EXPLORING VIRTUAL REALITY MECHANICS IN PUZZLE DESIGN

Taneli Nyyssönen

Master's Thesis
May 2020

DEPARTMENT OF FUTURE TECHNOLOGIES
UNIVERSITY OF TURKU

UNIVERSITY OF TURKU
Department of Future Technologies

NYYSSÖNEN, TANELI: Exploring virtual reality mechanics in puzzle design

Master's Thesis, 138 p., 8 app. p.
Interaction Design
May 2020

This study explores practical implementations of various custom virtual reality mechanics, developed specifically for this study, in the context of puzzle game design with an experimental approach. These mechanics include swimming, crawling, climbing, and hiding of objects into virtual spaces in different ways. There are two different variations for each of the mechanics, the more and less realistic one, the main goal of the study being to discover which one is more enjoyable to use by testing them within puzzle levels designed for the mechanics. The final result is determined by the preferences of the testing sample.

The testing sample consisted of 22 volunteers acquired using convenience sampling, while the evalution methods consisted of a questionnaire which utilised a Likert scale, divided into pre- and post-questionnaire parts, in addition to observations. The enjoyability of the mechanics was evaluated based on four different aspects: perceived realism, personal traits and abilities of the testing sample, the testing order, and perceived difficulty. A special interest in the study is to see whether real-life skills corresponding to the studied mechanics affect the enjoyment and performance levels in the respective mechanics.

According to the results, the more realistic mechanics were found to be more enjoyable by a significant margin, suggesting that they should be utilised in the future. Additionally, the testing order did play a role in the enjoyment, but especially in the performance, while some of the personal skills and traits also affected the enjoyment in the mechanics related to those skills, which was a highly interesting result.

Keywords: virtual reality, puzzle, game design, experimental, mechanics, interaction design

TURUN YLIOPISTO
Tulevaisuuden teknologioiden laitos

NYYSSÖNEN, TANELI: Virtuaalitodellisuusmekaniikkojen tutkiminen ongelmanratkaisupelien suunnittelussa

Pro gradu -tutkielma, 138 s. 8 liites.
Vuorovaikutusmuotoilu
Toukokuu 2020

Tämä tutkielma käsittelee virtuaalitodellisuusmekaniikkojen, jotka on kehitetty tätä tutkimusta varten, käytännön soveltamista ongelmanratkaisupelien suunnitteluun kokeellisesti. Kyseessä olevat mekaniikat ovat uiminen, kiipeäminen, ryömiminen, sekä objektien piilotus eri tavoin virtuaaliympäristöihin. Jokaisesta mekaniikasta on kehitetty kaksi erilaista variaatiota, enemmän ja vähemmän realistinen, ja tutkielman päätavoitteena on selvittää kumpaa mekaniikkaa on mieluisampi käyttää testaamalla mekaniikoita niitä varten kehitettyjen ongelmanratkaisutasojen avulla. Lopullinen tulos määräytyy testaajien mieltymysten perusteella.

22:n hengen testausryhmä valittiin mukavuusotantaa käyttäen, ja arviontityökalut koostuivat etu- ja jälkikäteiskyselyistä, joiden kysymykset käyttivät Likert-asteikkoa, sekä havainnoista. Mekaniikkojen mieluisuutta arvioitiin neljällä eri tavalla: mielletty realistisuus, testaajien henkilökohtaiset kyvyt ja ominaisuudet, testausjärjestys, sekä mielletty vaikeustaso. Erityisen mielenkiintoista oli nähdä, mikäli testaajien tutkittuja mekaniikkoja vastaavat tosielämän taidot vaikuttaisivat mieluisuuteen tai suorituskykyyn kyseisten mekaniikkojen kohdalla.

Tulosten mukaan realistisemmat mekaniikat olivat huomattavasti mieluisempia, mistä voidaan tehdä johtopäätös, että niitä kannattaisi käyttää tulevaisuudessa. Lisäksi testausjärjestyksellä oli osansa mieluisuudessa ja etenkin suorituskyvyssä, kun taas testaajien henkilökohtaiset kyvyt liittyen tutkittuihin mekaniikkoihin myös vaikuttivat mieluisuuteen kyseisten mekaniikkojen osalta, mikä oli hyvin mielenkiintoinen tulos.

Asiasanat: virtuaalitodellisuus, ongelmanratkaisupelit, mekaniikat, pelisuunnittelu, kokeellinen, vuorovaikutusmuotoilu

Acknowledgements

I would like to acknowledge my supervisors, Jouni Smed and Natasha Skult, for assisting me through this greatest academic endeavour in my life this far.

Additionally, I would like to thank my partner in life, Jessica Blackburn, for rigorous spellchecking, love and everything else that you have provided before, during and (hopefully) after this process as well.

This thesis is dedicated to *you*.

"You Are The Moon"

Shadows all around you as you surface from the dark
Emerging from the gentle grip of night's unfolding arms
Darkness, darkness everywhere, do you feel all alone?
The subtle grace of gravity, the heavy weight of stone.

