantuntija-arvioinnin viidellä modernilla kuluttajakäyttöön tarkoitetulla virtuaalitodellisuusjärjestelmällä ymmärtääksemme yhteyksiä subjektiivisen kokemuksen ja niiden fysikaalisten muuttujien välillä, jotka liittyvät virtuaalitodellisuusjärjestelmiin.

Esitämme tämän asiantuntija-arvioinnin löydöksiä, ja luomme niiden avulla keskustelua, jonka avulla teemme mainittuihin yhteyksiin liittyviä johtopäätöksiä. Tärkein johtopäätöksemme oli se, että virtuaalitodellisuuden kokemus on erittäin subjektiivinen ja riippuvainen siitä sisällöstä, jota virtuaalitodellisuudessa koetaan. Aiemman lisäksi toteamme, että merkittävimpiä kehitystä kaipaavia osa-alueita ovat näytön resoluutio, linssien suunnittelu, käyttäjäergonomia ja langattomuuden puute. Viimeisenä totesimme, että virtuaalitodellisuusjärjestelmiin liittyy kaksi optimointiongelmaa; ensimmäinen liittyy järjestelmän suunnittelussa tapahtuvaan optimointiin, ja toinen liittyy sellaisen järjestelmän valitsemiseen, joka sopii kunkin kuluttajan suosimaan tarkoitukseen.

Avainsanat: Virtuaalitodellisuus, virtuaalitodellisuuslasit, käyttäjäkokemus, subjektiivinen, kuluttaja, ehdotus

TABLE OF CONTENTS

ABSTRACT
TIIVISTELMÄ
TABLE OF CONTENTS
FOREWORD
LIST OF ABBREVIATIONS
1. INTRODUCTION ... 8
1.1. History and Motivation ... 8
1.2. Examples of Current Consumer VR Systems ... 9
1.3. Purpose, Objective and Scope ... 11
1.4. Contributions ... 11
2. BACKGROUND AND RELATED WORK ... 12
2.1. Displays ... 12
2.2. Lenses ... 13
2.3. Controllers ... 14
2.4. Tethered and Wireless Solutions ... 15
2.5. Audio ... 15
2.6. Tracking ... 15
2.6.1. Tracking Performance ... 17
2.7. User Experience ... 17
2.7.1. Cybersickness and It's Related Sicknesses ... 18
2.7.2. Motion-To-Photon Latency ... 19
2.7.3. Quality of Experience ... 22
2.7.4. Immersion ... 23
2.8. Content ... 24
2.9. Industry and Consumer Standards ... 25
3. SUGGESTIONS FOR FUTURE STUDY ... 26
3.1. Suggested Problem Statement and Research Question ... 26
3.2. Defining the Suggested Test Methodology ... 26
3.3. Defining the Suggested VR Systems to Be Used ... 27
3.4. Qualitative Test ... 27
3.4.1. Defining the Test Content ... 27
3.4.2. Defining the Test Subjects ... 28
3.4.3. Defining the Data Collection ... 29
3.4.4. First Pilot Study ... 29
3.4.5. Finalized Test Procedure ... 32
3.5. Quantitative Tests ... 34
3.5.1. Defining the Measuring Equipment ... 34
3.5.2. Defining the Variables to Be Measured ... 34
4. EXPERT EVALUATION ... 35
4.1. Expert Evaluation Problem Statement and Research Question ... 35
4.2. Adaptation of the Method ... 35
4.3. Defining the VR-System to Be Used ... 36
4.4. Defining the Test Tasks ... 36

4.4.1. Task 1: Videos 37
4.4.2. Task 2: Beat Saber 37
4.4.3. Task 3: Richie's Plank Experience 38
4.4.4. Task 4: SUPERHOT VR 40
4.4.5. Excluded Tasks 40
4.5. Testing Environment 41
4.6. Experts 41
4.7. Detailed Testing Method 42
4.7.1. Adapted Heuristics 43
4.7.2. Adapted Severity Classification 44
5. RESULTS AND EVALUATION 45
5.1. Displays 45
5.2. Lenses 45
5.3. Controllers 46
5.4. Tethered and Wireless Solutions 46
5.5. Audio 47
5.6. Tracking 48
5.7. User Experience 48
5.8. Content 49
5.9. System Rankings 50
6. DISCUSSION 52
6.1. Displays 52
6.2. Lenses 52
6.3. Controllers 53
6.4. Tethered and Wireless Solutions 53
6.5. Audio 54
6.6. Tracking 54
6.7. User Experience 55
6.8. Content 56
6.9. Summary 57
7. CONCLUSION 58
8. REFERENCES 59
9. APPENDICES 63

FOREWORD

We would like to thank Prof. Timo Ojala for great stories and guidance trough our work. We would also like to thank PhD Matti Pouke and Dr. Tech. Aku Visuri for their guidance and feedback.

Oulu, June 12th, 2020

Aleksi Seppänen
Markus Hirsimäki
Patrik Pyykkönen

LIST OF ABBREVIATIONS

Please note that we intentionally capitalize all abbreviations for the sake of clarity. Plurals of abbreviations use lowercase s: e.g. VEs.

CA	Chromatic aberration
DOF	Degrees of Freedom
FOV	Field of View
FPS	Frames per Second
HMD	Head-mounted Display
IMU	Inertial Measurement Unit
IPD	Interpupillary Distance
IR	Infrared
ITU	International Telecommunication Union
M2P	Motion-to-photon
MOS	Mean Opinion Score
PPD	Pixels per Degree
SDE	Screen Door Effect
VE	Virtual Environment
VR	Virtual Reality

1. INTRODUCTION

Currently, virtual reality (VR) systems lack comprehensive evaluation methods that can measure both the VR systems themselves and their respective user experiences. This is in contrast to related technologies, such as displays, that have well-established standards of measurement and no similar need for measuring the user experience.

Various methods for measuring single variables, like sensor, drifting, and motion-to-photon (M2P) latency, have been proposed [1, 2]. These methods, however, do not combine the measured variables into metric for describing the experience, nor do they take into account the subjectiveness of the VR experience. We state that the subjectiveness of the VR experience might be one of the reasons for the lack of standardization; this builds on the fact that the experience of VR is inherently humane, and humane experiences are hard to measure.

These problems in comparing VR systems and user experience lead to difficulties in consumer and professional comparisons of different systems [3, 4, 5]. The results are long comparison tables of physical variables, containing multitudes of technical data, with various relevancies. While this is acceptable in some professional situations it may not be optimal for consumers. This leads to our thesis where we perform an expert evaluation to understand the connections between the user experience and the physical attributes of the systems as well as present suggestions for a systematic study in the future.

1.1. History and Motivation

The history of VR depends partly on the way we define the concept of VR. For example, in his book about VR [6], LaValle reviews the evolution of VR from ancient cave paintings, first motion pictures, and simple computer games to the current consumer market head-mounted displays (HMDs) [6]. In this thesis, we are focusing on modern consumer-friendly VR systems.

The first VR system utilizing an HMD is widely considered to be invented and built by computer scientist Ivan Sutherland in 1968 [6, 7]. For a long time, HMDs were mainly used in industrial or research setups, and the prices were unreachable to consumers. The first attempt with commercial VR devices was by a company called VPL with the release of Eyephone HMD in 1988. In the 1990s there were many attempts with consumer VR, like Sega VR, and Nintendo Virtual Boy console. These VR systems had some familiar elements from current HMDs, but the technology was too rudimentary and unwieldy for consumers at the time.

The first wave of popularity ended because the technology was not ready to offer adequate user experience. VR, however, found its place in many military, aerospace, and other specialized applications. The latest generation started when a company called Oculus started a Kickstarter campaign to fund the development of the Oculus Rift VR system. This was significant since it was aimed towards general use at a price reachable to consumers. These new VR systems still weren't ready to offer the quality and ease of use that consumer masses wanted and the hype settled once again. However, consumer VR systems are steadily becoming more commonplace, and we except that in the near future the demand starts to rapidly increase as technology develops and the prices decrease. This means that creating a standardized way of measuring VR systems and understanding the metrics that create an enjoyable user experience is becoming increasingly important and useful.

1.2. Examples of Current Consumer VR Systems

In this section, we present some specifications of consumer VR systems in table form alongside pictures of the same consumer VR systems. The same systems are used in our suggested future study and consequently in our expert evaluation. The terms used in the table 1 are elaborated on in chapter 2.

The images a) through e) in the figure 1 present the VR systems seen in table 1. In order, the systems are Oculus Quest, Oculus Rift S, HP Reverb, HTC Vive Pro, and HTC Vive Pro with wireless adapter.

Table 1. Specifications of the VR-systems used. HTC Vive Pro with wireless adapter is not presented in the table due to it being duplicate of the HTC Vive Pro column. Data has been collected from manufacturers specifications and an online source [8].

Spesification	Oculus Quest	Oculus Rift S	HP Reverb	HTC Vive Pro
Release Date	2019-5-21	2019-05-21	2019-05-06	2018-04-05
Positional tracking	Inside-out	Inside-out	Inside-out	Room Scale Tracking
Display	OLED	LCD	LCD	AMOLED
Resolution (Per eye)	1440x1600	1280x1440	2160x2160	1440x1600
Refresh rate	72 Hz	80 Hz	90 Hz	90 Hz
Field of view	100°*	115°*	114°	110°
PPD	14.4*	11.13*	18.94*	13.09*
Weight	570g*	563g*	0.5kg	563g*
IPD adjustment	Physical	Software	Software	Physical
Platform	Standalone**	Oculus App, SteamVR	Windows Mixed Reality, SteamVR	SteamVR

* Estimation. Not specified by the manufacturer.
** Can be connected to a PC with appropriate cable.

(a) Components of Oculus Quest VR system. Charger is excluded from the image.

(b) Components of Oculus Rift S VR system.

(c) Components of HP Reverb VR system.

(d) Components of HTC Vive Pro VR system. Power cords for the base stations are excluded from the image for clarity.

(e) Components of HTC Vive Pro VR system with wireless adapter. PCI-E expansion card of the wireless transmitter is not shown in the image. Also, power cords for the base stations are excluded from the image for clarity.

Figure 1. The systems used in expert evaluation. All images by Patrik Pyykkönen.

1.3. Purpose, Objective and Scope

The purpose of this work is to present an expert evaluation of current consumer VR systems alongside suggestions for performing a systematic study in the future. We also produce a brief literature review where we show research on quantitative measurements of VR systems and qualitative measurements of user experience in VR.

The main objective of this work is to lay the groundwork for further research and to find connections between the VR user experience and the physical attributes of the VR system used.

The scope of this work is intentionally limited to suit the limitations of a bachelor's thesis. We want to clarify that the work on this thesis was interrupted by the COVID-19 pandemic and that our initial scope included performing the study we are now presenting as a suggestion in section 3. Due to the cancellation of the planned study, we opted to use expert evaluation in place of the previous plan, which resulted in changes to the structure of our thesis.

1.4. Contributions

Generally, the amount of work on our thesis was evenly split among the three authors. During writing, our workflow was for Aleksi Seppänen to produce a rough sketch of various sections which were then refined by Patrik Pyykkönen and finalized by Markus Hirsimäki. When testing was performed during the pilot of our suggested future study, all of us worked together. Finally, during the expert evaluation phase, we all provided equal contributions successively.

2. BACKGROUND AND RELATED WORK

In this chapter, we provide brief backgrounds for the different aspects of VR systems we later explore in our expert evaluation. Additionally, we provide a few detailed lists of variables related to these aspects. We also discuss existing research papers and their relationships with our work. This discussion can be broadly categorized into qualitative and quantitative sections that are intertwined.

2.1. Displays

Displays in VR systems and more specifically in HMDs are an important part of any VR experience. *Resolution* of the display defines a notable part of the sharpness of the image in an HMD. Increased resolution helps the user distinguish more details in the VE and lessens the pixelation effects. Communities revolving around VR generally accept 60 pixels per degree (PPD) to be the threshold where pixels will become imperceptible. To our understanding, this is based on the visual acuity of normal vision as defined by the Snellen chart. The chart defines that normal vision should be able to discern the orientation of optotype covering 5 arc minutes of the visual field. The smallest details in this optotype cover one arc minute which is equivalent to 1/60 of a degree. This in turn translates to normal vision's ability to resolve details being surpassed at 60 PPD [9]. HMDs in our tests are still way behind this with values between 10 and 20 PPD.

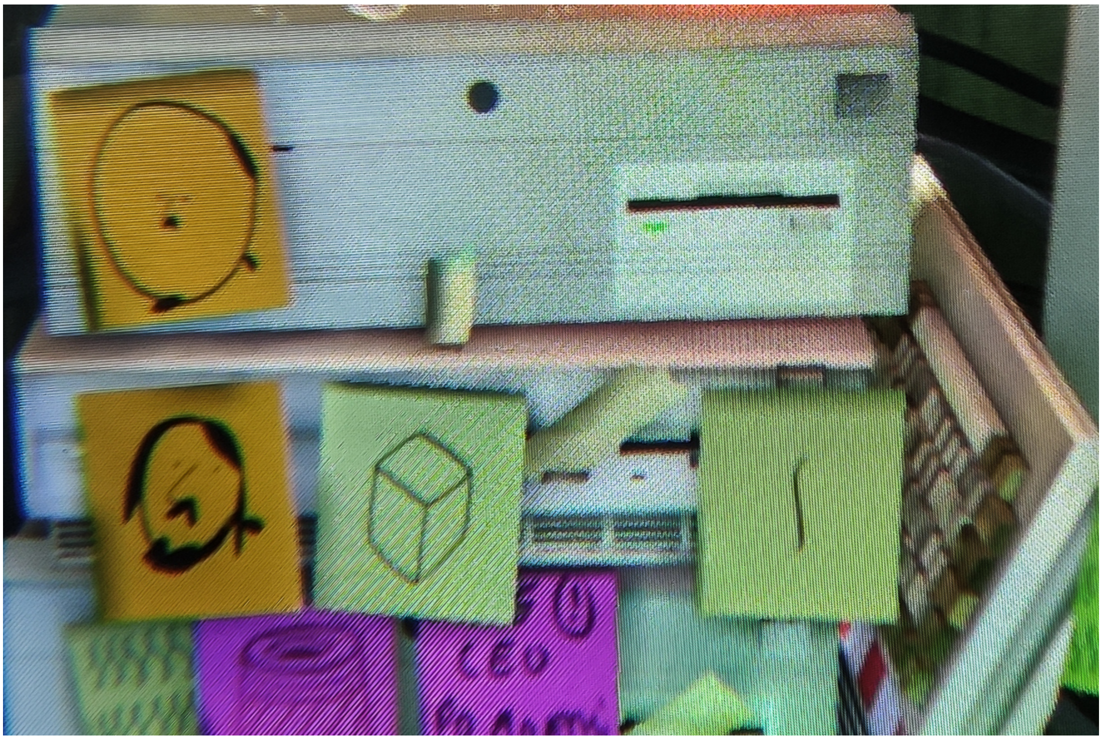


Figure 2. Effects of low resolution and screen door effect in an HMD's image.

Screen-door effect (SDE) is an aberration where the image in an HMD looks like it is being viewed through a fine anti-insect screen or mesh [10]. SDE is caused by black gaps between the subpixels in the display which become visible to the user due to the large magnification that is needed for the displays to cover a large field of view [11]. SDE can be reduced by increasing the display resolution, but also, for example, with optical methods such as diffractive films over the display acting basically as optical low-pass filters [11, 10]. Figure 2 shows a magnified HMD screen where low resolution and SDE affect the image.

Field of view (FOV) describes how large viewing angles the HMD's display covers. If the FOV is small enough, the user experience is might be similar to watching through binoculars. Conversely, high FOV can better replicate normal vision as dark borders are less noticeable. *Mura* in displays means a noticeable disparity of brightness between different sections of the display or even neighboring pixels. *Contrast* defines the difference between the darkest and brightest colors the display can show at the same time. Insufficient contrast can result in a lack of detail in dark scenes as the display might not be able to differentiate between dark grey and true black.

Other notable display related variables are response time, refresh rate, brightness, color accuracy, ghosting, motion blur, backlight period, and backlight frequency. These variables are not generally used in our thesis, and thus not elaborated on, but we want to acknowledge their existence as some distracting effects could have been noted in the displays if these variables had not been adequate already in our tests.

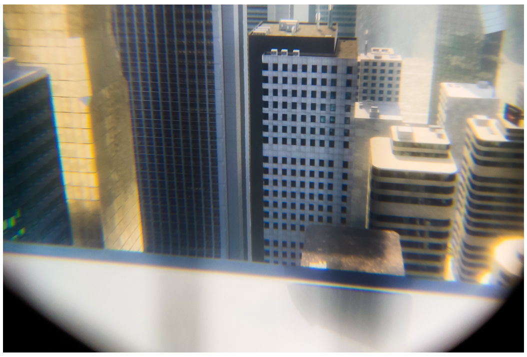
2.2. Lenses

Lenses are used in HMDs to help the eyes to focus on display and to properly project the view. Lenses are prone to causing various distortions or aberrations, especially in the edge areas of the visual field, and thus they should be carefully designed so as not to disturb the user experience [6]. The adjustability of the lenses is an important tool for achieving a proper focus. *Interpupillary distance adjustment* enables users to set the lateral distance between the lenses to match the distance between their eyes. This is important as the interpupillary distance (IPD) varies from person to person, the average being around 63 mm. [12].

Optical distortions, as an example, include barrel- and pincushion distortion, where the image seems to be either bulging outwards or inwards and straight lines appear curved. *Reflections* and flares inside the lenses distract and block the user from seeing the image clearly. They are caused by light from the display or by ambient light leakage to the lenses.

Chromatic aberration (CA) is a typical problem in high contrast areas of an image where colors in the opposite ends of the visible spectrum disperse due to refracting different amounts in the lens [6]. This can typically be seen as colored fringes in high contrast edges, and an example is present in figure 3b. *God rays*, also known as crepuscular rays, are also typically seen in high contrast areas that the lens makes glow in light ray or light beam fashion. The rays are more prominent in high contrast areas since they offer a background onto which the light can bleed and subsequently be seen. An example of this is provided in figure 3a.

(a) God rays



(b) Chromatic aberration

(c) Fresnel banding 1

(d) Fresnel banding 2

Figure 3. Examples of lens aberrations in HMDs. The images do not perfectly represent what can be seen with a naked eye. Images by Patrik Pyykkönen.

Lens type refers to either the material of the lens or the chosen lens technology. The chosen material is typically either glass or an optically suitable plastic. The lens technology is commonly either a traditional lens or a Fresnel lens. The choice between the Fresnel lens and a traditional lens offers a balance between possible optical distortions and higher weight [6]. Hybrid solutions that combine a Fresnel lens and a traditional lens also do exist [13]. Two examples of banding aberrations caused by Fresnel lens are shown in figure 3c and 3d.

Other noteworthy lens-related factors include the following; spherical aberration, coma, depth of focus, and clarity. These factors are not used widely in our work. This is why they are not further explained; we, however, want to acknowledge their existence as they are related to common distortions and problems.