You don't see what you possess, a beauty calm and clear
It floods the sky and blurs the darkness like a chandelier
All the light that you possess is skewed by lakes and seas
The shattered surface, so imperfect, is all that you believe.

I will bring a mirror, so silver, so exact
So precise and so pristine, a perfect pane of glass
I will set the mirror up to face the blackened sky
You will see your beauty every moment that you rise.

- Salpeter Greta Morgan (2006) (with permission)

Contents

1 Introduction

This thesis focuses on analysing real-world applied virtual reality (VR) puzzles based on test subject performance and feedback from the viewpoint of game design. The main goal of this thesis is to (loosely) determine whether it is more enjoyable to solve virtual reality puzzles with more realistic core mechanics or would it be better to use more of a gamelike approach. Additionally, the thesis looks into other kind of aspects which could affect the enjoyment, mainly testing order, testers' personal traits and abilities, and perceived puzzle difficulty.

Another goal is to see how much the testers' perception of their skills and traits corresponds with their performance in the puzzles and how this performance affects the enjoyment. Although the results are applicable only to the puzzles in question and only for this specific testing sample and cannot be generalised because of the small sample size, this thesis could act as a basis for further research in the field.

The mechanics which are studied in this thesis include VR-simulated swimming, crawling, and climbing in addition to studying different methods of hiding objects into virtual environments and comparing those with how they could be hidden in the real world. The mechanics are tested as a form of AB-testing, where testers get to test similarly themed puzzles with two different types of movement and/or other mechanics.

The thesis uses a mixed method of research containing a testing questionnaire which includes both multiple choice and open-ended questions, in addition to observations, which are gained during the testing and the filling of the questionnaires.

We will first go through the concepts of virtual reality, games, and game design, after which the puzzles developed for this thesis are presented along with the testing methodology. Finally, the results of the testing are presented, analysed and reflected upon with some thoughts left out for other future researchers of similar topics.

Similar publications in the field

There is little existing research related to combining virtual reality mechanics with game development currently, as VR games are still a young concept, and there are even less if any publications which would study exactly puzzle mechanics in VR, and none which combine these two and focus on the enjoyment aspect of realism. This was also a major reason for conducting this study in the first place.

The closest publications to the topic of this thesis which can be found, are projects which aim to enhance various real-life mechanics with VR. One of these is, for example, *AquaCAVE* [1], which is an augmented swimming environment combined with immersive surround-screen in VR, where the users are actually swimming while the experience is enhanced by a virtual environment that the users can see. Another one is called *Dungeons & Swimmers* [2], an interactive audio- and motion-based exergame for swimming which intends to create realistic swimming on dry land by using motion tracking, which is a similar idea as in this thesis although it is not done in VR. Also neither of these publications investigate swimming as a puzzle mechanic.

On the subject of publications which have researched puzzle mechanics in virtual reality, the closest one which this study found is *The collaborative cube puzzle: a comparison of virtual and real environments* [3], which investigates the differences of completing a Rubik's-type puzzle-cube in VR and in the real world and compares the result of those. While the publication is about a puzzle in virtual reality, it mostly focuses on the differences between real life and VR when trying to solve the cube, not in two alternative VR-mechanics for a puzzle which is the case for this thesis.

2 What is virtual reality?

Virtual reality (VR) refers to a computer-generated environment, which can be viewed and interacted with through the usage of some type of device, usually a VR headset. Unlike when viewing things through a screen, where the position of the display remains fixed, in VR everything is seen from a "true" first-person view, meaning the player directs what they can see by moving their head in real life. VR makes it possible for the user to experience extremely immersive three-dimensional worlds which can be rendered in all 360 degrees around the user. Figure 1 showcases the different dimensions of virtual environments in order to make it easier to see how virtual reality compares with the normal world.

Next, in order to get a sense of how virtual reality has developed from early concept to what it currently is, the history will be briefly recited here. After that, the reader will be familiarised with the basic concepts of games, game design and designing games for VR in Chapter 3.

Brief history of virtual reality

The following brief history is mostly focused on the most well-known pure VR-equipment and leaves out the various augmented reality devices and mobile add-ons. Some of the contents of the history are based on an article by Virtual Reality Society [5] with individual sources added when necessary.

Spiritual ancestors of virtual reality

The earliest forms of attempted virtual reality are thought to have been the 360-degree murals which are also known as panoramic paintings. These murals usually depict historical events and as they fill the entire field of view of the beholder, they can make them feel a resemblance of being actually present at the scene. An example of this can be seen in Figure 2 which depicts a famous Paraguayan battle scene.

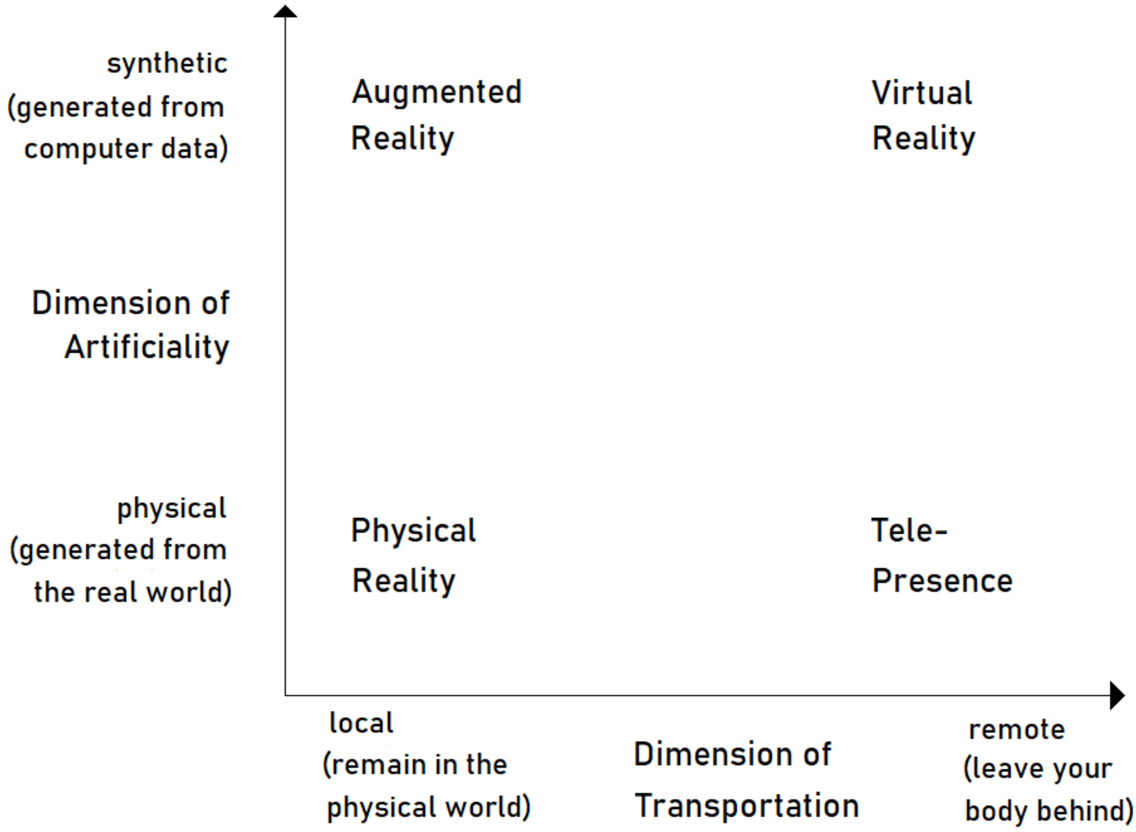

Figure 1: In virtual reality the user is separated from their body, as in they get to act as someone else (e.g., in a game) and their actions function differently than in the real world, while the world they are experiencing is completely artificial. As a comparison, augmented reality means that parts of the real world are replaced with computer-generated content in an attempt to enhance experiencing the real world. An example of augmented content could be long lost artifacts in museums, augmented for visitors to view in 3D similar to holograms. On the dimension of transportation, tele-presence refers to allowing a person to feel as if they were present at a place other than their true location. This can happen in, for example, various types of calls in which the physical location of the callers does not change, but their presence is somewhere else in a dimension where the two calls connect, so they are in a sense leaving their body behind. [4] p. 192.

Figure 2: A panoramic painting of the Battle of Avay from Paraguayan war in December 1868, which was fought between the forces of the Triple Alliance and the Paraguay, painted by Pedro Américo. [6]

The discovery

In 1838 Charles Wheatstone's research demonstrated that the brain processes the different two-dimensional images from each eye into a single object of three dimensions. Viewing two side by side stereoscopic images or photos through a *stereoscope* gave the user a sense of depth and immersion. The stereoscope was later succeeded by the *lenticular stereoscope* (David Brewster, 1849) and the *View Master* (William Gruber, 1939). [7]

Link Trainer, the first commercial flight simulation device in the world, was created by Edward Link in 1929. Largely used by the US military forces at the time, these "blue boxes" trained hundreds of thousands of pilots during World War II. [8]

In the mid-1950s, cinematographer Morton Heilig invented the *Sensorama* (patented 1962 [9]) which was a mechanical arcade-style theatre cabinet that would stimulate all the senses, not just sight and sound. It consisted of stereo speakers, a stereoscopic 3D display, fans, odor emitters and a motional chair. Sensorama is referred to as one of the first virtual reality systems and is displayed in Figure 3.

Figure 3: Morton Heilig's Sensorama simulator. [9]

Sensorama's creator also created the *Telesphere Mask* in 1960, which was the first example of a head-mounted display (HMD), although it did not feature any motion tracking. The first motion tracking HMD was called the *Headsight* and it was invented by two Philco Corporation engineers in 1961. Headsight allowed the user's head movements to move a remote camera, letting them inspect their surroundings quite naturally.