2.3. Controllers

Most consumer VR systems use handheld controllers for giving input into the virtual environment (VE) but other solutions do also exist. Examples of these different approaches include image recognition of the user's hands, controllers that grip the user's hands like a clothespin instead of being held, and controllers that have haptic force-based feedback built into a glove form. As an example of an image recognition based solution, Oculus Quest has hand tracking that enables rendering the users actual fingers moving in the VE. To an extent, the solution enables a user to interact with VE without controllers. More extensive control systems include tracking other parts of the user's body such as feet and some of the more extreme approaches track the user's entire body with the help of a specialized full bodysuit.

Different systems use different designs for their controllers and these systems sometimes offer vastly different control schemes. A typical consumer controller includes at least one of each of the following: joystick and or trackpad; button and trigger.

Various factors affect the general usability of the common handheld controller; grip, form factor, weight, battery life, handedness, and sturdiness.

2.4. Tethered and Wireless Solutions

Different VR systems offer different approaches to wirelessness. Some systems are natively standalone and do not require tether, an example of this sort of system is Oculus Quest. Other systems are natively tethered and require the presence of a tether and PC to provide the information to be rendered. Examples of these sorts of systems include HTC Vive Pro and HP Reverb. Aftermarket add-ons also exist to convert a tethered system into a wireless one with the addition of antenna and battery. Natively wireless smartphone-based VR systems also exist.

The system's design choice on having tether or on being wireless results in several differing requirements. Examples of possible requirements for the wireless system include antenna, battery pack, smartphone, or completely standalone design. Some example tethered requirements include beacon systems, desktop PC, and external primary display. It should be noted that some of these listed requirements are interchangeable between tethered and wireless designs.

2.5. Audio

An audio feed is an integral part of the VR experience. Commonly this is achieved by either having speakers built into the HMD or by having a folding over-ear headphone solution connected to the HMD. Many systems also offer the ability to connect external headphones straight into the headset. In VEs the user often able to move freely around, and so the distance and direction from audio sources change. For natural experience, it is important to mimic these movements by changing the audio level and directions of sounds. Even more in-depth approaches would include modeling the acoustics and propagation of sound waves in the VE [6].

Some common factors that affect the audio experience are noise cancellation, noise isolation, spatial audio tracking, surround sound, general sound quality, and audiovisual sync. These factors together contribute to the immersiveness of the audio which in turn contributes to the total perceived immersion.

2.6. Tracking

A VR system needs a reliable tracking method to resolve the position and movements of the user with high precision and speed. The reliability of this system is important as inaccuracies could interfere with the user's task or result in nausea. There are multiple types of *tracking technologies* used in different VR-systems. The amount of *degrees of Freedom* (DOF) denotes what sort movements the tracking system can differentiate. Systems with 3 DOFs can only track rotational movements around three axes. Systems with 6 DOFs are also able to track translational movements in 3D space which is necessary for an immersive experience in VR.

Outside-in tracking means that the sensors tracking the HMD and controllers are located externally to the HMD. In Oculus Rift CV1 for example, the outside-in tracking system consists of one or two infrared (IR) cameras that face the HMD and controllers. They track the known constellation of IR LEDs embedded in the devices. [14]

Inside-out tracking systems, in contrast, have the cameras built into the HMD. The cameras track recognizable features from the environment to solve the user's position and movements. Controllers are still embedded with LEDs but they are tracked by the HMD itself, which might limit the area where the controllers can be tracked in.

It should be noted that while the actual technical implementation of a given system might not be specifically outside-in or inside-out, the terms are used to colloquially differentiate between systems that have beacons or base stations and between systems that do not have tracking beacons or base stations. These different implementations also provide different tracking *area sizes*.

As an example approach, HTC uses a mixed solution to track its HMDs and controllers. The previous VR system generation's HTC Vive uses base stations sending alternating IR LED array pulses and sweeping IR laser pulses. The Vive headset has known constellation of photodiodes that are used to capture the different light pulses so that position and orientation of the HMD can be calculated from the time difference each photodiode sees. [14]

In addition, all HMDs and controllers typically include an inertial measurement unit (IMU). The IMU usually consists of an accelerometer, gyroscope, and magnetometer. IMU aids the tracking especially at times when the main system is unable to see the tracked object.

There exist several solutions to track various objects in addition to the HMD and its controllers. Additional trackers similar to those in controllers can be added to track various body parts such as legs and body joints. Full-body tracking suits have also been developed. With these solutions, a proper virtual avatar can be rendered in the VE instead of showing just virtual hands or controllers and head. *Eye tracking* is implemented in some HMDs to track the direction of the user's gaze. This could be used in the content being viewed in VR or in the rendering pipeline to prioritize the rendering efforts. This sort of rendering optimization is generally called foveated rendering.

The functionality of the tracking can be measured with multiple variables, but these are not normally available for consumer comparisons as they require highly specialized measurement devices. The amount of *Drifting* describes how much the tracked position drifts over time either rotationally or translationally. *Sensitivity* measures how small movements the system can detect. *Jitter* measures the unwanted small noise-like movement that the tracking system imposes in the image even though everything in the physical world remains still. Jitter can occur in different parts of the system independently from the other parts.

Movement prediction over and undershoot are artifacts resulting from a VR system predicting the user's motion incorrectly. This sort of predicting can be used to reduce M2P latency, since predicting the future the rendering can be started before the user reaches the predicted position. Undershoot and overshoot occur when the user's predicted motion differs from the motion performed by the user.

2.6.1. Tracking Performance

Multiple different approaches are possible when tracking performance is measured. Different aspects of tracking require different measurement methods; variables such as jitter and prediction overshoot or undershoot are measurable with a VR multimeter like BUDDY-3 [1]. Multimeters, however, may not be able to measure all tracking related variables. The below study presents an example of such variables.

Comparison of Oculus Rift and HTC Vive: Feasibility for VR-Based Exploration, Navigation, Exergaming, and Rehabilitation

As an example of prior quantitative research, we briefly highlight the above study by Borrego et al. The study takes a look into external position tracking in VR systems. The study compared Oculus Rift CV1 and HTC Vive in terms of tracking the working area, accuracy, jitter, and working range. Additionally, they measured how the tracking accuracy is affected by distance. The systems use different tracking methods but both can be considered to be outside-in based.

The researchers created a 10x10 meters sized grid with 0.5 meter spacing between lines. In total, they had 121 points where the position could be accurately measured using heights 1.3 and 1.7 meters from the ground. They estimated the accuracy as the mean difference between the position of the HMD measured in the real world and position estimated by HMD's tracking system. They also measured jitter as defined by the standard deviation in the estimated position in the time interval. [14]

Their results showed that the true working areas of tracking were much larger than what the manufacturers recommended. Oculus had a recommended area of 2.75 m^2 but it worked in an area of 11.75 m^2. HTC Vive had a larger recommendation of 6.25 m^2 and it was still able to track over twice the area of Oculus, resulting in 24.87 m^2. Accuracy from different heights inside the recommended working area ranged from 0.61 to 0.76 cm for Oculus Rift CV1 and 0.58 to 1.22 cm for HTC Vive. Outside the recommended area accuracy dropped slightly on both headsets, 0.92 to 1.03 cm for Oculus Rift CV1 and 0.85 to 1.49 cm for HTC Vive. HTC Vive also seemed to have a slightly larger error in the standing position (1.7 m) than in lower positions. Jitter acted similarly and Oculus performed slightly better than HTC Vive.

2.7. User Experience

Comfort and ease of use are partially subjective experiences. Many of the variables that measure these are often also dependent on context and they might change greatly based

[1] `https://www.optofidelity.com/offering/products/buddy-3`

on it. As an incomplete example list, these are some of the features affecting the user experience: user comfort when the system is being worn; ease of use when the system is (un)equipped, compatibility with eyeglasses or prescription HMD lenses, weights of the various parts of the system, amount of set-up required and safety features. As can be seen, creating a comprehensive list of all these variables is not practical. In our expert evaluation, we focused on the variables that we noted to affect the user experience within the systems and contents we used.

In addition to the previously discussed, it should be noted that a given VR system's user experience depends on the system's performance. This performance covers everything from the PC used to drive the system to the chosen lens technology and tracking. A systems frames per second (FPS) defines how smooth the movement of content appears to be. The FPS is dependent on the display's capabilities and the rendering power available. A study by Ramamami et al. states, that the minimum for immersive user experience is 60 FPS [15]. Usually, an even higher rate is preferred, as can be seen from the VR systems selected for this study.

2.7.1. Cybersickness and It's Related Sicknesses

The lack of cybersickness is crucial in providing a good VR experience. We want to clarify the use of the terms *motion sickness*, *simulator sickness* and *cybersickness* in the context of our thesis. In this section, we briefly describe our definitions. These definitions are based on established prior research [16, 17, 18]. It is worth noting that some studies use these terms interchangeably. These definitions are presented so that we can better discuss studies that refer to these terms.

All of these sicknesses happen when there is a mismatch between different sensory inputs, mainly between the vestibular and visual systems. Motion sickness happens when the vestibular system perceives motion but the visual system lacks sufficient indication of said motion. Cybersickness is the opposite of this effect; the vestibular system perceives little to no movement compared to the optical flow. This perceived illusion of self-motion in VR is called vection. Between these lies simulator sickness, where the vestibular system and visual system both perceive motion but they mismatch to a degree. It should be noted that other symptoms are also possible in addition to those related to vection. Examples of these are oculomotor problems such as headache and eyestrain. Next, we present some relevant studies related to cybersickness. [16, 17, 18]

Comparing the onset of cybersickness using the Oculus Rift and two virtual roller coasters

In the above study by Davis et al. the researchers focused on discussing the possible causes of cybersickness, as well as classifying it and measuring it objectively. The study also classifies types of distresses and symptoms related to simulator sickness. [19]

In the study an experiment was performed; participants were placed into two different VR roller coasters. The roller coasters had different designs, with the other one being more fast-paced and realistic. The test subjects were then asked to rate their experiences on a scale from 0 to 10, with 0 being no nausea and 10 being very nauseous. The roller coaster experiences lasted up to 14 minutes, with nausea being rated every 2 minutes. The subjects could also stop the experiment before the 14-minute mark if they felt too nauseous. [19]

The authors presented multiple research questions, with the most important in the context of our thesis being: 'Can cybersickness be simply induced to allow a more detailed study of the physiological changes that occur? [19]'. Among other things, they conclude that having a more realistic environment and higher optical flow in one of the roller coasters produced a greater chance of inducing cybersickness. This is supported by other studies, such as the one performed by Ashutosh Singla et al. [20], and the book by LaValle [6]. These results imply that it is possible to intentionally or accidentally create disorienting experiences even if the technology offered by the VR system is ideal.

These findings are relevant to our suggested study and expert evaluation as both measure variables that are subjective and dependent on the content being presented in VR. Both of these studies, the suggested one and expert evaluation, require using test contents to perform the study.

The effects of habituation and adding a rest-frame on experienced simulator sickness in an advanced mobility scooter driving simulator

In the above study by Heutink et al. the researchers measured the effects of habituation and rest frames in the context of simulator sickness. The study was conducted by having the participants drive on a mobility scooter simulator over two sessions with 24 hours in between. Their relevant research question was if habituation can decrease simulator sickness. The conclusions state that over the 24 hour period habituation did occur in measurable amounts [21]. This knowledge was used in the design of our suggested user study to minimize the possible bias from the habituation of test subjects.

Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness

Kennedy et al. performed the above study where they devised a simulator sickness questionnaire. The paper presents a justification for separating simulator sickness from motion sickness. These justifications are used in our definitions of cybersickness and its related sicknesses. [18]

Additionally, the study presented a simulator sickness questionnaire that has become an established way of measuring sickness also in the context of VR. In the section 3 where we present our suggested study methods the simulator sickness questionnaire is a noteworthy part of the measuring methodology.

2.7.2. Motion-To-Photon Latency

M2P latency is one of the most studied objectively measurable attributes of HMDs as it has an immediate and noticeable impact on user experience. It has been stated, that M2P latency should ideally be under 20 ms to avoid negative effects [15]. There has been discussion about even lower delays, especially for augmented reality solutions. Increased latency can induce an array of problems ranging from a sluggish feeling response time to disorientation. Various ways have also been designed to combat this latency, for example, predicting the user's movement. Another solution to this us using asynchronous time warp [15] where frames are produced by interpolating. Due to the importance of M2P latency, we want to highlight some studies that have measured this variable. The latency is measured in various ways within these studies. A more general approach to measuring it, alongside other variables related to it, like prediction overshoot and undershoot, is possible with VR multimeters [2]. The use of these multimeters is briefly discussed in section 3.

Time Sequential Motion-to-Photon Latency Measurement System for Virtual Reality Head-Mounted Displays

In their article, Choi et al. proposed a method for measuring M2P latency in HMDs at any moment of movement. They designed a way to artificially move HMDs in two DOFs with high accuracy direct current servo motors. The movement was measured using incremental encoders for precise measurements of axis movements. A test program was designed to show different measurements based on the angular position of headset calculated by the HMD's IMU, and overlay those as four grayscale areas on top of actual content. The movement of the image was then analyzed using photodetectors that measured the grayscale images. To calculate the actual M2P latency, the HMD was first moved with the direct current servo motors, and the movement was simultaneously measured with the incremental encoders using an oscilloscope. When the IMU sensed the movement of the HMD, the appropriate grayscale values were rendered to the measurement areas with the rest of the image. Afterwards, the difference between physical rotation measurement and photodetector based measurement was calculated to find the M2P latency. [1]

An experiment with the test setup was carried out using Oculus Rift DK2 HMD and powerful PC. Researchers performed the measurements using different scene complexities to figure out how the system workload affects M2P latency. At minimal workload (2 objects, 800 000 vertices) the average M2P latency was 46.55 milliseconds. The second scene contained 2.4 million vertices which resulted in an average M2P latency of 63.22 ms. Consequently, 4.8 million resulted in 101.29 ms and 9.5 million vertices already average latency of 154.63 ms. Therefore it can be deduced that rendering workload is a major factor M2P latency in HMDs and must be taken account when measuring devices. [1]

In the context of our future study suggestion, this results in an important note; at what sort of rendering workload should M2P measurements be performed? This question is left unanswered as it is not in our scope, it is however left in as an important note to consider as any sort of work towards creating a measurement standard should carefully assess it.

[2] `https://www.optofidelity.com/offering/products/buddy-3`

Performance Measurements of Virtual Reality Systems: Quantifying the Timing and Positioning Accuracy

Chang et al. created a test setup utilizing an accurate rotary platform with one DOF, external display to show contents of HMD, and a high-speed camera to record both the HMD and the monitor. With this, they measured both timing accuracy and positioning accuracy of Oculus Rift DK2, 3Glasses D2, and Samsung Gear VR and Google Cardboard using Samsung Galaxy S6. Experiments were repeated twenty times for each VR device with different angular movements and system workloads. Movement of device and image was recorded at 240 FPS using an Apple iPhone 6 as high-speed camera, and special calibration markers were added to the HMD or the rotating platform. This kind of setup enabled measuring HMDs using almost any VR application, unlike some other methods that need specialized content to be shown in glasses. It remained unclear whether they considered internal latencies of the external monitor when analyzing the results. [2]

After post-processing, the researchers calculated the initial delay between physical movement and HMD content, and likewise the settling delay when motion stopped. The initial delay for 3Glasses D2 and Oculus DK2 averaged around 44 and 48 milliseconds, and apparently, the speed of the movement made no clear difference in results. The scene complexity had again a negative impact on latency. Interestingly they found out that settling delays were longer than initial: 3Glasses D2 60 ms, Oculus DK2 75 ms, and for mobile VR 80 and 90 ms. Scene complexity had the same effect as in initial delay, but altering the rotation speed produced some differences that remained unclear to the research team. [2]

They next analyzed the overshoot latency, and results showed that the 3Glasses D2 had zero correction latency, meaning that it didn't implement any prediction in the movement. Oculus Rift DK2 instead had a latency of 40 ms, Gear VR 80 ms, and Google Cardboard 90 ms. Correction latency grows as the speed of rotation grows, but according to researchers dead reckoning still contributed to the system responsiveness positively. [2]

As previously mentioned in section 2.6 this sort of movement prediction introduces new types of errors into the system when the predictions prove to be incorrect. The movement prediction, however, should still be able to increase the responsivity of the system. As in the context of M2P latency and rendering workload, work towards measuring standards should decide on a systematic approach towards assessing movement prediction.

Estimating the Motion-to-Photon Latency in Head Mounted Displays

Zhao et al. mounted an Oculus Rift DK2 on a freely swinging pendulum to introduce a sinusoidal motion to measure both the translational and rotational M2P latencies. Similar to [1] they mapped the brightness of the display to change according to the movement of the HMD. In this research, the Oculus tracking camera was used to measure also the translational movements. The rendering workload and consequently its effects in latency were minimal in the study as only a changing grayscale image was rendered. Initial angles of 60, 75, and 90 degrees on the pendulum were used to get data with different accelerations and speeds. Tests were also conducted with dynamic prediction and time warping both on and off to see how they affected. Captured signals were processed to find out the phase difference between true movement and change on the screen, from which M2P latency could be calculated. [22]

The results showed that a bigger pendulum angle produced larger latencies, so fast movement in VR may result in stronger disorientation and cybersickness. Seemingly the rotation latency is more affected with acceleration speed than translation. Also, turning on the dynamic prediction and time warping did not reduce the latencies, which the researchers thought might be, because they cleared each frame, and no vertex or pixel shader was involved. Overall the Oculus Rift DK2 had very low latencies, maximum being from 90-degree drop angle: 7.2 milliseconds for translation and 10.3 ms for rotation. For 60 degree angle only 4.0 ms for translation and 1.3 ms for rotation latency. [22]

Similar to the two studies discussed above, this study also found changes in latencies as a result of changes in the context. This further highlights that even measurements on integral physical variables of the VR experience are not necessarily straight forward to measure. This in turn highlights the need to carefully assess the measurement protocols used when these variables and their changes are studied.

2.7.3. Quality of Experience

Quality of experience is one of the main variables measured in our thesis when talking about subjective experience. Since the below study by Ashutosh Singla et al. provided a starting point in the design of our suggested study and our expert evaluation we want to emphasize the study here.

Measuring and Comparing QoE and Simulator Sickness of Omnidirectional Videos in Different Head Mounted Displays

In the study, Singla et al. measured three aspects of 360-degree video experiences at two resolutions (4K and Full HD) on two HMDs: integral quality, simulator sickness, and behavioral analysis. The integral quality was defined as the HMD's ability to present audiovisual information, and behavioral analysis was based on the amount of head movement when subjects were watching the 360-degree videos.

The six videos used in the study were categorized into three sets based on their motion: slow, medium, and fast, with two videos in each speed category. Roughly one minute long pieces of the selected videos were used. Similar categorization can be derived from multiple sources, as it seems to be highly related to the amount of simulator sickness test subjects might experience [23]. Only one minute long pauses were used between the video clips to minimize the effects of a video on the following one. Each individual had their playlist to counterbalance any learning effects. In total each content was rated 4 times (2 different HMDs and resolutions), and each subject rated 24 videos in total.