A few years later, Ivan Sutherland and his student Bob Sproull created the *Sword of Damocles* [10] in 1968, a heavy and uncomfortable HMD which was suspended from the ceiling and was capable of displaying very simple wireframe rooms and objects. Their creation is displayed in Figure 4.

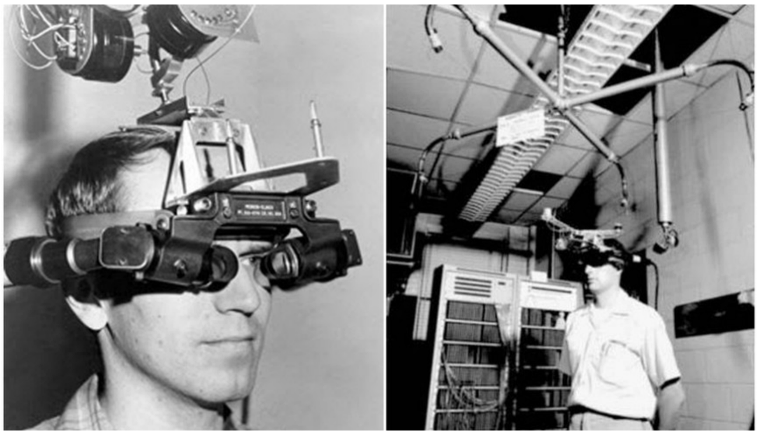

Figure 4: Ivan Sutherland's Sword of Damocles. [10]

In 1969, Myron Kruegere developed a series of experiences which he called "artificial reality". These experiences allowed users to communicate with each other through a responsive virtual environment despite being miles apart. This technology was called VIDEOPLACE, and is widely referred to as the first interactive VR system, although it was technically a form of augmented reality as it did not feature a headset. [11]

The convention

In 1987, the term *virtual reality* was coined by Jaron Lanier, founder of the visual programming lab (VPL). The VPL was the first company to sell virtual reality goggles (such as EyePhone) and gloves. In 1991 the Virtuality Group launched a range of arcade games and machines which were played using VR-goggles with less than 50ms latency stereoscopic 3D visuals. Some of them were even capable of multiplayer experiences through networking.

In 1993, Sega announced the Sega VR headset for the Sega Genesis console at the Consumer Electronics Show. The wrap-around prototype glasses had head tracking, stereo sound and LCD screens in the visor and were aimed to be sold at a price of 200 USD. Unfortunately, the headset was never released for commercial use due to technical difficulties and remains incomplete.

1995 was the release year of Nintendo Virtual Boy which was supposed to be the first ever portable console capable of displaying stereoscopic "3D" graphics. The product unfortunately failed commercially and was discontinued the following year due to complaints about its lack of colours (only black and red) which diminished its graphical outlook, price, and comfortability among the main complaints.

Figure 5: The first Oculus Rift designed by Palmer Luckey in 2016. [12]

Virtual reality today

During the 21st century, virtual reality development has taken giant leaps forward, major ones mentioned here, starting with the Google Street View in 2007, which consists of interactive panoramas taken from all around the globe. In 2010, Palmer Luckey designed the first prototype of the Oculus Rift (in Figure 5), which was the first VR-headset able to produce a 90-degree field of view, but not without distortion issues however.

In 2013, Valve Corporation discovered and freely shared the breakthrough of low-persistence displays which made latency-free VR experiences possible. This breakthrough was used in the creation of most (if not all) VR-headsets after its discovery, one being the *HTC Vive* (shown in Figure 6), which Valve created together with HTC Corporation (previously a mobile phone company) on 5th of April 2016. The first Vive featured what is called a "room-scale" tracking technology. This meant that users could freely move in the play area and their real-life movements would be reflected inside the virtual reality. Vive also had a quite realistic resolution (1080×1200 pixels per eye) with a refresh rate of up to 90Hz and a field of view of 110 degrees.

The headset and its two controllers function using tracking technology called Lighthouse, which utilises wall-mounted "base stations" for positional tracking using infrared light. The combination of the headset, controllers and the tracking system is called SteamVR, although SteamVR can be utilised via other types of headsets as well thanks to its OpenVR SDK, an open source programming interface created by Valve to allow communication with a VR system, allowing for creation of VR games and applications suited for almost any VR device. Meanwhile in 2014 Sony announced *Project Morpheus*, which later on in 2016 became the Playstation VR, a VR headset for the Playstation 4 gaming console.

Figure 6: HTC Vive created by Valve and HTC in 2016. [13]

On December 29, 2017, a device called *TPCAST Wireless Adapter* ([14]) was released for HTC Vive. TPCAST allows users to experience VR completely wirelessly and claims to have a sub 2 millisecond latency without affecting the image quality. Next year the HTC Vive Pro, an enhanced version of the HTC Vive was released on April 5, 2018, and it featured improved resolution (1440×1600 pixels per eye) and came with an additional camera which allowed users to see their real-life surroundings when stepping outside of the designated play area, making the device safer to use.

The future

On June 28, 2019, Valve released their first self-created VR-headset, the *Valve Index* for 1079 euros. Index features vastly improved technology with an experimental up to 144Hz refresh rate and a way of tracking the positions of each finger without the need for gloves in its controllers.

On October 15th the same year, a Finnish company called Varjo released their latest headset, *VR-2 Pro*, which among other features claims to feature a human-eye resolution (20/20 Eye Tracker^{TM}). The headset is said to be the most advanced one on the market to date, but it is only designed for enterprise use, having a quite steep pricetag of near 6000 euros just for the headset. It is also designed to function with SteamVR with its built-in SteamVR tracking. [15]

Only the future holds what shall be next in line for VR.

3 Designing games in virtual reality

Games are the most popular medium of today (as proven, for example, in this report from Reuters [16]), surpassing even the film industry in annual revenue with more and more games and gamers created every second. When it comes to virtual reality, VR-games are a very new addition to the vast repertoire of game genres, and maybe one of the most challenging ones to design. In order to get a glimpse of what is required for designing games for virtual spaces, the reader will be indulged in the definition of a game, followed by defining game design and then looking at game design from the perspective of virtual reality in more detail. The latter includes getting familiarised with the basic and more experimental types of movement possibilities in virtual spaces in order to be able to identify and compare them later on with the types used in the puzzles this thesis is focused on.

3.1 What are games?

Games *are a type of play activity, conducted in the context of a pretended reality, in which the participant(s) try to achieve at least one arbitrary, nontrivial goal by acting in accordance with rules.*
-Ernest Adams [17]

The following summary about various aspects of games is largely based on information found in *Fundamentals of Game Design, Third edition* by Ernest Adams [17].

Games consist of three main aspects:

- the core mechanics
- the user interface (UI), and
- the player, as illustrated in Figure 7.

Core mechanics implement the rules and restrictions that create the challenges for the player. These can be, for example, physics, internal economies, or progression mechanisms.

The user interface presents the game to the player. Through user interface the player interacts with the game world and experiences the

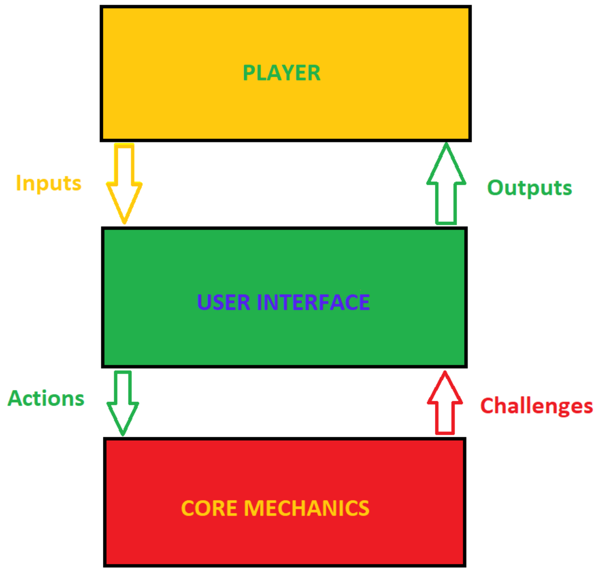

Figure 7: Building blocks of games, [17] p. 38.

game's aesthetics.

The player is the one who plays the game and has to abide by the rules and restrictions of the core mechanics which are presented to them through the user interface.

The user interface can be further divided into two parts, which are the *camera model* and the *interaction model* as seen in Figure 8. The interaction model describes how the player is capable of interacting with the game world. These models can be, for example, *avatar-based*, which means that the player is a character inside the game and will be able to perform actions through that character, which usually means that the player's actions are tied to the character's (physical) location within the game world, or *omnipresent*, which means that the player is not inside the game world but rather gets to view and influence it from the outside.

The second part of the user interface, camera model, refers to the way the game world is presented to the player visually (through a camera). The camera model can be static, which means it will always be fixed to one position, or dynamic, which means it will change based on the player's interaction with the game world, like movement. Common camera models include first-person, where the player sees the game world through someone's eyes (e.g., most 3D horror games), third-person where the

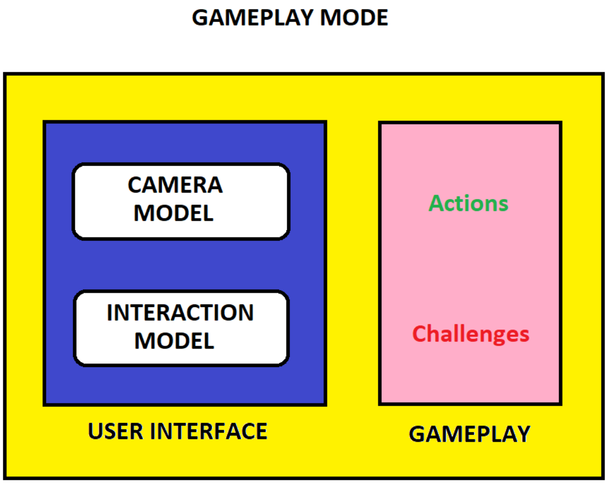

Figure 8: A gameplay mode consists of the camera and interaction models combined with gameplay (challenges and actions), [17] p. 41.