The main method of getting information about the integral quality was asking the test subjects to rate their experience from 1 to 5, with 5 being excellent and 1 bad. The relevant conclusions of the study are as follows: One of the headsets offers slightly better integral quality, and there is a significant contribution of content and resolution to the final quality ratings, suggesting a connection between content and device [20]. The study also states that greater simulator sickness scores are caused by certain types of content and that females are more prone to sickness. There was also a strong connection between the subjects turning their head, content type, and amount of simulator sickness. [20]

The key point that we transferred to our suggested study design as well as the expert evaluation was varying the content used. We implemented this by including videos varying in intensiveness and by including other contents that also varied in intensiveness within themselves and between each other.

2.7.4. Immersion

The general immersiveness and presence of a VR experience crystallize many other factors into a single concept. A study, presented below, by Papachristos et al., focuses on the immersiveness of VR experience. We present this study to provide an example approach to measuring immersion. When discussing immersion and presence we use the definitions by Slater. Slater states that immersion is an objective characteristic of a VE system; this immersion is defined by the systems objective performance capability which provides the ground on which the subjective presence of being in the VE is built on [24].

A comparison between Oculus Rift and a low-cost smartphone VR Headset

The study compared two systems, one HMD based on a smartphone and a plastic casing, and one driven by a PC. The main research question focused on changes in spatial presence, usability, simulator sickness, satisfaction, workload, and learning outcome when test subjects were using the HMDs. The test subjects used the two HMDs to learn the order of planets in the solar system using an application called ’Titans of Space’. The actual testing consisted of measuring the spatial presence, usability, workload, simulator sickness, and satisfaction. These were measured with Temple Presence Inventory, System Usability Scale, Task Load Index, Simulator Sickness Questionnaire and a simple question of ’how pleasant was the virtual experience?’ on a scale from 1 to 5. [25]

The study concludes that there was little difference in learning outcomes when using the different HMDs and the immersion level was also similar. The workload measured might indicate that smartphone-based HMDs have a greater strain on the visual system. Learning outcomes showed a small difference in favor of the PC based HMD, but this result was inconclusive partly due to the sample size. [25].

Similarly to other studies we have highlighted, this study also approaches measuring the experience via various questionnaires presented to the users. It should be noted that other approaches, however, are possible. As a counterexample, these measurements could be based on measuring physical variables from the test subjects. While these are also dependent on the subjective experience of the individual they are more subconscious. Examples of these variables are heart rate, skin conductivity, movement patterns, and brain activity. This sort of biometrics-based approach is not further elaborated on our study as it was deemed to be outside the scope of our thesis. We have decided to leave this note as an acknowledgment of the existence of this approach.

Assessing User Experience in Virtual Reality – A Comparison of Different Measurements

In the above study, Wienrich et al. perform related user experience and immersion measurements. The study had a setting where subjects played in multi-user adventure on the Immersive Deck of Berlin-based Illusion Walk. The adventure has the subjects working as a team in a story that plays out in VR. The environment has mixed reality elements such as doors and buttons that exist in both the virtual and real realities. The users have true freedom of movement provided by VR systems built into backpacks. [26]

In the study, several positive correlations were found. In our context, most important of these were a correlation between VR presence and post-experience user experience as well as a correlation between VR social presence and post-experience user experience [26]. These findings hint towards different aspects of user experience affecting other parts of user experience indirectly. In the expert evaluation we performed, these sorts of connections were not specifically assessed but we suspect some were present. Due to the format of our study, this does not directly affect our findings that are presented at a later time. We, however, think that acknowledging the possible existence of these sorts of connections that are not immediately clear should be taken into account when designing a study.

2.8. Content

The content that is viewed in VR is a key part of the experience and special attention should be paid on both creating and selecting the content as different contents might introduce very different amounts of discomfort. Within this selection process, attention must be paid platform support as all content is not available on all platforms.

In contents that require moving the user's virtual avatar within the virtual world the choice of movement mechanic affects the immersiveness of the experience. When available, the most natural approach to moving around is walking. Conversely, this is commonly only a partial solution since VR system installations are often limited by the space they are set up within. Different workarounds to this have been presented, such as teleporting within VR, omnidirectional treadmills, and redirected walking. In redirected walking user is coaxed to walk in a large circle which is interpreted as walking in a straight line within the VR. The goal is to keep the user unaware of them walking physically in a circle. This approach, however, also requires a considerable amount of physical space. [6]

Movement within the VR, in general, has to be designed so that the conflicts between physical and virtual movements are minimized. This is essential in avoiding cybersickness. These possible conflicts are not limited only to movement; acceleration can produce similar conflicts [23]. Generally, it is beneficial to avoid creating a high disparity between optical flow and user's movement; teleporting within VR avoids this to a degree as the act of teleporting is instantaneous which consequently limits the disparity only to a brief instant of time.

This is in contrast to moving around with a joystick. As an example, the user might either walk, fly or glide in VR while being physically still. This sort of movement creates a large disparity over long periods, which should be avoided. Workarounds to these joystick movement problems do exist. For example, some contents artificially limit the user's FOV by blocking their entire peripheral vision when they are moving in 3D space if they remain physically still [23, 6]. This limits the stimulus for the user's peripheral vision and thus helps to reduce vertigo.

The movement of the user's FOV should always be controlled by the users to avoid cybersickness [23]. An example of a situation where this principle is violated could be loss of tracking which might results in the image on the display remaining static despite users turning their head. In our experience, this sort of problem will result in immediate disorientation. The perceived discomfort of disorienting experiences is directly related to their duration [23]. This in turn creates a need to have the VR be rendered with minimal delay and without dropping frames.

2.9. Industry and Consumer Standards

In our literature review, we have not found any largely accepted standardized ways of measuring VR experience. Two noteworthy existing approaches are consumer comparisons and work-in-progress standards by the International Telecommunication Union (ITU). From the consumers' standpoint the de-facto comparison method for different VR products is to use metrics like platform, screen resolution, refresh rate, and FOV. As for the ITU standards, they are work-in-progress at the time of writing this and we will not comment on them besides acknowledging their existence [3].

A large number of consumer comparisons can be found with search engines. The metrics used in these sorts of comparisons give information about picture quality, graphical performance, and other technical aspects but they will not necessarily give information about variables like immersion, risk of cybersickness, or quality of experience in the user's preferred content.

The choice of content is a noteworthy factor when measuring the quality of experience. This results in further problems when a standard is being decided on. As an example, the perfect experience for watching a 3D movie might constitute of very different elements than the perfect experience of playing an immersive VR role-playing game. These differences raise from their possibly very different requirements; a video experience might be perceived to be of high quality when it has very high resolution and the display used to view it has minimal distortions. However, the role-playing game might be perceived to be of high quality when the player has complete freedom of movement and is not limited by tethers or the space they are playing the game in.

[3] https://www.itu.int/ITU-T/workprog/wp_search.aspx?q=13/12

3. SUGGESTIONS FOR FUTURE STUDY

In this section, we present suggestions for performing a future study to combine qualitative and quantitative metrics of a VR experience. We first present the study's problem statement followed by test methodology. We also present two small pilot studies that we performed based on the suggestions.

3.1. Suggested Problem Statement and Research Question

We present the following suggestion for future study problem statement: *Any comprehensive comparison of VR systems requires correlating the physical attributes of the system with the subjective experience provided by said system, which has not been systematically performed before (at the time of writing).* This leads to our suggested research question; *What kind of correlation can be found between the physical attributes of a VR system and the subjective experiences the system provides?* The following sections will describe our user study pilot test and a plan for the larger test. The methods for measuring the physical attributes of VR systems are also discussed.

3.2. Defining the Suggested Test Methodology

Based on our suggested problem statement and research question we propose that the most reasonable approach would be to combine both existing and original quantitative and qualitative studies. The purpose of the data aggregation that would be performed is not to test or define recommendations for any singular variable but to understand the bigger picture.

In the qualitative part, the goal is to find out if it could be determined which HMD offers the best experience for the user and why. The overall design of the study is loosely based on the research presented by Ashutosh Singla et al.[20]. We propose three different types of content selected by varying the speed and amount of interaction required from the participants. After the three experiences, the participants would be presented with four short questionnaires to measure simulator sickness, presence, and immersion for statistical analysis. We suggest carrying out the same test 20 times for each of the HMDs selected for the study. This would result in a total of 80 answer sets giving the study a reliable *N* and statistical significance. We discuss a set of example VR systems that could be used in section 3.3.

For the quantitative part of the study, we present meaningful metrics for consumer VR systems in section 2. Some of these metrics can be directly found from manufacturers' specifications while some can be measured only with specialized equipment. Measurement devices are briefly discussed in section 3.5.1. Special attention should be directed towards the M2P latencies which are one of the main contributors to cybersickness. The end goal here should be to gather a comprehensive table comparing the attributes of HMDs in a way that is not normally accessible when compared to consumer reviews. Examples of these sorts of variables are M2P latency, various jitters, audio-video sync, display brightness, and sensor drifting.

3.3. Defining the Suggested VR Systems to Be Used

A basic criterion for the VR system selection process that we argue should be decided on first is choice between tethered and non-tethered headsets. We suggest a comparison that contains both wired and wireless HMDs. Non-standalone HMDs should be powered by a high-end workstation PC to avoid any bottlenecks outside of the VR systems themselves. Furthermore, we suggest a choice to equalize the number of tethered HMDs with the number of standalone ones so that meaningful correlations can be derived from the effects of a tether on user experience. We argue that a reasonable scope in the sense of HMDs for this study would be four or more HMDs. This should provide enough data to find statistically relevant correlations. The chosen VR systems should also be easily available for consumers to make the study easier to replicate. At the time of writing, we suggest using HTC Vive Pro, HTC Vive Pro with a wireless adapter, Oculus Quest, and Oculus Rift S for this sort of study.

If the suggested study is performed in future this recommendation is very likely to be partially outdated and should be refined on. As a last note regarding these suggestions, we advise against using a mobile phone-based HMD as they seem to be a dying trend after Google discontinued its Daydream platform [27]. A different argument against mobile phone-based systems can be made based on their varying computational capabilities which make the choice less generalizable. A factor worth considering is also the ease of software development for these systems as using VR multimeters is likely to require using content specifically designed for performing the measurements.

3.4. Qualitative Test

In this section, we discuss how the user study could be performed in practice. We constructed our example test like earlier relevant study [20]. Some important changes were, however, made compared to this study. We argue that carryover effects from using multiple different HMDs within such a short time period might have a large impact on questionnaire results when measuring the subjective experiences of non-experience users. Counterbalancing can be performed to avoid this to some extent but we further argue that counterbalancing will not prevent diminishing the possible subtle variations in the results that are being studied. To avoid carryover effects especially when measuring simulator sickness we decided to allow for one person to test only one VR system. This is in contrast to an earlier study where each test subject used many systems within a short time.

3.4.1. Defining the Test Content

We suggest for the test to contain a variety of short tasks so that the test subjects can enjoy a variety of different experiences in VR within a manageable time frame. We deem the variety to be an important aspect of the test protocol as the test is intended to be comprehensive. The test content should be something simple and relatively easy to handle so that the participants will not get overwhelmed by the difficulty of the task. Furthermore, the tasks should be presented to test subjects in order of increasing intensity so that the risk of experiment getting interrupted due to nausea remains minimal.

Content used for the tests also has to be highly available as VR systems are designed for very different platforms. This is why we suggest using a stereo 360-degree video and two games that can be acquired for all of the different systems. We present the 360-degree video titled *Reimagine Etihad A380* [28, 29] as an example for this purpose. This video was chosen for multiple reasons; it was used in an earlier study [20], it is short, it does not contain any sudden movements, and it is available in high quality. We suggest using a selected short section of footage from the video chosen. Examples timestamps for this particular video are as follows; starting from 2:30 mark and ending at 4:30 mark.

For the next content, we suggest the popular VR game titled *Beat Saber* [30]. In this rhythm game the player slices cubes that slide towards them on a platform with two lightsabers. The player also dodges walls that slide towards them by physically moving in the space they are playing in. Within the game, we suggest using the most widely available songs such as *Beat Saber*. Extra care should be directed towards selecting reasonably easy settings such as and no-fail -mode and relatively low difficulty.

For the last content we further suggest a game titled *Richie's Plank Experience* [31]. In this game, the player can experience a variety of VR demos but we focus on a very specific part of the game. The player starts on a street next to a skyscraper and on the side of the skyscraper they enter an elevator that takes them to the 80th floor of the building. The elevator opens directly from the side of the skyscraper and in front of the player is a plank from which they are supposed to walk off and experience free fall in VR. The experience is enhanced by the fact that the players can be placed in front of an actual physical plank that has the exact placing and dimensions of the virtual plank sans the height. This experience in particular is good at measuring the amount of immersion via the players' possible hesitation to jump [32]. However, it suffers from the caveat of very fast habituation so the experience is not repeatable per person when results are being measured.

3.4.2. Defining the Test Subjects

Our suggested method for selecting and gathering participants is meeting people in ad-hoc style. A suitable place for this could be a university campus, a library, or a mall. Participants should also be able to book a time during a specified day so that they could arrive at a specific time suitable for them; this approach should be combined with gathering people spontaneously to fill any possible gaps. To avoid carryover effects participants should be excluded if they have used VR systems within last week or if they already been tested. This is to reduce any possible bias. Basic information should also be collected on the test subjects; information such as gender, age, and previous VR experience are important factors when statistical analysis is being performed to evaluate which system provides the best user experience. Ultimately this information is then joined with information gathered from quantitative analysis to understand why physical attributes led to these user experiences. When measurements are being performed, the basic variables like gender should also be balanced for if possible, so that each VR system receives equal distributions in relation to these variables.

3.4.3. Defining the Data Collection

In the appendix, we present a set of example questionnaires for collecting data from the test subjects' experiences. The example questionnaires are as follows: simulator sickness questionnaire (SSQ) [33], two mean opinion score (MOS) questionnaires [34, 35], and Slater-Usoh-Steed presence questionnaire (SUS) [36]. In addition to these questionnaires, the events should also be recorded so that they can be verified later if necessary. We also suggest taking notes manually, as the participants might suddenly exclaim important information. The participants should also be given business cards with contact information so that they can inform the testers in case of late onset of cybersickness. Next, we briefly describe the questionnaires and their purposes.

The SSQ is used to measure total simulator sickness as the combination of three sub-scales that are called nausea, oculomotor, and disorientation. The questionnaire includes 16 variables, such as headache. These variables are rated on a four-step scale of none, slight, moderate, and severe. The results are then weighted to obtain total scores in the sub-scales which are further combined into the final total score.

The two MOS questionnaires include a total of 11 questions that are rated on a scale from one to five based on their verbal equivalents presented to the test subjects. The equivalents are bad, poor, fair, good, and excellent. The two question sets contain questions such as *'How would you rate the responsiveness of the system?'* and *'How would you rate image contrast?'*. These questions were acquired from an earlier study [34] and from ITU recommendation titled: ITU-T Rec. P.910 (04/2008) Subjective video quality assessment methods for multimedia applications [35].

Finally, the SUS questionnaire measures presence based on questions, such as *'To what extent were there times during the experience when the virtual environment was the reality for you?'* that are rated on a scale from one to seven when seven represents the normal experience of being in a place.

Some studies have also collected biometric data for further analysis [37]. This information can be especially useful when cybersickness or other symptoms of discomfort are measured clinically in addition to questionnaires. We do not provide any suggestions regarding this option as we have gathered no information on the methods specific to this approach. This is due to our initial decision of leaving biometric data collection outside of the scope of our study.

3.4.4. First Pilot Study

To test our plan for the qualitative study, we decided to run at least two pilot studies. The first pilot involved only us and the second one used test subjects. These pilots were intended to provide us information on how to improve our study design, as the initial plan was to perform the study. During the first pilot, the final set VR systems to be measured was not yet decided on. We instead opted to use a diverse set of three systems with different qualities. One was tethered Oculus Rift S, which uses inside-

Figure 4. One of the researchers testing the plank before pilot study.

out tracking based on IR cameras in the headset. The second HMD was also a tethered one called HTC Vive Pro that utilized external IR laser beacons for the tracking. The third headset was Oculus Quest which was selected to give us hands-on experience on testing standalone headsets. We selected a decent-sized testing site at the University to set up these tests. The room was dimly lit with natural light from one side. This gave some light for the inside out tracking cameras to work. For running the tethered HMDs we set up a powerful desktop PC, with Intel i7-9700K central processing unit, Nvidia RTX 2080 graphics processing unit, 16 gigabytes of DDR4 random-access memory, 500 gigabytes M.2 solid-state drive, running Windows 10 operating system. We also brought a large display into the room to easily follow the participants' progress when they were using tethered HMDs. The display was also used to show the participants a slide show which helped us present the experiment the same way for everyone. An overview of the room setup can be seen in figure 4.

We first set up all the software and hardware required for the systems to work and then calibrated the VE play area to fit the room. In Richie's Plank Experience the physical and virtual planks were calibrated to precisely match each other. The placement of the physical plank was marked to the floor with masking tape. To practice performing the study we first ran the whole test by having one of us act the part of a test subject. We then found the first three actual test participants for the pilot study next to the test room. The participants were our acquaintances that had happened to be working next to the CAVE. The first pilot test was then performed by having each one of the acquaintances to test one of the systems.