camera is positioned to spectate the main character from a short distance (e.g., role-playing games like MMORPGs), top-down, where the game world is seen straight from above as a two-dimensional field (e.g., various 2D-games), or isometric, in which the top-down view is at an angle which makes the two-dimensional game world appear as three-dimensional. There are also games which do not have a visual representation of their game world and for those the term camera model is redundant, and the layout of the screen needs a different kind of explanation.

The actions implemented by the user interface and the challenges created by the core mechanics can be combined to represent *gameplay*, which together with the user interface forms the *gameplay mode* as seen in Figure 8. Gameplay modes restrict the player's ability to interact with certain parts of the game world while simultaneously allowing interaction to others. The gameplay usually swaps they player in and out of various gameplay modes, such as combat, and each of those game modes brings their own subset of rules and restrictions which the user interface presents to the player. For example, the player may be able to walk around in the world when in the walking mode but when they mount a vehicle they no longer have access to the same actions as they did before, for example, they can no longer jump, but they might have gained some new actions, like changing gears.

3.2 What is game design?

Game design refers to the thought processes behind everything to do with game development. Unlike programming, which can be considered as engineering, and graphics or sound design and creation, which can be considered as art, game design can be seen as a craft. The reason behind this is that games consist of both functional and artistic elements: A game must be aesthetically pleasing but also functional and enjoyable to play in order to deliver an optimal experience. Game development is a joint effort involving both programmers (engineers), who take care of the functional side, and artists who create the outlook, so naturally the game designers must be able to handle both of those aspects to some degree in order to be able to design a game with them in place. That is why game design is not *just* engineering or art, it is both, it is a *craft*. [17]

The role of a game designer in the process of developing a game consists of the following aspects:

1. Imagining a game (into a design document).

2. Defining the way it works (core mechanics).

3. Describing the elements that make up the game, these include the game concept, functional, visual and audio aspects (the aesthetics, presented through the UI) among others like the story if the game has one.

4. Acting as an informant between the team members who are building the game and making sure that the vision of the game stays intact.

5. Refining and revising the design constantly during the development progress and testing until the game is completed (and sometimes afterwards as well).

Game design is often thought to be the most important aspect of game development, because without it there is no game, even if all the other aspects of the game would be of excellent quality. That is also why game designers are required to have a moderate amount of knowledge about almost all the other fields in game development, as they have to be able to understand all the possible problems that can occur when implementing their design for real, and be able to think around them.

3.3 Designing puzzles in virtual reality

Virtual reality offers a new way of exploring the game world and thus new ways of creating puzzles which can be solved by this exploration. The main aim in this thesis is to test mechanics which are not widely in use in virtual reality by having created problems for the player to solve using these mechanics. This section discusses the advantages of camera-movement in VR alongside with what kind of possible designs there currently exists for VR-movement. After that the chapter finishes with discussion about the already known design challenges that utilizing virtual reality creates.

3.3.1 Advantages and disadvantages of virtual reality's player-controlled camera

In virtual reality games, it is this far extremely common to use the first-person camera model with avatar-based interaction model, although the "avatar" is rarely visible apart from its hands as the avatar is the player themselves. The puzzles in this thesis also utilise the first-person model combined with avatar-based interaction model, in which the only parts of the avatar the player can see are its hand models. Reasoning for this is that it would be very difficult to realistically simulate the mechanics otherwise.

This brings us to one major aspect about virtual reality in comparison to traditional games: the camera movement. In traditional games, the developers have to always think carefully about what the restrictions for the camera movement are and due to this, which parts of the game world the player actually gets to see. In virtual reality, this problem does not exist in a similar way as the player can move the camera freely to whatever angle and direction they choose to by default. This in turn creates opportunities for developers with, for example, hiding information or objects in locations which in traditional games would be either unreachable or too clearly indicated to the player by the camera restrictions.

For example, if a player was told to search for a key in a room full of boxes which had holes in them, in traditional games the player would immediately know whether it is possible that a hole in a box could contain the key or not, just by testing whether they were able to enter it. In virtual reality, the problem is a lot less clear, as it is technically possible to go inside anything as long as the player's head fits through, creating almost endless amount of possibilities for hiding objects out of sight without giving away any hints of their whereabouts (the Seeking and finding puzzle version A, introduced in Section 4.4.1, explores this in practice).

On the other hand, the real-life connection between the player and the camera does come with its subset of problems as well, including heightened sensitivity to input lag and problems related to physical fitness in activities requiring rapid camera movements, which can also cause nausea.

3.3.2 Designing movement in virtual reality

Creating realistic movement inside a virtually created environment poses challenges for game developers, and one big reason for this is the fact that unlike the room where the playing is taking place, virtual space is technically endless. In order to attempt solving this problem, several different types of VR-movement have been created.

Room scale

The most basic and natural one is the room scale movement, which basically lets the player move around in the virtual space in the same way they are moving in real life. While this option definitely is realistic, it comes with problems such as limiting the explorable area inside the virtual environment drastically and the constant worry for the cable staying connected when trying to reach for the limits of the designated play area.

Teleportation

Another common way of movement is teleportation, which is often combined with the room scale movement in order to explore really wide areas without worrying about the physical space so much. The flaws of teleportation are reduced realism and possible nausea caused by rapid leaps in virtual space, which can be made less nauseating by dimming the player's view when a teleport takes place.

Locomotion

Locomotion means that the player can move, usually at around walking speed, inside the virtual environment via pressing or holding down some buttons in the controllers. This way of moving can feel more immersive than teleportation, as it can trick the user's brain to think that they are actually moving in real life as well, with a downside that it can often cause motion sickness.

Experimental

More experimental ways of moving in VR are for example *drag-world*, which allows the user to drag themselves to any direction by grabbing the space around them, almost mimicking the motion of skiing. Another example is a mechanic included in this thesis, swimming in VR, which is introduced in the Swimming puzzle in Section 4.2.

When it comes to physical devices which enhance VR-movement, various treadmills like *Virtuix Omni* [18], have been experimented with for simulating real-life running in virtual environments. With Virtuix Omni the users can convert real life workout to movement inside VR. The disadvantages of these kind of devices are the still quite steep prices and possible injury risks in cases of losing balance, although they often come with safety straps and/or barriers included.

3.3.3 Known design challenges in virtual reality development

In virtual reality, the player is "mapped to themselves", so to speak, meaning that their real-life actions mirror the actions of their character inside the VR, thus they *are* the character, as a comparison to traditional games, where the player *plays as* another character or entity. This means that the player's natural aspects, especially height and posture, are fully present with their character as well. This causes problems, for example, when the player is required to reach out or into something, as different players have different reaching capabilities based on their armlengths and general heights. The problem is unique to VR, as in traditional games the character's height and all other aspects are always set by the developers. For VR though, the player height is immutable, and rather than trying to change it (which would cause a major disconnection with the player and the game world), the games have to change to accomodate this "physical" challenge.

One way to deal with the innate player height problem is to make sure that the shortest possible players (of the intended target audience at least) are able to complete every required action. This was the design used when developing the puzzles Crawling and climbing and Seeking and finding (in sections 4.3 and 4.4), as both of those require reaching for something (climbable surfaces and keys), while the distance of the reach has to be possible for all player sizes. This, in turn, naturally causes the tallest of players to have a (slightly) less challenging experience, which could have an effect on their enjoyment level.

Another, more elegant solution, would be to create multiple different game versions, in which the height and the length of key objects would be scaled up or down based on the player. This option was not used in this thesis because it would necessarily mean that not all testers would be solving the exact same puzzles, causing it to be another factor reducing the reliability of the results. Additionally, measuring the exact height and reach values of the testers during the testing would be both time-consuming and impractical.

On the subject of "physical" design challenges, virtual reality is infamous to cause motion sickness, and because of that the design of VR-games (and other applications) have to take that into account in order to minimize it or otherwise the games can become unplayable just due to that. The main cause for the motion sickness is generated by the user's brain not being able to process two mixed signals: one sent by the eyes that the user is moving, and another sent by the vestibular system (in the inner ear), which is responsible for the sensation we like to refer to as balance, that we are definitely not moving. Some design solutions for this challenge are avoiding moving the camera without initiation by the player, minimizing the input lag, and blurring the player's field of view when necessary, for example, during teleportation transitions (this is utilised in the teleportation used in the Swimming puzzle version B, introduced in Section 4.2.2). Additionally, it is possible to ease the confusion caused to the player's brain by visualising the movement in the air using "speed stripes", which the brain might understand as the environment being the one which is moving, not the player.

Another "physical" challenge is related to the actual physical constraints, because as long as the headset is still attached with a cable and the room space remains quite limited, the experience is going to suffer because of cable or the surrounding environment (walls, furniture) stopping the player from moving or reaching out. This challenge is a very difficult one for game developers to solve and is more in the area of expertise of VR-hardware developers, but the effects of this can still be attempted to be mitigated a little. The easiest way to do this is naturally by creating games which require no moving around horizontally, major rotation or reaching out. An example of this kind of design is Beat Saber [19], which is a VR rhythm game in which the player does not need to move more than a step or two sideways during the gameplay in order to have the full experience, although the game has a 360-degree mode which does suffer from these problems. Currently, there are few better solutions to the problem, other than recommending the use of something like a TPCAST [14] for a wireless experience, which is a problem as VR-equipment is already so expensive that very few players are willing to invest on additional VR-enhancement add-ons as well.