To ensure repeatability in the test procedure, we created a detailed script to be followed when working with the participants. Since each participant would be using

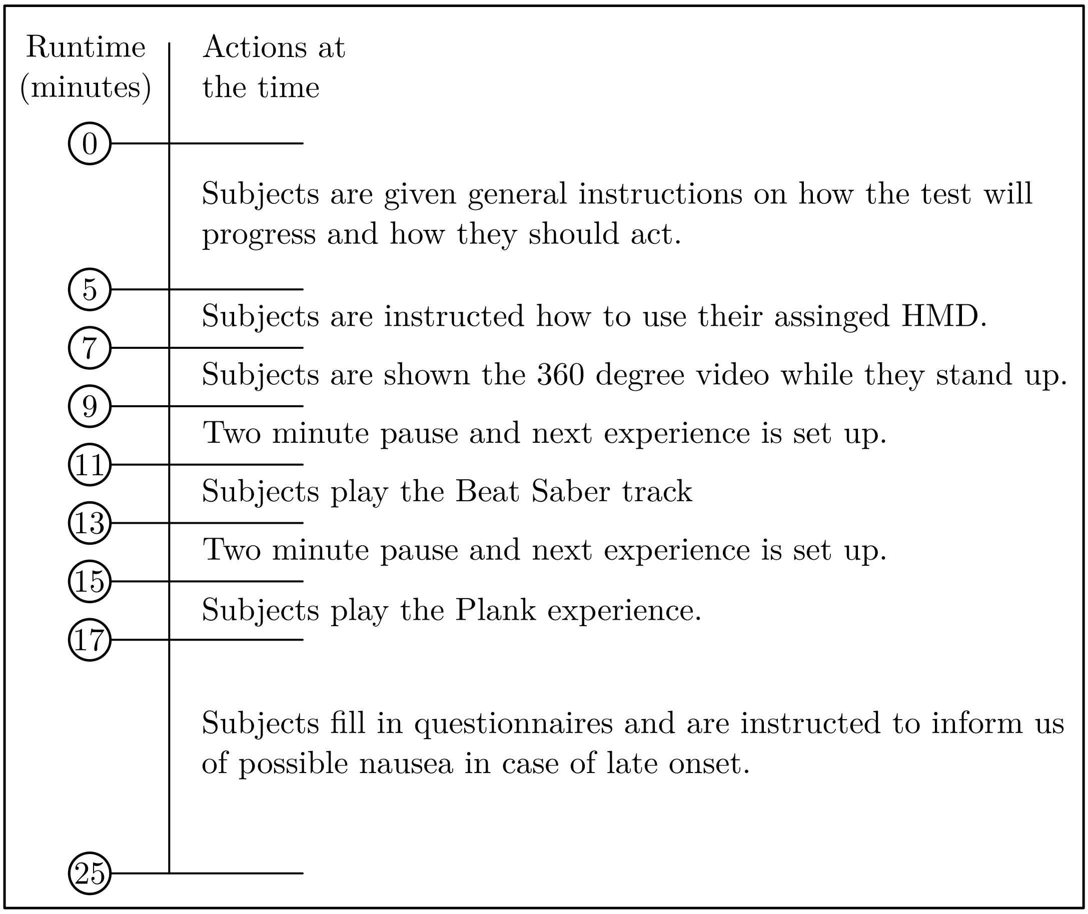


Figure 5. Timeline of an example test subject performing the test.

only one of the four selected HMDs there was no need to worry about counterbalancing the order of the systems besides making sure that every system would be tested equal number times with both genders. We made a conscious decision to test the contents in the same order for each participant, starting from the least intense. The goal was that one participant would complete the test in around 25 minutes. In the actual test, we planned to record all tests on video and for one of us to take manual notes during tests.

After the participants of the second pilot arrived we welcomed them in and asked to sit down for the general information slide show. This slide show was used throughout the test to present the tasks. The participants then signed a consent form and filled short demographics questionnaire regarding age, gender, experience with VR, and other details. The question set can be found in appendix 2.

We then introduced the first task, which was watching a short 360-degree video while standing up. Next, we helped the participants to put on the HMD and explained the necessary controller inputs. When the participants were ready they would start viewing the 360-video by themselves by pressing play-button with the controller. When the video ended, the participants removed the HMD and sat down for a short break, while we presented the next task.

We showed a short video [38] (from 1:00 to 2:15) to visualize the Beat Saber gameplay to the participants, so it would be clear how the game works. During this time, one of us prepared the Beat Saber, so that the participant could start the level easily. The participants did not complete any tutorial level before starting the actual test. After the level ended, we also wrote down the Beat Saber score for possible use in the evaluation.

After Beat Saber, there was also a short break while we introduced the last task, and again the participants sat down with the HMD off. The final task, Richie's Plank Experience, was explained while one of us again prepared the game for the participants. The participants could then start the experience by simply putting on the HMD, which after they would get in the elevator, and push a button to get to the floor in the building where the plank was.

To heighten the experience on the virtual plank, we created a physical plank, which rested 5 centimeters above the ground on top of styrofoam pieces that allowed the plank to slightly bend from side to side while walking on it. We adjusted the settings in Richie's Plank Experience so that the virtual plank in the game matched the size of the physical plank we created. We also adjusted the position of the plank so that virtual and physical location matched, and marked the correct position with tape to the floor of the room. Measurements and images of the plank can be found in appendix 1.

While the participants were in the virtual elevator, we positioned this real-world plank in the marked spot. When the elevator was up and the doors opened, we helped the participants onto the plank so that they would not trip. After that, one of us positioned to stand at the beginning of the plank to keep it still, while the other two stayed alongside the participants to help if they lost their balance. The test participants were asked to jump off the end of the plank and the test ended either when participants would fall to the ground or decided that they did not want to jump. We recorded timestamps for this content to later review time spent hesitating.

When participants were ready after the test, they answered the questionnaire regarding the experience with a laptop via Google Forms. This question set comprised of 4 different questionnaires described already in subsection 3.4.3. After the questionnaire, the test was over. We told the participants that we would like them to contact us with email or Whatsapp if they experienced any nausea later that day after the tests, but it was not a required part of the test.

3.4.5. Finalized Test Procedure

While performing the pilot tests, one of the first issues we encountered was adjusting the audio volume to be appropriate for all three contents. Especially the 360-degree video volume was criticized for being too low so we have to find a way to control it properly. The next problem we encountered was an environmental one. In our tests, the inside out tracking in the Oculus Quest was lost a few times while trying the contents. This was most likely because the room was quite dimly lit and the tracking relies on cameras to recognize features to lock on to. Another problem with the standalone system was that the play area and by extension the positions of the physical and virtual planks were not properly saved when exiting the game. This meant extra waiting time for the participant while we performed recalibration. A Possible workaround for this could for us to run the test with two HMDs and not exit the Plank Experience between subjects.

As mentioned in content subsection 3.4.1 the graphics of the Plank Experience are worse in the standalone system. We, however, argue that this can be seen as a partial limit of the standalone ecosystem in general. We also noticed that users could accidentally drop off from the plank very quickly if they moved their head too far from the plank when reaching out to look at the view. We assert that users should not be warned about this as it might break the immersion of the experience if they have to focus on our instructions. In the 360-degree video, we noticed some problems with the projection in tethered HMDs which made straight lines appear as being curved within the view of the HMD. We used the same Viveport video player from Steam for both the Oculus and the HTC. Furthermore, we accidentally played the 360-degree video in the tethered glasses using a 3D source file while the standalone HMD's video was in 2D. This did not cause any problems as our approach changed and we initially did not plan on using the results from the pilots but it should be noted that careful attention should be paid to these sorts of details.

The questionnaire part of the pilot test revealed many details that can be improved on. An example of this is the need for small improvements in our translations of the question sets. We also noted a need for visualizing certain uncommon terms when speaking about images, for example, *smearing*. More importantly, we need to emphasize to the participants that they should evaluate the HMDs, not the content that was shown. This was mainly a problem in the ITU question set measuring the image quality as the participants asked what they need to assess. For our next testing of the questionnaire, we are planning to explain that the subjects should view the experiences as tools instead of the goal. We also need to stress that in the SSQ they should consider only the symptoms that were not supposed to be felt. By this, we mean that if a participant experiences vertigo while standing on a virtual plank 80 stories high the feeling should be considered positive while vertigo during the 360-degree video should not.

We also learned that preparations are needed for a situation where participants might feel sick by having some drinking water and bucket to puke into, should they be needed. Additionally, we learned that contact information cards should be ready at hand so that the participants can inform the personnel performing the test easily if they start to feel nauseous hours after the test.

Finally, we learned how a question in the questionnaire about the amount of experience the test subjects have using VR systems, should be implemented. This realization came to us after finding out about the relatively high amount of VR experience the test subjects had beforehand. We recommend implementing this by asking the subjects to evaluate how many hours they have spent using VR systems on a logarithmic scale of 0-1 hours, 1-10 hours, 10-100 hours, 100-1000 hours, 1000-10000 hours and 10000 or more hours.

Later in our testing, we also realized the importance of the correct IPD setting for the experience with the HMD. Therefore we strongly recommend, that the IPD should be measured from all participants and set correctly from the hardware or software side before the test is started or else the results might be erroneous. However, if this is too unpractical for a large study, the average IPD values for males and females should be used instead.

3.5. Quantitative Tests

Quantitative measurements were the second major part of our original research question. Our measurement method planning was interrupted in early stages, and therefore, instead of giving a detailed description of how the measurements could be done we briefly introduce the measurement tools and the variables that could be focused on.

3.5.1. Defining the Measuring Equipment

The topic for this thesis was introduced to us with the idea of using the university's newly acquired VR measurement tools. Therefore it was clear from the beginning, that we would use those professional-level tools instead of our methods. The tools in question were made by a Finnish company called Optofidelity, and they are named as BUDDY 1 and BUDDY 3 based on the DOF they can measure. These VR multimeters can measure multiple HMD-specific variables, such as M2P latencies, tracking accuracy, drifting, and jitter.

3.5.2. Defining the Variables to Be Measured

Based on our initial plans for the quantitative measurements we present a list of variables that could be measured or acquired relatively easily with either a VR multimeter, from manufacturer's specifications, or with other measurement tools related to displays. Due to the changes in scope discussed at section 1.3 this list is not intended to be comprehensive; we have decided to include it as a base for more accurate future work: *platform support, resolution, PPD, FOV, display refresh rate, weight, display type, included sensors, amount of DOFs, IPD adjustment, wirelessness, tracking type, display brightness, backlight frequency, backlight period, jitter related to movement, M2P latency, movement prediction over and undershoot, audiovisual synchronization and color accuracy*.

4. EXPERT EVALUATION

In this section, we introduce and discuss the expert evaluation method we adapted to suit VR systems. First, we discuss the adaptation itself, followed by information on the VR systems and on the content. Lastly, we present the testing method in detail.

4.1. Expert Evaluation Problem Statement and Research Question

For our expert evaluation, we present the following problem statement: *Designing and using VR systems requires balancing different aspects that contribute to the user experience. This balancing requires understanding connections between physical attributes and user experience.* This leads to our research question; *In the contexts of general and specific use cases, what kind of subjective connections can expert evaluation find between user experience and physical attributes of a VR system?*

4.2. Adaptation of the Method

Based on our literature review, expert evaluations are commonly used to assess the functionality of user interfaces. Conversely, expert evaluations are rarely used to compare different devices and experiences. At first, we attempted to develop our method to perform the expert evaluation for the HMDs. The method was loosely meant to subjectively compare different quantitative and qualitative metrics between the HMDs and rate the HMDs from best to worst in each metric. We would have then rated the HMDs overall quality of experience using similar contents as introduced in our suggested human study also from best to worst. This data would have been then used to see whether any clear correlations between the individual ratings and rating for the whole experience could be found.

There was some uncertainty about whether our method would produce enough useful data. This resulted in us next looking more closely into some well-established methods for evaluation of VR user interfaces. One of the methods was utilizing extended cognitive walkthrough and a model of action [39]. While the method seemed to be effective for evaluating the usability of VE it was suboptimal for hardware analysis.

Following the previous, we analyzed the paper *Heuristic evaluation of virtual reality applications* by Sutcliffe et al. [40] The paper presents a method for evaluating VE's user interfaces using a method that is based on the well established Jakob Nielsen's usability heuristics. The method consists of using the predefined usability principles and heuristics, to assess the nature of the problems encountered while going through predefined test cases with a system. The study was selected as a base of our version that we adapted for VR system expert evaluation. This is due to the testing method being easily performed by one expert at a time and the ease of replication of the tests.

The most significant changes were made to the definitions of the heuristics and the severity classification. Careful thought was put into adapting the principles concerning usability in VEs to principles defining the experience with VR systems. Single heuristics that could not be directly adapted were substituted with an additional heuristic that was more characteristic of VR systems. Additionally, a new class for positive findings was added to the severity classes to help gather important data for our research question. A list of our heuristics can be found below in subsection 4.7.1 and the list for severities is found in subsection 4.7.2. A part of the original expert evaluation was omitted; the part concerned a technology audit and we were already aware of the capabilities and functions of our chosen systems.

An additional methodology was added to include a ranking of the tested VR-systems to help better answer our new research question. We added steps to rank the VR-systems from best to worst after each task was completed, and at the same time numeric grade for the VR system was given with the following scale: *5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), 1 (Bad)*. Overall ranking after all the tests were finished, was also added to the evaluation method. The findings would not be classified based on the heuristics. Instead, we would later work together to combine all the findings and categorize them in the same manner that we used in section 2 to introduce different factors in VR-systems.

We will next present the VR systems selected for the evaluation, detailed definition of the tasks, and finally the detailed testing procedure.

4.3. Defining the VR-System to Be Used

All HMDs we originally planned to use in our human study, were automatically chosen to be part of the expert evaluation. As we were not so limited by time in our expert evaluation, we decided to introduce two additional HMD for testing. Eventually, HP Reverb and HTC Vive Pro with wireless adapter were selected. HP Reverb was chosen mostly because its screen resolution was significantly higher than in the other HMDs. This gave us a wider range of variables we could try to evaluate. HP Reverb was also the only HMD utilizing Windows Mixed reality platform. HTC Vive Pro with a wireless adapter package was chosen as a companion for its wired counterparts because we wanted to include a system that could be used to strictly consider the effects of wirelessness.

We considered including Valve Index as one VR system for our study since it represents one of the newest high-quality consumer HMDs. Unfortunately, the units were sold out at the time of selection so this particular system was excluded due to external factors. We also considered a high-end VR system from Varjo. This system was decided against as it was not aimed for consumer use, unlike our other chosen systems.

4.4. Defining the Test Tasks

We based the choice of test contents on our previous suggestions for the user study. Because the number of test runs would be significantly lower than in our originally planned study, we opted to expand the tests to cover longer time spans to better match real use cases. The goal for this was to include a variety of different types of contents and use cases to diversely test the systems because different contents could benefit from different properties of the systems. As all the experts had previous experience of using different VR systems we decided to increase the complexity of the tasks from those previously planned for the novice users. Due to our experience, we also chose content that could cause nausea in less experienced users. Based on these decisions we finally defined four separate contents that included one or more accurately repeatable experiences.

4.4.1. Task 1: Videos

For the first task, we selected three different types of videos. The objective of the video content was especially to evaluate the quality of the display and lenses, smoothness and accuracy of tracking, eye strain, and user comfort. We asserted that these qualities would be easiest to assess in a slow-paced experience.

First video was a typical 2D high-quality 4K-video (3840px by 2160px) video with 16:9 aspect ratio. The video shows aerial scenery from Iceland with rich colors and good contrast, and it runs 8:11 minutes [41]. A typical video was selected to best evaluate the resolution and SDE of the display since the whole 4K image is seen from the user's viewpoint. The video is also not stretched like 360-degree videos would be. We could also expect to see a rectangular view while watching this video so any distortions could be seen easily.

For the next test video, we chose a 2D 360-degree video showcasing scenery from different seasons of Alaska [42]. The video's resolution is also 3840 x 2160 pixels and it lasts 3 minutes and 23 seconds. Being a 360-degree video, only a small part of the video file rendered to the user's viewpoint so the image has to be scaled up. One purpose of this video was to test the FOV of the HMDs and so the increased immersion compared to a normal video. Additionally, the purpose was also to see whether there would be clear differences between tethered and wireless headsets since 360-degree videos will prompt the user to look around and turn from side to side.

The third and last video was both a 3D stereo and 360-degree video where two friends go inside a video game while traveling in their car [43]. This source was a square 4K video (4096x4096 px) and length 4:21. The meaning of this video was on top of the other qualities to evaluate to the stereoscopic effect produced by the HMDs.

It is also important to note, that these 360-degree videos have many stitching errors and other distortions, so we made sure to disregard those imperfections while assessing the quality of the HMDs. All the videos were to be watched fully from beginning to end. Videos in a tethered HMDs were viewed via Steam VR and Viveport Video using local files ripped from the YouTube with the highest available quality. In Oculus Quest the videos were streamed with YouTube application in the highest possible quality.

4.4.2. Task 2: Beat Saber

For the second task we chose the game titled Beat Saber. The game is introduced earlier and shown in figure 6. As all of the experts had previous experience with the game we wanted to increase the complexity and intensity of the task by selecting three different songs to be played with increasing difficulty each round.

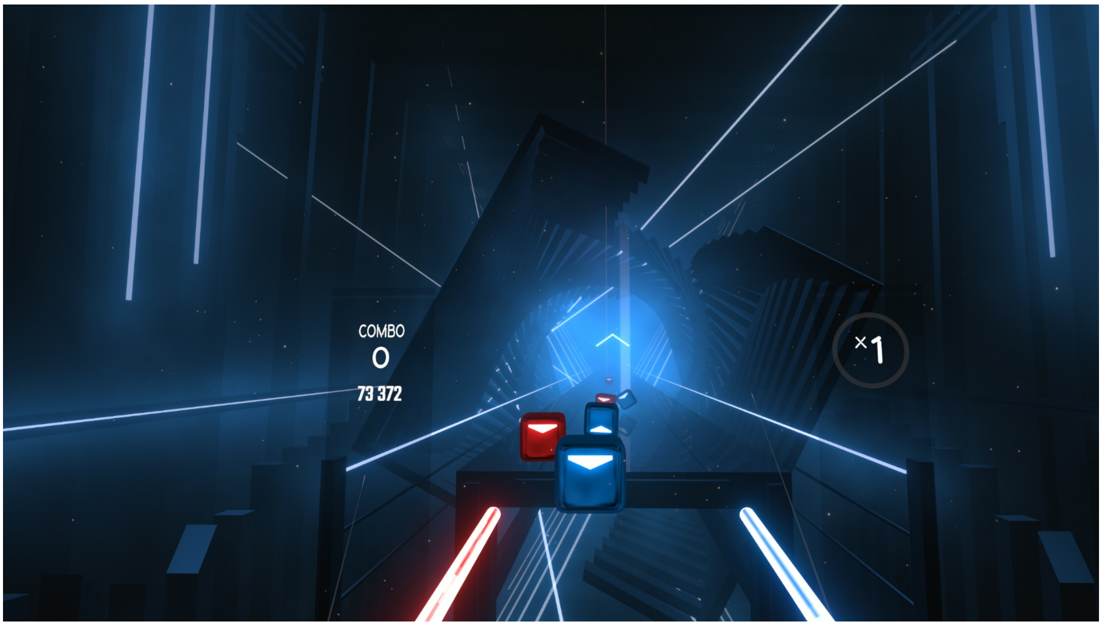


Figure 6. Screenshot from the gameplay of Beat Saber.

Beat Saber is a very fast-paced game, so our main goals were to evaluate the responsiveness of the image and tracking as well as agility while wearing the system, stability of the headset and controllers grip in quick movements. The detailed songs and settings used are listed below:

1. Song: Lvl Insane
   Settings: Normal difficulty, No fail selected, others options unchecked
2. Song: Escape ft. Summer Haze
   Settings: Hard difficulty, No fail selected, others options unchecked
3. Song: Beat Saber
   Settings: Expert difficulty, No fail selected, No arrows selected, others options unchecked

4.4.3. Task 3: Richie's Plank Experience

In the previous content, the user's position in the virtual and in the real world remains relatively stable. Richie's Plank Experience was selected to introduce 3D-movement in the VE, to see how strong feelings of movement the HMDs produce and to assess vertigo. One important aspect of the game was that in the Oculus Quest version the level of graphical detail is noticeably lower than in the PC version. This can be seen in the figure 8. We wanted to test whether that would be a major factor in the immersiveness of the experience and if the full wirelessness could replace the inferior details. It can be argued, that this content is not fair for Oculus Quest due to these differences, but in our opinion, the limited amount of computing power and therefore less rich content is a major attribute of the standalone HMDs and should be included as a factor.

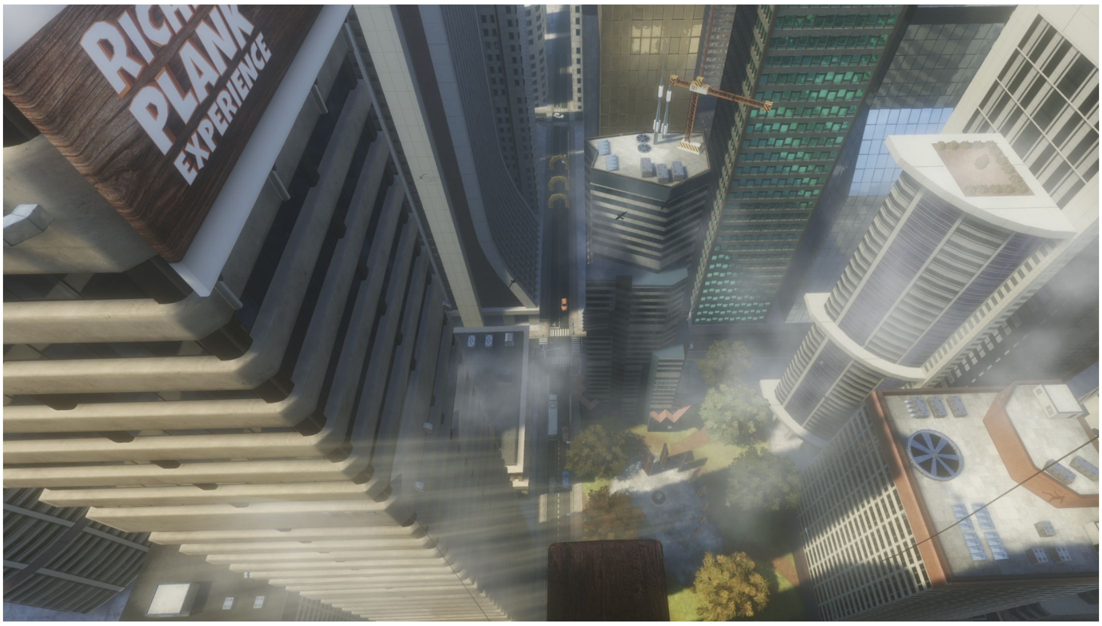


Figure 7. Screenshot from the Richie's Plank Experience while standing on the plank.

For the first mission in the game, we wanted to include the walking off the plank, like in the suggested user study and shown in figure 7. It was included because the user physically walks a short distance in the VE, and the height and the free fall can give a quite strong effect, especially when hitting the ground. It is important to mention again, that the jumping on the plank quickly starts to lose its effect after the first few attempts. This was not seen to be an issue in our case since we have performed the jump multitude of times with different glasses while designing the user study. Unfortunately, we had to exclude the use of the real-world plank in our expert evaluation since we were forced to do the experiments in our apartments and the space was not adequate for the large plank.

We selected three more missions from *Hero Academy*-section of the game. All of these are quite similar experiences as in all of them the player will be flying around modern city skyscrapers using hand-held jet-gloves. In the first *Rocket race*-mission, the task is to complete a racecourse by flying through hoops floating in various locations around the city. In the *Firefight*, the player is in hurry to search for fires and extinguish them by changing back and forth between the rocket gloves and a pair of fire extinguishers by momentarily placing their hand behind their back over the shoulder. In the last *Missile defence*-mission, the player protects the city by flying and smashing into the missiles that circle the city skies.

These flying missions were selected to introduce content where the user stands still in the real world, but in VE there is a lot of movements in all directions. This type of movement in the VEs is generally seen to cause nausea due to the mismatch of perceived movements between the eyes and the vestibular system. These flying missions were also thought to be good experiences for evaluating the immersiveness and veracity.

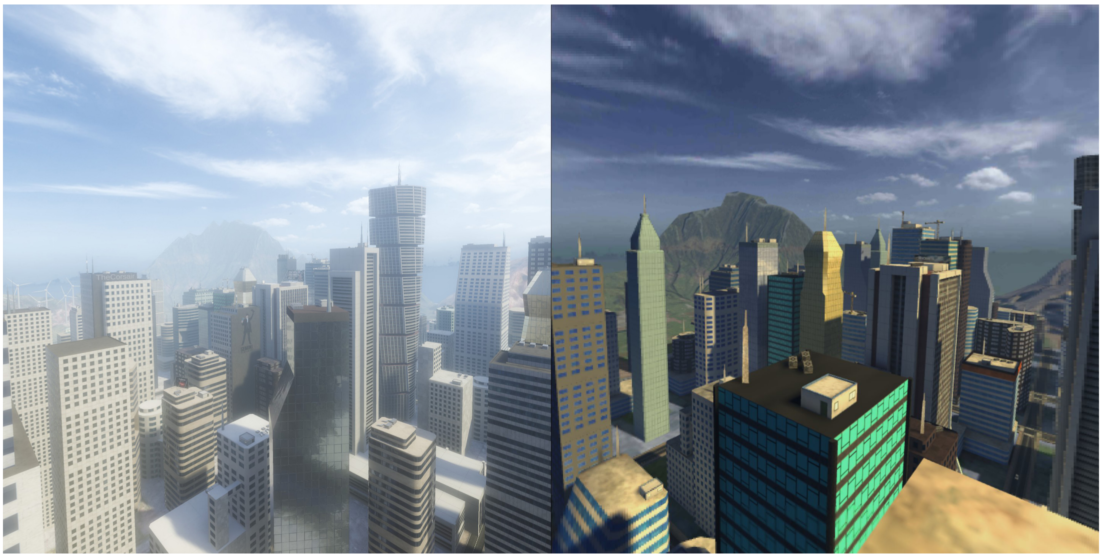


Figure 8. Comparison of the high-fidelity version *(on the left)* of Richie's Plank Experience on Steam and the low-fidelity version *(on the right)* in the Oculus Quest.

4.4.4. Task 4: SUPERHOT VR

Fourth content was added to further enrich our test design; we selected a game called SUPERHOT VR. It is a level-based first-person shooter puzzle game where time moves only when the player moves in the VE. In SUPERHOT VR the player must destroy enemies with various methods including shooting, punching, and throwing objects, while also avoiding enemy bullets and punches. The graphics in the SUPERHOT VR are cartoon-like and low-poly as can be seen in figure 9. SUPERHOT VR offered a very different control system compared to the other contents which would later help to broaden our evaluation.

One purpose of the game was to evaluate eye-hand coordination in VR when aiming, shooting, and throwing objects. Interaction with objects in the levels, shooting, and throwing objects were seen as a good way to test the also the functionality of the controllers. The game permitted us to freely move around while avoiding bullets and enemies even though our testing spaces were quite limited for performing extensive movement.

In the SUPERHOT VR, a so-called party-mode was used to run a replicable demo level in all of the HMDs. This mode was started unconventionally by selecting play in the Steam desktop application and then selecting the party mode. In Oculus Quest the same demo can be started within the game itself.

4.4.5. Excluded Tasks

Several other candidate contents were researched in addition to the four that were finally included. A content that we were the most interested in including was Valve's recently released Half-Life: Alyx, which is a first-person shooter game designed especially for VR. The game would have allowed us to test highly detailed and close to

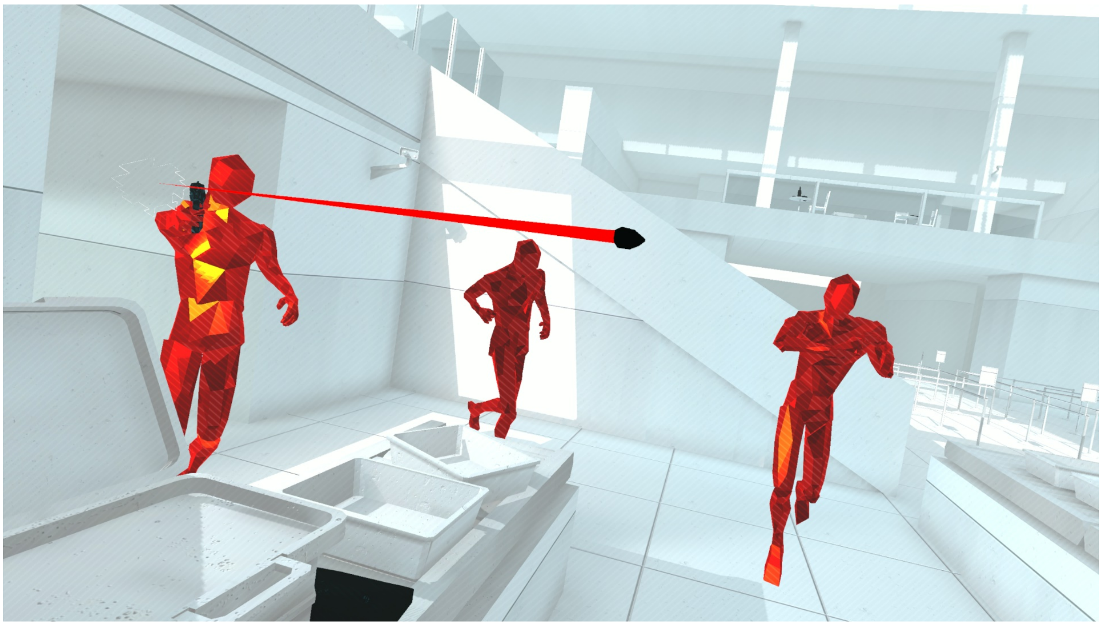


Figure 9. Screenshot of gameplay and avoiding a bullet in SUPERHOT VR.

the photo-realistic environment and navigation using teleportation. Additionally, the game would also have allowed testing of complex interactions within the VE using the controllers and accurate spatial audio.

We considered using a tether or streaming solution to test Alyx on Oculus Quest but this was decided against as it would have required either equipment we did not have access to or software solutions not approved by Oculus.

4.5. Testing Environment

Each one of us individually set up our testing environment. Thus a total of three different setups were created in our apartments of differing sizes. Manufacturers' instructions and recommendations were followed as best as we could when setting up the systems. Interference between the different systems was resolved by plugging only one system to the computer at any given time and by unplugging the HTC Vive Pro Base Stations when they were not used. Changes in environmental variables were kept as small as possible by keeping the windows and blinds closed and indoor lights on. We replicated the setup to the best of our abilities in all three spaces but inevitably some differences ensued such as differences in play area size.

4.6. Experts

We as the authors acted as experts. The resulting expert panel composed of three male computer science and engineering students with ages ranging from 24 to 25. All of our IPDs were measured to be around 61 millimeters. It is important to note that we had various amounts of experience with certain VR-systems that could have caused bias in our tests that were left unaware of. In section 6 where we discuss our findings, we have aggregated the findings to minimize any unknown biases.

One expert had around two months of experience with Oculus Rift S, a few weeks with HTC Vive Pro, and less than one hour with the others. Another expert had a few weeks of experience with Oculus Rift S and less than one hour with other VR-systems. The third expert had a couple of months of experience with Oculus Quest and less than one hour with the other systems. These time spans represent for how long each one of the experts had a particular system in their possession under occasional use.

Only one expert used eyeglasses during the testing process. The other two experts also had eyeglasses, but with such minor corrections that they preferred using the systems without the glasses. The expert that used the glasses was able to do so with all the systems.

All the experts had a similar amount of experience from the test contents: a couple of hours playtime in Beat Saber, around ten jumps from the plank, and less than one hour flying in Richie's Plank Experience.

4.7. Detailed Testing Method

The detailed steps to replicate our test are presented here. Before the test, we familiarized ourselves with all the test contents that were new to us so that the first test runs would not be negatively impacted by learning experiences. Before starting the test, the experts measured their own IPDs using a mirror and a ruler. The corresponding IPD was then set to the HMDs before using them. IPD for those HMDs that did not have a physical control were set from the software used.

Only one task was performed with all HMDs in one day to prevent fatigue and the accumulation of cybersickness effects. We randomized the starting order of the VR systems in tests for each expert. The order was then rotated per expert by shifting the last system in the first place in the testing order after each task.

- Each test content was then played using the following testing loop with one HMD at a time:

  1. Set up the appropriate HMD and turn on the computing unit on or restart it
  2. Check that system is working and play area is set correctly
  3. Launch the particular test content and perform the set task in its entirety
  4. Memorize any observations (negative and positive) and write them down after content is finished. Writing brief notes during testing is also allowed.
  5. After the test run, use the heuristics to walk through the experience again so that all aspects will be covered, and write down any new observations (negative and positive)
  6. Give negative observations a severity classification based on our definitions in section 4.7.2
  7. Grade the experience with the HMD using a scale of 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor), 1 (Bad)

- After completing one task with all the tested VR-system, the expert would then set all of them in order from best to last.
- At this point, any findings that came up as a result of experiencing all the HMDs could be still added to the findings.
- Expert would also ponder the reasons why a particular system performed better or worse than the others and write any notes down.

After finishing all the tests with all the VR-systems, the expert would then rate the overall experience with the system and rank them from best to worst.

4.7.1. Adapted Heuristics

In this section, we present the heuristics list that we adapted to be used in the research of VR systems. The order of the heuristics is intended the match the order of original heuristics as well as possible.

1. *Natural engagement and match to reality.* Being and acting in the virtual environment should feel as natural as possible and match the reality when it is intended.
2. *Compatibility with the task at hand.* System should be adequate to use in multiple types of VR experiences.
3. *User's depiction in the virtual world.* Users should be comfortable with the way their form is presented in the VE.
4. *Depiction of users actions.* Users should be able to naturally interact with the VE using the controllers and other features offered. The content should support this interaction where applicable.
5. *Feedback given of users actions.* The system should have means of giving the user clear feedback of actions performed and events in the VE.
6. *View to the virtual world.* User's view should be clear, properly scaled, and wide enough to give a good sense of presence in the VE. The view should be accurately mapped to the user's real viewing direction and updated with minimum delay.
7. *Responsiveness and accuracy of user's action and representation.* The position of the HMD, controllers, and other components of the system should be accurately determined and transformed to the VE. Position should be updated with minimum delay and noise.
8. *Optical adjustability and distortions.* Lenses needed to project the screen should be designed so that optical distortions, aberrations, and other negative effects would not distract the user from the experience. Different range of IPDs should be covered with the design or adjustability of the position of the lenses.
9. *Other distortions and defects.* There should not occur any unnatural glitches, deformations or displacements.

10. *Intuition of usage.* Using the VR-system should not require excessive learning and training from the user.
11. *Physical comfort.* System components should be comfortable to wear or hold for long periods of time.
12. *Sense of presence.* While using the system the user should feel like actually being in the VE.

4.7.2. Adapted Severity Classification

This section presents the classifications we used for our findings during expert evaluation. Positive findings were sufficiently described as only being positive instead of further subdividing.

1. *Severe*. The problem completely disrupts the user from the task or experience and must be dealt with immediately.
2. *Annoying*. The problem distracts the user from the task or experience and can not be ignored but does not require immediate addressing.
3. *Distracting*. The problem distracts the user from the task or experience but can be ignored for short periods of time.
4. *Inconvenient*. The problem could distract the user from the task or experience but can be ignored for long periods of time.

Positive. VR-system includes an additional positive feature or outperforms its rivals with different design choices

5. RESULTS AND EVALUATION

In this section, we list our findings from the expert evaluation. In addition to the scope defined before it should be noted that we intentionally did not include the system's set up as a part of our evaluation nor did we include the user experience of using system menus or tools. The terms inconvenient, distracting, annoying, and severe are used as defined in section 4.7.2. The categorization of the findings follows the order seen in section 2.

5.1. Displays

We made two important remarks on the visual displays offered by the tested systems. These were the choice of display technology and display resolution. By *displays* we specifically mean the LCD, OLED or AMOLED panels of the HMDs that form the pixel grid as opposed to, for example, lenses.

The most frequently reoccurring problem that was noted regarding displays was the visibility of SDE. This effect was noted on each one of the systems to a varying degree; it was least noticeable on HP Reverb and most noticeable on Oculus Quest with HTC Vive Pro and Oculus Rift S being somewhere in between of the first and second system. The rating for this problem was generally deemed to range from less than inconvenient to distracting depending on the content and VR system. Generally, high contrast, high brightness, and low movement were key elements when the visibility of SDE was noted. Video content was a clear leader in the number of findings related to SDE while other contents received occasional remarks.

The display technologies themselves resulted in the second major difference in the experience, as OLED and AMOLED based displays, such as Oculus Quest and HTC Vive Pro were able to provide much higher contrast and thus color definition. These differences were perceived to be noteworthy in our contents, albeit less important than high display resolution. These differences were most noticeable in high contrast scenes and scenes that contain black areas. The physical brightness of the displays was sufficient for all HMDs and contents except for Oculus Quest, as one expert noted that the display was noticeably dimmer than it's competitors, while videos were being watched. A different expert also noted severe flickering of the display in his peripheral vision, especially in videos.

5.2. Lenses

Similarly to displays, few remarks on lenses were repeatedly made on all the HMDs. These were CA, Fresnel banding effect, and god rays. Again similarly to displays, when referring to *lenses* we are only referring to the optics of the system and not the combination of optics and display technology.

CA was noted on all the system to an extent. Excluding HP Reverb CA was mostly noted on the edges of the image and the aberration was deemed to be usually only inconvenient and distracting at most. However, in the case of HP Reverb, two experts noted that the CA was very notable and present throughout the FOV of the HMD. In the context of videos, the two described CA as being severe on the Reverb, and in other contents, it was described as being up to annoying.

Fresnel banding was found on all systems and it was generally perceived as being more striking than CA but less annoying, being described only as an inconvenience in most cases. Different systems had different Fresnel banding styles depending on their construction. Banding was most notable on HTC Vive Pro, followed by HP Reverb and Oculus Quest. The banding can also be somewhat noticeable on Oculus Rift S but it received little to no remarks.

God rays were perceived as being similar severity on all the systems. God rays were mostly noticed in areas with high contrast in lightness and they always have an endpoint towards the center of the lens. They were commonly rated to be inconvenient and sometimes distracting.

Slight barrel distortion was also present in Reverb and Rift but this was barely noteworthy. The only significant distortion was found by two experts in the HP Reverb. Experts described it as a static warping or undulation of the image, and it was mostly noticed when turning the head relatively slowly.

Lastly, a circular variation of brightness of the image, with the brightest part being in the middle, was noted by one expert in both Rift S and Reverb. Caused by the same Fresnel lens banding, this respective mura was noted as being inconvenient and distracting. Other systems had no noteworthy mura effects.

5.3. Controllers

Unlike in notes regarding displays or lenses where experts tended to focus on different aspects, they were unanimous in their findings on the controllers. The different controllers offered various types of feels in practical use but the controllers present in Oculus Rift S and Oculus Quest were generally liked the most.

Furthermore, all experts unanimously agreed on HP Reverb's controllers offering the worst user experience. Two experts also noted the battery hatches of HP Reverb's controllers opening during intense gameplay when Beat Saber was being played. This was perceived as being a severe problem. The controller also provided the lowest perceived haptic feedback and the worst quality in the buttons.

The controllers of HTC Vive Pro received some positive notes and little to no negative notes. A notable exception to this trend was the control scheme present in SUPERHOT VR which required using the button built into the controller's touchpad when performing a throw in VR. Two experts described this as being an unnatural feeling design choice and rated it annoying. This problem can, however, only be partially attributed to the controller as it had an unused grip-based button left unbound in the game.

5.4. Tethered and Wireless Solutions

The effects of a tether and the effects of wirelessness are both presented here as they contrast each other. The findings on the effects of tether were focused mainly on its negative impact and vice versa for the wirelessness. Different HMDs had very different sets of tether related findings.

HP Reverb's tether was noted as being the heaviest and rigid which limited the user's movement by physically slowing them down which was perceived to be either inconvenient or distracting depending on the content. Compared to this Oculus Rift S' asymmetric tether placement induced a very different set of problems; it resulted in less limitation on the experts' movements but the tether also flung around onto the experts' hands during gameplay. This required repeatedly placing the tether behind the back which was found to be similarly inconvenient or distracting depending on the use case.

The experts noted HTC Vive Pro's tether to be the least obtrusive of the tethers. The tethered version, however, still produced generally lower immersion compared to wireless the counterpart. Vive Pro's wireless solution included using a wireless antenna add on which in turn required a battery pack that the user carries on their person. The experts noted the battery pack warming up but not to a harmful level. The battery pack was connected to the HMD via a cable, which induced problems similar to a tether but at a greatly lesser degree. These effects from the battery were considered to be less than inconvenient.

An indirect meta-finding on wirelessness can be derived from the results in table 2. Due to the testing including both wired and wireless versions of HTC Vive Pro, it can be seen that the wireless version was perceived to a better choice in all use cases, but only to an extent that increased it's ranking by specifically and exactly one position in each use case.

5.5. Audio

Our audio-related findings can be categorized to concern audio quality or user experience. Two of the systems offered hinge-based over-ear headphones while the other two offered audio that was not directly aimed at the user's ears. These different design approaches created very different user experiences.

The HTC Vive Pro and the HP Reverb both had hinge-based over-ear speaker solutions and they were generally deemed to offer better audio experience. They were also deemed to offer better immersion due to the over-ear speakers blocking outside sounds. However, this came at the expense of ease-of-use. Especially HTC Vive Pro was noted by one expert to be a somewhat clumsy feeling when the expert was equipping the HMD and adjusting the over-ear headphones. Similar problems were also present on HP Reverb but to a lesser extent. These adjusting related experiences were viewed as being inconvenient.

The other two systems, Oculus Quest and Oculus Rift S were noted to be easier to equip. This ease of use, however, had the caveat of offering notably more hollow and phone-speaker like feeling in the sound. These audio quality-related findings were noted to be only inconvenient by two experts, and annoying at most by one expert.

5.6. Tracking

We categorize tracking into two broad groups: inside-out tracking and outside-in tracking. The only outside-in tracked system was HTC Vive Pro, albeit there were two instances of it in the form of a tethered and wireless solution. Our findings regarding tracking were mixed as two experts reported generally more problems related to outside-in tracking and one expert noted more problems in inside-out tracking.

The non-mixed findings were largely dependent on the mechanics used by the contents. Some contents were found to be tremendously more susceptible to tracking related problems than other contents. This resulted in the severity of the ratings ranging from inconvenient to severe, with video content mostly receiving notes rated to be inconvenient and SUPERHOT VR receiving notes rated as being up to severe.

The most problems with tracking were caused by controllers and rated at most to be severe. In inside-out tracked systems this was true especially in movements where the controllers were behind the HMD. Less severe notes were related to the tracking of the HMDs themselves, but one expert noted occasional jitter and jumping in certain movements. One expert also noted small jitter in the Oculus Rift S's tracking system, and general problems with the inside-out systems when turned towards a big flat-screen TV and plain white wall.

Two experts had several problems with tracking of HTC Vive Pros, and less but still some problems with inside out systems. One of these experts noted that dimming of bright light reduced the number of problems encountered. The third expert instead had almost no issues with the HTC Vive Pros' tracking and a only few problems with the tracking of inside-out HMDs. He, however, had also several issues with the tracking of the controllers in inside-out systems, and especially in Oculus Quest's and HP Reverb's systems.

In general, from the inside-out tracked systems Oculus Rift S received least notes on tracking problems followed by Oculus Quest and HP Reverb in no specific order.

5.7. User Experience

This section contains a wide variety of different findings that relate to the overall user experience or otherwise do not fit other sections. The findings are loosely grouped based on their theme

The weights of the HMDs, controllers, and tethers were also perceived to be an important factor in how comfortable the system was. In addition to the weight, the weight distribution in itself was also an important factor that could significantly affect the general immersion. Regarding these aspects, the experts preferred systems that had balanced weight distribution and low weight alongside light tether when it was present. The balance and weights of different parts of the system were all perceived as being equally important. The experts noted several factors that affected general ease of use, such as some systems having physical sliders for IPD or physical buttons for adjusting volume. The lack of physical IPD adjustment was attributed to generally worse user experience due to eye fatigue. Two of the experts also noted that two systems also had rubber band -like fastening which made the HMDs easier to equip and unequip. The third expert, however, found these two systems to be the most uncomfortable to wear, as all weight of the HMD was focused on the front.

The test included one standalone system, Oculus Quest, which received positive notes for being generally easy to use for all the contents due to multiple reasons like the lack of tether and independence from external software. It, however, also received negative notes for reasons such as lower screen brightness, lower resolution and lower render fidelity in one of the test contents, Richie's Plank Experience. The lower brightness was rated as being inconvenient and the lower render fidelity is further elaborated on in the next section, Content.

The reflections and light leakage were perceived to be inconvenient at most and they were only visible in dark scenes. The overall fit of the HMDs was also noted to be related to the perceived FOV. Additionally, the FOV was perceived as being slightly narrow in all of the systems which in turn reduced immersion.

One expert noted systematic cybersickness that was attributed to using HP Reverb. The expert listed vertigo, eye fatigue, and changes in depth-perception as the most notable symptoms. The expert did not attribute the cybersickness to any specific content. A different expert noted problems in having both eyes focus at the same time on the display of HP Reverb. The same expert also noted a fogging of lenses in Reverb.

Finally, one expert noted that a useful safety feature was present on Oculus Quest by default. Similar features would have also been available on the other platforms via software solutions. The safety feature in Oculus Quest can be described as an automatic real-life video feed pass-trough or overlap when the user approaches the edge of the designated play area.

5.8. Content

Two findings repeated themselves throughout our expert evaluation results. Firstly, the perceived severity of the findings is heavily context-dependent; this is true to the extent of all perceived severities depending on the selected content. Secondly, the subjectiveness of VR experience makes objectively classifying the goodness of an individual experience an extremely difficult task. This can be seen in table 2 as different experts preferred different systems due to perceiving different aspects as being the most important ones.

It was also noted that different contents objectively required different types of performance from the VR systems. Freedom of movement provides a concrete example of this; video content had the smallest benefit from the expert being able to move their hands comfortably whereas Beat Saber required extensive hand movement. Similarly, the freedom of head movement was more valuable in Richie's Plank Experience and SUPERHOT VR than it was in videos and Beat Saber. The experts also noted that Richie's Plank Experience was commonly associated with feelings of cybersickness.

Additionally, the choice of the content itself required careful planning as this aspect also had its context-dependency. *Half-Life: Alyx* was considered as a content to be used for testing but was left out due to it not being available for the standalone system. The severity of this cannot be reasonably determined but we consider this to be a noteworthy and important factor as the limitations of the standalone system had an indirect but large effect on the study.

Two specific notes were also made regarding Richie's Plank Experience. The distinctly lower fidelity on the standalone system was rated as being either annoying or distracting and the experts unanimously noted the experience to be the most cybersickness inducing one. This is related to the high optical flow which is discussed in section 2.7.1.

Some game mechanics were also noted to be more error-prone than others. The "pull an item from behind your back" -mechanic was extensively used by Richie's Plank Experience and it oftentimes induced hand tracking errors which resulted in the experts' virtual hands moving erroneously and sometimes traveling several virtual meters away. Mechanics involving crouching also had similar effects. These were more apparent in SUPERHOT VR as the experts sometimes crouched while dodging enemy bullets. These specific tracking problems were perceived as being up to severe as they had a large negative impact on the gameplay.

5.9. System Rankings

Our findings on rankings are best presented in the table form as seen in table 2. Each group of three lines under the different categories shows the three experts in the same order. The leftmost system was ranked the best and the rightmost system was ranked the worst on each line.

After each VR system name, there is a number in parentheses; these numbers are shown to highlight the relative performances of the systems in addition to ranking them from best to worst within the specific content. The verbal equivalents for the numeric rankings from five to one are the following: excellent (5), good (4), fair(3), poor (2), and bad (1). The discussion on these findings can be found in its corresponding discussion section, 6.9.

Table 2. Expert evaluation rankings. The asterisk symbol (*) denotes the use of wireless adapter.

Videos				
Quest (4)	Vive Pro* (3)	Vive Pro (3)	Rift S (3)	Reverb (2)
Vive Pro* (4)	Vive Pro (4)	Rift S (3)	Reverb (2)	Quest (1)
Reverb (5)	Quest (4)	Rift S (4)	Vive Pro* (3)	Vive Pro (3)
Beat Saber				
Quest (5)	Rift S (4)	Vive Pro* (3)	Vive Pro (3)	Reverb (2)
Vive Pro* (4)	Vive Pro (4)	Rift S (3)	Quest (2)	Reverb (1)
Vive Pro* (5)	Vive Pro (4)	Rift S (3)	Quest (3)	Reverb (3)
Richie's Plank Experience				
Rift S (4)	Quest (4)	Vive Pro* (3)	Vive Pro (3)	Reverb (1)
Vive Pro* (4)	Vive Pro (4)	Rift S (3)	Quest (2)	Reverb (1)
Vive Pro* (5)	Vive Pro (4)	Reverb (4)	Rift S (3)	Quest (2)
SUPERHOT VR				
Quest (5)	Rift S (5)	Vive Pro* (3)	Vive Pro (3)	Reverb (3)

Table 2 continues below

Continuation of Table 2				
Rift S (4)	Reverb (3)	Quest (2)	Vive Pro* (1)	Vive Pro (1)
Vive Pro* (5)	Vive Pro (4)	Reverb (3)	Rift S (3)	Quest (3)
Overall rank by expert				
Quest	Rift S	Vive Pro*	Vive Pro	Reverb
Vive Pro*	Vive Pro	Rift S	Quest	Reverb
Vive Pro*	Vive Pro	Quest	Reverb	Rift S

6. DISCUSSION

In this section, we discuss the findings within the same order and categorization as they were presented before in section 5. Similarly like before, the terms inconvenient, distracting, annoying, and severe are used defined in section 4.7.2. Multiple parts of these following discussions have been intentionally left out from their respective sections and placed into the sections of user experience, 6.7, and content related effects, 6.8, to keep the discussions systematic.

6.1. Displays

In the findings section, we highlighted the resolution of the display and choice of display technology. The current state of consumer VR technology seems to be trending towards higher resolution and sometimes higher refresh rate which is a positive sign [3, 4]. However higher display refresh rates might not be required to produce good VR experiences, as this was something none of the experts paid much attention to. Additionally, selecting OLED or AMOLED as display technology to bring natural contrast in the VE might be more valuable of a choice than it is commonly presented to be in consumer comparisons.

We state that the trend towards higher display resolution is a desirable choice. Getting closer to achieving a resolution where individual pixels become imperceptible is, in our opinion, one of the primary tasks to further develop VR experiences. This should provide users with greater immersion and heightened user experience. We further assert that the efforts in performing this leap in display technology should be focused on platforms that are able to provide high contrast similar to the theoretically infinite contrast of OLED or AMOLED displays. This is due to the increase in contrast alongside pixel density providing a closer approximation of how to world is normally perceived.

In conclusion, a sufficient refresh rate has mostly been achieved to match the commonly enjoyed VR content, a sufficient display resolution however has not been achieved, and finally, a sufficient color fidelity has been achieved but has not been implemented for all solutions.

6.2. Lenses

Similarly to displays, few attributes of lenses received the most attention. These attributes were CA, Fresnel banding, and god rays. The systems offered different variations of Fresnel banding and CA but god rays were generally viewed as only being inconvenient while Fresnel banding and CAs received more severe ratings.

As opposed to our conclusions on displays, there are little recommendations we can give on lenses as the choice between Fresnel lenses and traditional lenses creates an optimization problem between lens artifacts and weight. Fresnel lenses can be built to be considerably lighter and thinner than traditional lenses but they also introduce more chances for light to scatter. They also have the disadvantage of having visible banding. [44, 45, 13]

We state that the lens design of VR systems should receive improvement efforts at priority similar to the displays themselves as the displays and lenses work in conjunction to produce the visual field. This is because a system with an optimal display will not provide optimal visual experience if lenses introduce distortions like god rays. The opposite is true for a system with optimal lenses and display that introduces artifacts like SDE.

6.3. Controllers

Our findings and opinions on controller designs were unanimous regarding which design goals were considered to be crucial. In conclusion, the goals that experts valued the most were as follows: compact form factor, ergonomic design, rigid construction, haptic feedback, low weight, and handedness.

The controller design used by Oculus Quest and Oculus Rift S provided a user experience that we found no complaints on. The controller's form factor was noticeably smaller compared to some other controllers which in turn provided lower weight. The ergonomic grip and solid construction resulted in the overall feeling of being durable. Additionally, the mirrored design between the left and right hand allows for good feel and users to easily place the controllers in their respective hands for contents where this matters. The handedness in the design is obvious enough to be noted even if the user is not able to directly see the controllers due to wearing an HMD. Finally, the controller's battery compartment hatch also withstood rigorous movements without becoming loosened.

A Counterexample in the overall design is provided by HP Reverb's controller. The comparatively large and clumsy form factor and front weight resulted in large inertia when either moving the controller or turning it. Additionally, a good grip on the controller was harder to achieve due to the controller not being rounded enough, and due to how buttons were placed. The overall feel of the controllers was weak and cheap, which probably contributed to the battery hatch becoming loose under rigorous movement.

The feedback of the controllers can be split into two equally important parts; the haptic feedback and the feedback provided by buttons and joysticks. The haptic feedback, such as vibrations, should be strong enough to be noticed even during intense movement. The feedback provided by buttons and joysticks should also be clearly distinguishable.

6.4. Tethered and Wireless Solutions

The effects of wirelessness were found to be positive and wirelessness generally offered better immersion. The increase in immersion was directly correlated to the amount of movement in the content and the reduction of immersion due to tether ranged from imperceptible to being rated annoying.

Regarding the tethers themselves, we conclude that three aspects generally determined the amount of immersion reduction from the tether. These aspects were the rigidness of the tether, the weight of the tether, and the position of the point where the tether connects to the HMD. A light and flexible tether should be chosen over heavy and rigid, as it is mostly achieved already. Furthermore, the tether should be placed in the middle of the back of the user's head. This should minimize the torque induced on the user's head and minimize the chance for the tether to collide with the user's hands, which was an issue in Oculus Rift S.

Additionally, we conclude that wireless solution was always preferred over tethered solution even when wirelessness required an additional battery and battery connector. This combined with the effects of tether ranging from imperceptible to annoying results in tether being sometimes non-factor and generally being a design choice that should be avoided when possible as the added weight from battery and antenna were perceived to be less severe than the effects of the tether.

6.5. Audio

Of the categories we have defined in findings, audio had the least amount notes. Some experts were more focused on the experience around the audio system rather than the audio quality itself, which suggests that the general ease of use is also important.

In conclusion, there was a balance between the built-in easy to use speakers and between the adjustable over-ear speakers. The built-in speakers provide noticeably lower audio quality but they do not require any sort of attention from the user. The over-ear speakers require adjusting every time the HMD is equipped but they also provide noise isolation and better audio quality. Especially in noisy environments, this reduction of ambient noise plays an important part in creating immersion.

Similar to the other various findings that have user experience related aspects, the balance between ease of use and audio quality is hard to define optimally. This highlighted the subjectiveness of the VR experience as different experts preferring different configurations. A viable solution is to offer an easy external audio connection in the HMD so the user can upgrade the audio experience with higher quality headphones if they wish to do so.

6.6. Tracking

Tracking received the largest number of mixed notes. We were able to narrow down the probable causes for these differences into a few items; reflective surfaces and amount of indoor lighting.

Our hypothesis on the problems of outside-in tracking is that most of its problems can be solved by covering IR-reflective surfaces such as mirrors, window panes, display cases, or TVs, and turning off indoor lights that are exceptionally bright and by.

Our hypothesis on the problems of inside-out tracking is that most of its problems could be solved by turning off indoor lights that are exceptionally bright while still making sure that adequate lighting is present. Ensuring that detailless surfaces do not cover large portions of the view of the cameras also proved to be important.

These hypotheses are based on troubleshooting performed when the experts were setting up the systems in their respective environments and after the test. Least tracking error was encountered by the expert who had large reflective surfaces covered. Another expert instead could limit the number of after turning off powerful ceiling lighting.

Tracking related problems were most often triggered by movements and game mechanics that somehow blocked either the user's hands or head. In the case of inside-out tracking, movements that placed the user's hands outside of the field of view of the cameras presented the largest amount of problems as the cameras are used to track the controllers, while the position of user's head was less relevant. In the case of outside-in tracking, movements that placed either the user's head or hands in a blind spot from the point of view of the base stations presented most problems as the base stations are used the give the light signal to the controllers and the HMD.

For the systems that had inside-out tracking the number of cameras had a negative correlation with the number of tracking errors noted. Oculus Rift S with its five cameras had the least problems followed by Oculus Quest with its four cameras. HP Reverb had the most problems and it had only two cameras. This hints towards a causal negative relationship between the number of cameras and the amount of tracking errors. We suspect that a similar relationship could be found between the number of base stations and tracking errors when outside-in tracking is used.

The set-up phases for the different systems were not directly in the scope of our study but it should be noted that inside-out tracking required a more permanent approach in the VR system's installation as the base stations have clear guidelines for optimal placement. In some cases, this optimal placement requires rearranging furniture or acquiring mounts for the base stations.

6.7. User Experience

This section contains a variety of discussion that relates to the user experience or does not fit other discussions. The various subjects are loosely grouped based on their themes.

The meta-level finding of the experts not making notes on M2P latency suggests that a sufficiently small latency has already been achieved to avoid inducing any immediately notable side effects like disorientation. However, further improvement might heighten the user experience by making the system feel more responsive.

The FOV was generally noted to be either just large enough or slightly narrow. This leads us to believe that a reasonably large increase in immersion could be achieved by increasing the FOV.

We assert that the ability to adjust the IPD of an HMD physically is an important aspect of user comfort. Systems that had physical IPD adjustment generally received fewer findings that were related to feelings of cybersickness or discomfort in eyes. This is highlighted by one expert reporting systematic cybersickness symptoms including changes in-depth perception when using HP Reverb, a system that had no physical IPD adjustment.

The subjectiveness of VR experience can be extended to cover the preferences of users in regards to the physical attributes of the HMD. One expert noted preference towards low weight and streamlined use, a different expert preferred high resolution even if lens problems were noted by other experts and the third expert placed more importance on general image quality and less importance on weight. Similar preferences were found in relation to multiple other factors like audio quality, wirelessness, presence of physical buttons, and fastening method of the HMDs.

Due to the large subjectiveness of user experience, we conclude that from a manufacturers' point of view the user experience might be best optimized by relying on the wisdom of the crowd that can be acquired by user studies as no perfect solution exists. It should be noted that the chosen content also has a large effect on many variables. These are discussed directly below.

6.8. Content

This section discusses aspects that have varying severities based on the content they are inspected with. The section also discusses the conclusions that can be drawn from table 2.

We state that at least the following vary in their perceived severeness or importance depending on the content: SDE, Fresnel banding, CA, god rays, need for contrast, tracking problems, controller problems, effects of wirelessness and risk of cybersickness. Additionally, we state that these variables can be categorized based on their severities changing depending on either content or the user's movement.

In our opinion, the first five variables listed, SDE, Fresnel banding, CA, god rays, need for contrast, fall in the first category that changes in perceived severity depending on content. All of these variables decrease in perceived severity as the movement speed of the content increases. This decrease is a result of the user paying less attention to these details as the content demands more focus from the user due to its speed. Three of these variables; the need for contrast, CA, and god rays are additionally dependent on either the brightness or contrast of the content. Increased brightness and contrast of content make CA and god rays more prominent whereas decreased brightness of content makes the need for (screens ability to provide) contrast greater.

Furthermore, we assert that the tracking problems, controller problems, and the effects of wirelessness all vary depending on the user's movement. Tracking related problems and benefits of wirelessness are amplified by content that requires for the user to move around and especially rotate, whereas controller related problems are amplified by content that requires the for the user to give inputs to interact with the content.

As an example based on the contents used, videos and Beat Saber provide polar opposites in their requirements. In the context of videos, all the problems related to display and lenses increase in severity as the user is relatively static and is focused on the viewing experience whereas the problems related movement decrease in severity for similar reasons. In the context of Beat Saber, the user is focused on moving their body during intense gameplay; this results in decreased perceived severity of display and lens-related problems as the user is not focused on these details. As opposed to this, problems related to movement are amplified as the user might fail a level if any of their inputs or movements are measured incorrectly. Lastly, wirelessness will act similarly as it's benefits increase based on the amount of required movement.

Additionally the expert's notes on Richie's Plank experience being associated with cybersickness further support the association with high vection and optical flow [6, 19, 20]. The experts did not make similar notes on the other contents. This leads us to believe that the experts might have a tolerance to some amounts of vection and optical flow.

In conclusion, these relationships create an optimization problem similar to the one presented in section 6.7. This problem, however, is distinct as an individual consumer might know their preferred content which places this problem more at the consumers' hands as opposed to those of the manufacturers.

6.9. Summary

Finally, we will discuss the conclusions that can be drawn from table 2. As mentioned before, different experts valued different parts of the VR experience. Examples of these differences are variables such as; how well the HMD fits the expert's head, general ease of use, lack of tracking related problems, overall visual performance, and resolution. This can be seen especially clearly in Oculus Quest being placed in opposing ends of the chart in the various contexts. A similar example is HP Reverb being seen in opposing ends of the chart in the context of videos alone.

Drawing a singular all-encompassing conclusion is impractical in the context of our study as the aspects listed in our findings varied on three levels; on the content, on the system, and on the expert. While the experts agreed on the strengths and weaknesses of different systems they valued different aspects at different levels which resulted in the variance seen in table 2. Instead of further condensing the overall rankings that were presented in the table 2 we now crystallize the main points on each one of the different systems.

HTC Vive Pro, especially when the wireless adapter was used, exceeded in providing a good overall experience with its most prominent strength is the lack of notable problems and its most prominent weakness is its lack of ease-of-use. In contrast to this, Oculus Quest had the most easily approachable user experience as its strength and the least convincing graphical performance as its weakness. Oculus Rift S had the least positive and negative notes; this places in the middle ground of HTC Vive Pro and Oculus Quest. Lastly, HP Reverb offered a conflicting experience with its different software platform, highest display resolution, and the largest amount of visual distortions; these together resulted in Reverb receiving the overall lowest rank.

Finally, we briefly emphasize on the most important findings. Displays need the most improvements on their resolution as their other variables are satisfactory already. Lenses need overall improvements to lessen the number of aberrations present. Controllers and audio have both achieved satisfactory solutions that are present on some systems already. Wirelessness should be always preferred over tether even with its drawbacks. Tracking systems generally work to a satisfactory degree but need further proofing against errors. Finally, all of the previous are largely dependent on the content chosen and these intertwined dependencies create optimization problems in designing VR systems and in choosing a VR system for a specific use case.

7. CONCLUSION

In our thesis, we presented suggestions and guidelines for performing a future study to correlate user experience with the physical attributes of VR systems. Additionally, we performed a small scale pilot study to refine these guidelines. We also presented an expert evaluation of the aspects related to VR systems. Both of these approaches aimed towards understanding the correlations between the user experience and the physical attributes of VR systems, both in general as well as in specific use cases.

We learned valuable information on the optimization problems related to balancing different aspects of a VR system to suit the system for a wide variety of use cases. We also learned valuable information on the optimization problems related to choosing a VR system suitable for the user's specific needs. Both of these optimization problems further highlighted the subjectiveness of the VR experience. This is crystallized as different systems being more suited for different contents and for different users.

While different users can prefer different systems and contents, some variables, when improved upon, will result in improved overall user experience. The next goals for these variables should be achieving higher resolution, lesser distortions in lens design, creating natively wireless systems with good platform support, and improving the user experience with ergonomic design of the HMDs and controllers.

8. REFERENCES

[1] Choi S.W., Lee S., Seo M.W. & Kang S.J. (2018) Time sequential motion-to-photon latency measurement system for virtual reality head-mounted displays. Electronics 7, p. 171.

[2] Chang C.M., Hsu C.H., Hsu C.F. & Chen K.T. (2016) Performance measurements of virtual reality systems: Quantifying the timing and positioning accuracy. In: Proceedings of the 24th ACM international conference on Multimedia, ACM, pp. 655–659.

[3] PC Magazine UK (2019), The best vr headsets for 2019. https://uk.pcmag.com/virtual-reality/75926/the-best-vr-headsets. [Online; accessed 10-November-2019].

[4] Niora.net (2019), Vr headsets. https://www.niora.net/en/vr_headsets. [Online; accessed 10-November-2019].

[5] Wikipedia contributors (2019), Comparison of virtual reality headsets — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Comparison_of_virtual_reality_headsets&oldid=925143267. [Online; accessed 10-November-2019].

[6] Steven M.L. (2016), Virtual reality.

[7] Mazuryk T. & Gervautz M. (1996) Virtual reality-history, applications, technology and future .

[8] The 360 Guy (2020), The ultimate vr headset comparison table: Every vr headset compared. https://www.threesixtycameras.com/vr-headset-comparison-table/. [Online; accessed 07-June-2020].

[9] Kalloniatis M. & Luu C. (2007) Visual acuity by michael kalloniatis and charles luu. Webvision. The Organization of the Retina and Visual System 5.

[10] Sitter B., Yang J., Thielen J., Naismith N. & Lonergan J. (2017) 78-3: Screen door effect reduction with diffractive film for virtual reality and augmented reality displays. In: SID Symposium Digest of Technical Papers, vol. 48, Wiley Online Library, vol. 48, pp. 1150–1153.

[11] Cho J.m., Kim Y.d., Jung S., Shin H. & Kim T. (2017) Screen door effect mitigation and its quantitative evaluation in vr display. In: SID Symposium Digest of Technical Papers, pp. 78–81.

[12] Dodgson N.A. (2004) Variation and extrema of human interpupillary distance. In: Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, International Society for Optics and Photonics, vol. 5291, pp. 36–46.

[13] Wheelwright B., Gollier J. & Geng M. (2018), Hybrid fresnel lens with reduced artifacts. US Patent 10,133,076.

[14] Borrego A., Latorre J., Alcañiz M. & Llorens R. (2018) Comparison of oculus rift and htc vive: feasibility for virtual reality-based exploration, navigation, exergaming, and rehabilitation. Games for health journal 7, pp. 151–156.

[15] Ramamani B. & Anala M. A brief understanding about quality metrics in virtual reality .

[16] Kolasinski E.M. (1995) Simulator sickness in virtual environments, vol. 1027. US Army Research Institute for the Behavioral and Social Sciences.

[17] LaViola Jr J.J. (2000) A discussion of cybersickness in virtual environments. ACM Sigchi Bulletin 32, pp. 47–56.

[18] Kennedy R.S., Lane N.E., Berbaum K.S. & Lilienthal M.G. (1993) Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. The international journal of aviation psychology 3, pp. 203–220.

[19] Davis S., Nesbitt K. & Nalivaiko E. (2015) Comparing the onset of cybersickness using the oculus rift and two virtual roller coasters. In: Proceedings of the 11th Australasian Conference on Interactive Entertainment (IE 2015), vol. 27, vol. 27, p. 30.

[20] Singla A., Fremerey S., Robitza W. & Raake A. (2017) Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays. In: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, pp. 1–6.

[21] Heutink J., Broekman M., Brookhuis K.A., Melis-Dankers B.J. & Cordes C. (2019) The effects of habituation and adding a rest-frame on experienced simulator sickness in an advanced mobility scooter driving simulator. Ergonomics 62, pp. 65–75.

[22] Zhao J., Allison R.S., Vinnikov M. & Jennings S. (2017) Estimating the motion-to-photon latency in head mounted displays. In: 2017 IEEE Virtual Reality (VR), IEEE, pp. 313–314.

[23] Porcino T.M., Clua E., Trevisan D., Vasconcelos C.N. & Valente L. (2017) Minimizing cyber sickness in head mounted display systems: design guidelines and applications. In: 2017 IEEE 5th international conference on serious games and applications for health (SeGAH), IEEE, pp. 1–6.

[24] Skarbez R., Brooks Jr F.P. & Whitton M.C. (2017) A survey of presence and related concepts. ACM Computing Surveys (CSUR) 50, pp. 1–39.

[25] Papachristos N.M., Vrellis I. & Mikropoulos T.A. (2017) A comparison between oculus rift and a low-cost smartphone vr headset: immersive user experience and learning. In: 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT), IEEE, pp. 477–481.

[26] Wienrich C., Döllinger N., Kock S., Schindler K. & Traupe O. (2018) Assessing user experience in virtual reality–a comparison of different measurements. In: International Conference of Design, User Experience, and Usability, Springer, pp. 573–589.

[27] The Verge (2019), Google is discontinuing the daydream view vr headset, and the pixel 4 won't support daydream. https://www.theverge.com/2019/10/15/20915609/google-pixel-4-no-daydream-support-view-vr-headset-discontinued. [Online; accessed 19-February-2020].

[28] Atanasio + Martinez (2018), Reimagine | etihad a380 virtual reality experience featuring nicole kidman | 360 3d vr | directed by atanasio + martinez. https://vimeo.com/304661051. [Online; accessed 19-February-2020].

[29] Atanasio + Martinez (2018), Reimagine etihad a380 virtual reality experience 360 video 4k highbit. https://www.youtube.com/watch?v=xpOe7cFeY1k. [Online; accessed 19-February-2020].

[30] Beat Games (2018), Beat saber. https://beatsaber.com/. [Online; accessed 19-February-2020].

[31] Toast VR (2017), Richie's plank experience. https://toast.gg/. [Online; accessed 19-February-2020].

[32] Slater M., Usoh M. & Steed A. (1995) Taking steps: the influence of a walking technique on presence in virtual reality. ACM Transactions on Computer-Human Interaction (TOCHI) 2, pp. 201–219.

[33] Kennedy R.S., Lane N.E., Berbaum K.S. & Lilienthal M.G. (1993) Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. The international journal of aviation psychology 3, pp. 203–220.

[34] Brunnström K., Sjöström M., Imran M., Pettersson M. & Johanson M. (2018) Quality of experience for a virtual reality simulator. Electronic Imaging 2018, pp. 1–9.

[35] SECTOR S. & ITU O. Series p: Telephone transmission quality, telephone installations, local line networks audiovisual quality in multimedia services .

[36] Usoh M., Catena E., Arman S. & Slater M. (2000) Using presence questionnaires in reality. Presence: Teleoperators & Virtual Environments 9, pp. 497–503.

[37] Malińska M., Zużewicz K., Bugajska J. & Grabowski A. (2015) Heart rate variability (hrv) during virtual reality immersion. International Journal of Occupational Safety and Ergonomics 21, pp. 47–54.

[38] Craig1287 (2018), My mom playing beat saber. https://www.youtube.com/watch?v=Y2gXM-ir-JE. [Online; accessed 19-February-2020].

[39] Sutcliffe A.G. & Kaur K.D. (2000) Evaluating the usability of virtual reality user interfaces. Behaviour & Information Technology 19, pp. 415–426.

[40] Sutcliffe A. & Gault B. (2004) Heuristic evaluation of virtual reality applications. Interacting with computers 16, pp. 831–849.

[41] Stefan Forster (2019), Elements of iceland - 4k. https://www.youtube.com/watch?v=F-1weFCiYBA. [Online; accessed 29-May-2020].

[42] William Briscoe Photography (2019), A year in alaska (2018). https://www.youtube.com/watch?v=a8w5uYujhkY. [Online; accessed 29-May-2020].

[43] SoKrispyMedia (2018), Video game vehicle (360 vr video!). https://www.youtube.com/watch?v=8ieIar09dRk. [Online; accessed 29-May-2020].

[44] Scott Hayden (2018), Facebook patents hybrid fresnel that could make vr optics more clear. https://www.roadtovr.com/facebook-patents-hybrid-fresnel-make-vr-optics-clear/. [Online; accessed 03-June-2020].

[45] Joseph Flynt (2019), How do lenses in vr goggles work? https://3dinsider.com/vr-lenses/. [Online; accessed 03-June-2020].

9. APPENDICES

Appendix 1	The Physical Plank
Appendix 2	The final set of Questionnaire
Appendix 3	Expert evaluation findings

A physical real-world plank was created to enhance the Richie's Plank Experience. The top board length was 188 cm, width 19,5 cm, and thickness 2 cm. To the starting end, we perpendicularly screwed a thicker plank to the counterbalance whole structure. To the other end of the plank, we attached a sculpted curving piece of Styrofoam which enabled the plank to bend and sway slightly when a person was walking on it.

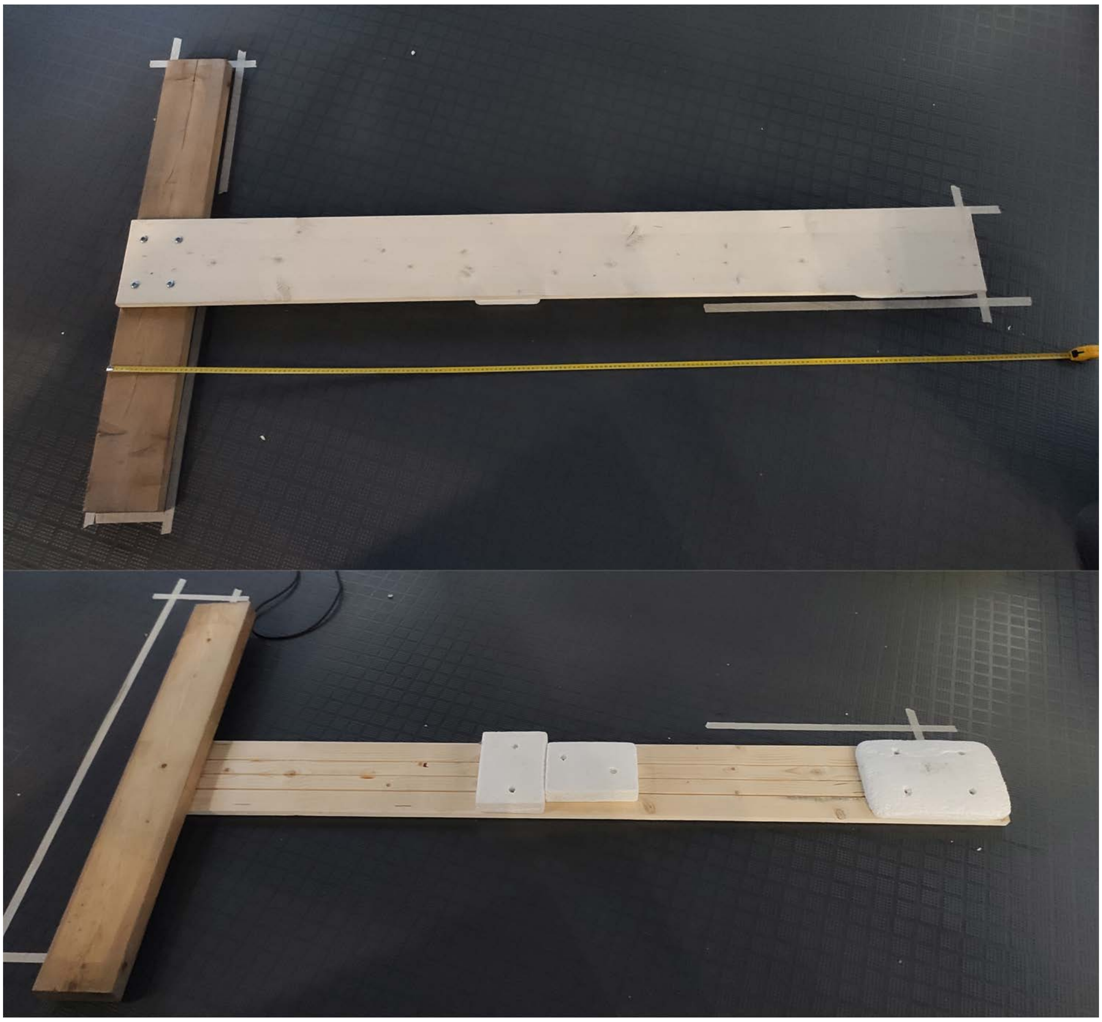


Figure 10. The physical plank created to enhance the experience.

Table 3. Demographic questionnaire.

Finnish	**English**
ID (filled by us)	ID (filled by us)
HMD (filled by us)	HMD (filled by us)
Lista, jossa HMD:t	List of the HMDs
Ikä	Age
Sukupuoli	Gender
Mies / Nainen / Muu	Male / Female / Other
Onko sinulla aiempaa kokemusta VR-laseista?	Do you have previous experience in using VR headsets?
Arvioi suunnilleen kokemuksesi määrää tunteina?	Estimate your experience in hours if you have any.
Ei yhtään / 0-1 / 1-10 / 10-100 / 100-1000 / >1000	None yhtään / 0-1 / 1-10 / 10-100 / 100-1000 / >1000
Oletko käyttänyt VR-laseja viimeisen viikon aikana?	Have you used VR glasses in the past week?
Onko sinulla silmälasit/piilolinssit?	Do you have glasses/contact lenses?
Onko sinulla todettu värisokeutta?	Do you have any kind of color blindness?
Onko sinulla muita näköongelmia jotka voisivat vaikuttaa VR-lasien käyttöön	Do you any other difficulties in eyesight that might affect wearing VR headsets?
Onko sinulla mitään sydänongelmia tai muita terveydellisiä tiloja joista meidän tulisi olla tietoinen?	Do you have any heart problems or other medical condition that we should know about?
Onko kokemusta Beat Saber VR-pelistä?	Do you have experince playing Beat Saber VR-game?
Onko kokemusta Richie's Plank Experience VR-pelistä?	Do you have experience playing Richie's Plank Experience VR-game?

Symptom scale				
SSQ Symptoms	**0**	**1**	**2**	**3**
General discomfort	None	Slight	Moderate	Severe
Fatigue	None	Slight	Moderate	Severe
Headache	None	Slight	Moderate	Severe
Eyestrain	None	Slight	Moderate	Severe
Difficulty focusing	None	Slight	Moderate	Severe
Increased salivation	None	Slight	Moderate	Severe
Sweating	None	Slight	Moderate	Severe
Nausea	None	Slight	Moderate	Severe
Difficulty concentrating	None	Slight	Moderate	Severe
Fullness of head	None	Slight	Moderate	Severe
Blurred vision	None	Slight	Moderate	Severe
Dizzy (eyes open)	None	Slight	Moderate	Severe
Dizzy (eyes closed)	None	Slight	Moderate	Severe
Vertigo	None	Slight	Moderate	Severe
Stomach awareness	None	Slight	Moderate	Severe
Burping	None	Slight	Moderate	Severe

Figure 11. SSQ in English

SSQ Pahoinvointikysely				
SSQ Oireet	**0**	**1**	**2**	**3**
Epämukava olo	Ei yhtään	Vähän	Jonkin verran	Paljon
Väsymys	Ei yhtään	Vähän	Jonkin verran	Paljon
Päänsärky	Ei yhtään	Vähän	Jonkin verran	Paljon
Väsyneet silmät	Ei yhtään	Vähän	Jonkin verran	Paljon
Vaikea tarkentaa silmiä	Ei yhtään	Vähän	Jonkin verran	Paljon
Lisääntynyt syljeneritys	Ei yhtään	Vähän	Jonkin verran	Paljon
Hikoilu	Ei yhtään	Vähän	Jonkin verran	Paljon
Kuvotus	Ei yhtään	Vähän	Jonkin verran	Paljon
Vaikea keskittyä	Ei yhtään	Vähän	Jonkin verran	Paljon
Paineen tuntu päässä	Ei yhtään	Vähän	Jonkin verran	Paljon
Sumea näkökenttä	Ei yhtään	Vähän	Jonkin verran	Paljon
Pyörrytystä/huimausta silmät auki	Ei yhtään	Vähän	Jonkin verran	Paljon
Pyörrytystä/huimausta silmät kiinni	Ei yhtään	Vähän	Jonkin verran	Paljon
Tasapainohuimaus	Ei yhtään	Vähän	Jonkin verran	Paljon
Lievä kuvotus/ruokahaluttomuus	Ei yhtään	Vähän	Jonkin verran	Paljon
Tarve röyhtäistä	Ei yhtään	Vähän	Jonkin verran	Paljon

Figure 12. SSQ in Finnish

Slater-Usoh-Steed Questionnaire (SUS) - Question on scale from 1 to 7
1 .Please rate your sense of being in the virtual environment, on a scale of 1 to 7, where 7 represents your normal experience of being in a place.
2. To what extent were there times during the experience when the virtual environment was the reality for you?
3. When you think back to the experience, do you think of the virtual environment more as images that you saw or more as somewhere that you visited?
4. During the time of the experience, which was the strongest on the whole, your sense of being in the virtual environment or of being elsewhere?
5. Consider your memory of being in the virtual environment. How similar in terms of the structure of the memory is this to the structure of the memory of other places you have been today? By 'structure of the memory' consider things like the extent to which you have a visual memory of the virtual environment, whether that memory is in colour,the extent to which the memory seems vivid or realistic, its size, location in your imagination, the extent to which it is panoramic in your imagination, and other such structural elements.
6. During the time of your experience, did you often think to yourself that you were actually in the virtual environment?

Figure 13. SUS in English

SUS skaala 1 - 7
1. Anna numeroarvo tuntemuksellesi omasta läsnäolostasi virtuaaliympäristössä asteikolla 1 – 7. 7 vastaa tavanomaista kokemustasi läsnäolosta. Virtuaalitodellisuudessa koin "olevani siellä": *Ei lainkaan - Hyvin paljon*
2. Missä määrin kokemuksessasi oli hetkiä jolloin koit, että virtuaaliympäristö oli todellisuutta? Tunsin, että virtuaalitodellisuus oli todellisuutta minulle *Ei koskaan - Melkein koko ajan*
3. Kun ajattelet virtuaalikokemusta, ajatteletko, että virtuaaliympäristö oli pikemminkin kuvia joita katselit, vaiko paikka jossa vierailit? Virtuaaliympäristö oli minulle enemmän: *Kuvia joita katselin - Paikka jossa vierailin*
4. Kokemuksen aikana, kumpi seuraavista tunteista oli kokonaisuutena voimakkaampi: tunne, että olet läsnä virtuaaliympäristössä vai tunne, että olit läsnä muualla? Koin enemmän olevani: *Muualla - Virtuaaliympäristössä*
5. Ajattele muistikuvaasi virtuaaliympäristössä käynnistä. Kuinka paljon tämä muistikuva muistuttaa rakenteellisesti muistoja muista paikoista missä olet käynyt tänään? Ajattele "rakenteellisen muistikuvan" tarkoittavan asioita kuten näkömuistia, ovatko muistikuvat värillisiä ja elävän sekä realistisen näköisiä, muistikuvan kokoa ja sijaintia mielikuvituksessasi, onko muistikuva panoraamamainen, sekä muita rakenteellisia elementtejä. Muistikuvissani virtuaaliympäristö on paikkana samantyyppinen kuin muutkin tänään vierailemani paikat: *Ei lainkaan - Hyvin paljon*
6. Ajattelitko usein itseksesi kokemuksen aikana, että olit todella virtuaaliympäristössä? Kokeen aikana ajattelin todella seisovani virtuaaliympäristössä: *En juurikaan - Hyvin paljon*

Figure 14. SUS in Finnish

MOS scale 1 to 5
1. How would you rate the picture quality? (circle the verbal option)
2. How would you rate the responsiveness of the system?
3. How would you rate your ability to accomplish your task?
4. How would you rate the immersion of the experience?
5. How would you rate your overall experience?

Figure 15. MOS1 in English

MOS skaala 1 - 5
1. Kuinka arvioisit kuvanlaatua? (ympyröi vastaus)
2. Kuinka arvioisit järjestelmän responsiivisuutta?
3. Kuinka arvioisit kykyäsi suoriutua tehtävästä?
4. Kuinka arvioisit kokemuksen immersion?
5. Kuinka arvioisit kokemuksen yleensä ottaen?

Figure 16. MOS1 in Finnish

MOS scale 1 to 5
1) How would you rate image colours?
2) How would you rate image contrast?
3) How would you rate the image borders?
4) How would you rate the movement continuity?
5) Did you notice any flicker in the sequence? Yes/No? If you noticed flicker, please rate it on the scale below
6) Did you notice any smearing in the sequence? Yes/No? If you noticed smearing, please rate it on the scale below

Figure 17. MOS2 in English

MOS skaala 1 - 5
1) Kuinka arvioisit kuvan värejä?
2) Kuinka arvioisit kuvan kontrastia?
3) Kuinka arvioisit kuvan rajoja?
4) Kuinka arvioisit liikkeen sulavuutta?
5) Huomasitko kuvassa välkyntää? Kyllä/Ei? Jos huomasit, arvioi se taulukkoon.
6) Huomasitko kuvan suttautuvan? Kyllä/Ei? Jos huomasit, arvioi se taulukkoon.

Figure 18. MOS2 in Finnish

This appendix contains the condensed notes of the experts on per-system and per-content basis in no specific order. The findings, their correlations, and implications are discussed in section 4. This list is not meant to convey the entirety of the knowledge the experts acquired during their testing, rather it represents the most notable findings from their combined notes. The full list of findings per expert is available in Finnish by contacting the authors.

HTC Vive Pro

General

- Light leakage and internal reflections of light are noticeable on dark backgrounds. Inconvenient.
- Two out of three experts had problems with head and hand tracking. Distracting.
- Two out of three experts had to cautiously adjust lighting conditions and visible reflective surfaces during the setup phase. Annoying.
- Fresnel banding effect is noticeable in the entirety of the visual field and CA is noticeable on the edges. Inconvenient.
- God rays are distinct in all high contrast areas. Distracting.
- An SDE was noticeable at all times. Inconvenient.
- The FOV was perceived to be slightly narrower than that of other systems despite it being similar according to the system's specifications. Inconvenient.
- The AMOLED display provided high contrast especially on scenes containing dark elements. Positive
- Over-ear headphones made the system feel somewhat unwieldy. Inconvenient.
- Over-ear headphones, however, also provided better sound quality and isolated outside noise. Positive.

Videos

- The relatively high weight of the HMD and tether make the system's suitability for long video or movie sessions questionable. Long sessions were not, however, specifically tested for. Inconvenient.
- The perceived sound quality was agreeable. Positive.

Beat Saber

- The shape of the controllers was well suited for the content and their weight caused no problems. Positive.

Richie's Plank Experience

- The tether's weight noticeably reduced immersion during flight for one expert. Inconvenient.
- Having to work around the tether during flight movements reduced immersion. Distracting

- Excluding the tether's effects immersion was perceived to be of good quality. Positive.
- Hand tracking acted erroneously when "pull an item from behind your back" -mechanic was used.
- SDE was most noticeable in this content against the bright sky, similarly to SUPERHOT. Inconvenient.

SUPERHOT VR

- SDE was most noticeable in this content against the bright background, similarly to Plank Experience. Inconvenient.
- Occasional jitter and jumping produced by the hand tracking when crouching has a large negative impact on the gameplay due to the game's movement mechanic. Severe.
- The game's controls require using the button combined into the controller's touchpad due to the small number of buttons on the controller. This feels unnatural and unwieldy. Distracting.

HTC Vive Pro with wireless adapter

General

- Light leakage and internal reflections of light are noticeable on dark backgrounds. Inconvenient.
- Two out of three experts had problems with head and hand tracking. Distracting.
- Two out of three experts had to cautiously adjust lighting conditions and visible reflective surfaces during the setup phase. Annoying.
- Fresnel banding effect is noticeable in the entirety of the visual field and CA is noticeable on the edges. Inconvenient.
- God rays are distinct in all high contrast areas. Distracting.
- An SDE was noticeable at all times. Inconvenient.
- The FOV was perceived to be slightly narrower than that of other systems despite it being similar according to the system's specifications. Inconvenient.
- The AMOLED display provided high contrast especially on scenes containing dark elements. Positive.
- The wireless antenna and its battery can get noticeably warm which might require careful placement. Inconvenient.
- Wireless antenna removes the need for tether but adds a battery cable which sometimes requires maneuvering around if not positioned wisely. Inconvenient but net positive.
- The perceived sound quality was agreeable. Positive.
- Over-ear headphones made the system feel somewhat unwieldy but provided better sound quality and isolated outside noise. Positive.

Videos

- Compared to tethered version wirelessness provides slightly better immersion and freedom of movement when turning head. Positive.
- The relatively high weight of the HMD makes the system's suitability for long video or movie sessions questionable. Long sessions were not, however, specifically tested for. Inconvenient.

Beat Saber

- Compared to tethered version wirelessness provides somewhat better immersion and freedom of movement during playing. Positive.
- The shape of the controllers was well suited for the content and their weight caused no problems. Positive.

Richie's Plank Experience

- Compared to tethered version wirelessness provides clearly better immersion and freedom of movement during flight. Positive.
- Immersion was perceived to be of high quality. Positive.
- Hand tracking acted erroneously when "pull an item from behind your back" -mechanic was used.
- SDE was most noticeable in this content against the bright sky, similarly to SUPERHOT. Inconvenient.

SUPERHOT VR

- Compared to tethered version wirelessness provides clearly better immersion and freedom of movement during crouching and dodging. Positive.
- SDE was most noticeable in this content against the bright background, similarly to Plank Experience. Inconvenient.
- Occasional jitter and jumping produced by the hand tracking when crouching has a large negative impact on the gameplay due to the game's movement mechanic. Severe.
- The game's controls require using the button combined into the controller's touchpad due to the small number of buttons on the controller. This feels unnatural and unwieldy. Distracting.

Oculus Quest

General

- Light leakage and internal reflections of light are noticeable on dark backgrounds. Inconvenient.
- Two out of three experts (those who had problems with the tracking of HTC Vive Pro) evaluated Oculus Quest's inside-out tracking to be noticeably less error-prone than that of HTC Vive Pro's. Positive.

- Physical volume button improved ease of use. Positive.
- Rubberband like fastening system improved ease of use when removing or equipping the HMD. Positive.
- One expert noted the HMD to be front-heavy which reduced overall user experience. Inconvenient.
- The relatively low weight of the HMD was perceived to contribute to overall better experience and immersion. Positive.
- Automatic image pass-through when approaching the edge of the play area improved security. Positive.
- Fresnel banding effect and CA are noticeable at the edges of the display. Inconvenient.
- God rays are distinct in all high contrast areas. Distracting.
- An SDE was noticeable at all times. Inconvenient, some times distracting
- The OLED display provided high contrast especially on scenes containing dark elements. Positive.
- The standalone nature of the system provides easy to approach experience. Positive.
- The brightness of the display was perceived to be slightly lower than that of other systems. Inconvenient.
- The perceived sound quality is lower compared to that of HTC Vive PRO. Inconvenient.
- Perceived resolution of the display is relatively low compared to other systems. Inconvenient.
- All experts unanimously preferred the controller design found in Oculus Quest and Oculus Rift S. Positive.

Videos

- One expert noticed a flickering of the image. Inconvenient.
- The general ease of use of the system makes it suitable for casual video watching. Positive.

Beat Saber

- The accuracy of hand tracking suffers from movement performed outside of the FOV of the cameras. Inconvenient.
- The handedness of the controller design makes it easy to place the controllers in their respective hands which is important for this content. Positive.

Richie’s Plank Experience

- The accuracy of hand tracking suffers from movement performed outside of the FOV. Repeatedly cause by "pull item from behind your back" -mechanic. Distracting.

- Specifically in this content the graphical fidelity of all textures and overall experience was noticeably lower than that of other systems. The game itself is rendered with lower quality due to the limitations of the platform. Annoying.
- Flying provided a liberating experience due to the low weight and wirelessness of the system. Positive.
- After the experience, one expert noted mild cybersickness in the form of feeling ill.

SUPERHOT VR

- High freedom of movement and low weight of the system increased immersion greatly by allowing careful maneuvering around enemy bullets without restrictions. Positive.

Oculus Rift S

General

- Light leakage and internal reflections of light are noticeable on dark backgrounds. Inconvenient.
- Automatic image pass-through when approaching the edge of the play area improved security. Positive.
- All experts unanimously preferred the controller design found in Oculus Quest and Oculus Rift S. Positive.
- No physical IPD adjustment is present. Inconvenient.
- The tether is placed asymmetrically, which induces a slight tilt to the user's head if the tether is not carefully positioned. Inconvenient.
- Image colors were perceived as being bleak or mildly washed out. Inconvenient.
- An SDE was noticeable at all times. Inconvenient.
- Sound quality was perceived to be descant heavy and lacking bass. Inconvenient.
- Fresnel banding effect is noticeable in the entirety of the visual field except for the middle. Inconvenient.
- CA is noticeable on the edges. Inconvenient.
- God rays are somewhat distinct in high contrast areas. Distracting.
- The HMD offered the highest general comfort when wearing. Positive.

Videos

- Slight barrel aberration is present in all video content. Inconvenient.

Beat Saber

- One expert noted jitter in the tracking. Inconvenient.
- The handedness of the controller design makes it easy to place the controllers in their respective hands which is important for this content. Positive.

- The asymmetric tether required careful placing so that it would not hit the expert's hands. Inconvenient.

Richie's Plank Experience

- Hand tracking acted erroneously sometimes when "pull an item from behind your back" -mechanic was used. Inconvenient.
- One expert noted jitter in the tracking. Inconvenient.
- Excluding the tether's effects immersion was perceived to be of good quality. Positive.

SUPERHOT VR

- The experts made no findings specific to this content besides reduced immersion from the tether. Inconvenient.

HP Reverb

General

- Light leakage and internal reflections of light are noticeable on dark backgrounds. Inconvenient.
- The HMD offers the least noticeable SDE and the best resolution. Positive.
- Image colors were perceived as being bleak or mildly washed out. Inconvenient.
- Two out of three experts noted that the HMD has immense CA in the whole FOV which greatly reduces perceived image quality. Annoying.
- One expert noted high mura, with the brightest part being in the middle. The mura has a circular form of concentric rings correlating with the Fresnel lenses. Distracting
- The FOV has a warped distortion that is noticeable when turning head. Distracting
- The tether used by the system was most rigid and heavy compared to other tested systems. Distracting.
- One expert noted systematic cybersickness and eye strain related to this system. The cybersickness presented in the form of vertigo and slight changes in depth perception. Severe.
- Lack of physical IPD adjustment reduced the perceived quality of experience. Annoying.
- Over-ear headphones made the system feel somewhat unwieldy, but to a lesser extent than in HTC Vive Pro. Inconvenient.
- Over-ear headphones, however, also provided better sound quality and isolated outside noise. Positive.
- Rubberband like fastening system improved ease of use when removing or equipping the HMD. Positive.

- Display related problems reduce overall immersion. Distracting.
- All experts unanimously disliked the controller design due to it's unwieldy and frail-feeling design. Distracting.

Videos

- Slight barrel aberration is present in all video content. Inconvenient.
- Two experts noted the controller design being slightly impractical even for watching videos. Inconvenient.

Beat Saber

- Controller battery hatches opened repeatedly when the controllers were waved around intensely. Severe.
- The controllers were unnecessarily weighty and large for the content. Distracting.
- The handedness of the controller design makes it easy to place the controllers in their respective hands which is important for this content. Positive.
- Haptic feedback in the form of vibration of the controllers was inadequately weak. Distracting.

Richie's Plank Experience

- One expert noted fogging of the HMD after fastening it.
- Inside-out based hand tracking acted erroneously when "pull an item from behind your back" -mechanic was used.
- Overall immersion was low compared to other systems due to display related problems and tether. Inconvenient.

SUPERHOT VR

- High CA and mura were distinct in the high contrast environment. Distracting.
- Hand tracking acted erroneously when moves were performed at odd angles; such as during crouching or behind the head. Distracting.
- The controllers were perceived to be needlessly large for the content. Inconvenient.