Other than these three quite obvious "physical" problems, there are many other types of "mental" problems which VR-developers need to consider.

The first "mental" problem that often causes serious issues is: what to do in the event of the player pushing their real-life head and/or body through a virtual wall? In this case doing nothing is probably the worst option, as this leads to players being able to walk through walls making the entire game world kind of pointless, as it cannot restrict the player at all. One way to answer this problem is by having only room scale movement enabled, with a large open space as the virtual play area, making it impossible for the user to ever reach a single virtual wall in the game. However, as mentioned in Section 3.3.2, this limits the game scale to a very small room size, and is thus not the most elegant solution. Another, maybe the most commonly used solution, is to teleport the player a small distance away from the entered wall in the virtual environment, making it impossible to ever get through it in theory. While this solution can be effective, it still comes with problems such as causing motion sickness with the sudden uninitiated teleportation, and it can also be immersion-breaking.

Unfortunately, apart from these two methods, there has not been any groundbreaking progress in solving this problem at the time of writing this thesis. One possible future technology that could solve this could be actually restricting the player's physical movements with the headset when a collision occurs, maybe by altering the player's sense of touch so they could actually "feel" the virtual surfaces? Although in that case the device would be more of a "neurogear" (like Neuralink [20]) than a VR-headset, and in any case that future seems quite distant as of yet, although technically anything is possible. The puzzles in this thesis utilise the wall teleport (rewind) mechanic as the puzzles require the player to move around quite a lot in restricted space, thus making the option of never getting close to a wall not possible.

The second problem is related to interacting with objects inside the virtual reality and is basically as follows: how should objects be grabbed or touched in VR without losing realism or breaking the game? There are only two possible ways of grabbing an object in VR: either the object is grabbed from the position it is touched from, or it is grabbed from a specific point (hinge) in the object regardless of where it was originally touched. The first option preserves realism, while the second usually results in less errors and awkward object positioning. By using the first option, developers have to be careful with key items, as using them the wrong side up or down can cause really unrealistic interaction. It can also make simulating realistic physics (mainly weight and friction) more difficult, as it would be difficult to differentiate the relationship(s) between an object's grabbing location, its center of mass and the parts of the hand(s) currently touching it. The second option then again can cause problems with interacted objects going inside walls or other objects, as the position of the object is fixed to a specific location in relation to the controller (hand) which picks it up, ignoring the surroundings. It is also worth considering what happens if the player tries to put their hand through a solid surface while holding an object, although it can be relatively easily "fixed" by just force-dropping the handled object in this scenario.

The third problem is related to the second, specifically about interaction with objects which need to move slower than the hand that is interacting with them. In traditional games, opening doors slowly is not a problem at all, as in most (if not all) games the player does not have perfect independent control over their arms and hands at all times, but rather usually toggles starting an interaction with a door or a switch while the interaction itself happens with an animation. In virtual reality, it would feel quite strange to the player to see their hands being animated on their own while just standing and watching, as they are perceived to be their actual real-life arms as well. This is why creating friction to opening doors or pulling levers poses a challenge, as the movement needs to be done by the player and every player has their own speed and method of operating objects such as doors and levers. It is of course possible to have the rotators follow the player's hand movements precisely, but this requires giving up friction completely, making the rotators function very unrealistically, as, for example, heavy iron doors should not be able to be rotated with one small wrist movement. Unfortunately, the only other realistic option is basically slowing the rotator down at the cost of ending the interaction in case the player's hand is no longer touching the rotator. This second option forces the player to move their hand more slowly in order to be able to interact at all, but it does not really create the feeling of friction too well, as in real life heavy objects need more force, not less.

4 Practical application

For the purpose of this thesis, three different VR puzzles, all of which have 2 variations, have been created and tested. The movement mechanics in the puzzles have been partly developed by modifying pre-existing mechanics from a piece of open source software called the Virtual Reality Toolkit 3.3.0 (VRTK), which is a non-profit project aiming to aid VR-developers around the world getting started in their development, hosted on Github ([21]). This chapter consists of first presenting the basic functionalities, namely the player and the tracking system, followed by the three puzzles, their mechanics and variations. The testing results are displayed and analysed later in Chapter 6.

4.1 Basic functionalities

Let us begin by introducing the player's composition and the interactive parts, which are the headset and the controllers, and some information about the detection mechanisms in place.

As aforementioned, in these puzzles the VR-device used was the HTC Vive, which in total consists of two base stations utilizing the Lighthouse-technology, which detect the player's headset and controllers using infrared light, two controllers, and the actual headset. The controller can be seen in detail in Figure 9 with all three buttons necessary to complete the puzzles in this thesis named.

The player

The player is formed by three different main colliders: head, body, and feet. Head collider's position is calculated based on the real location of the headset, measured via physical "lighthouses" which track the VR-headset and controllers, while the body collider is created based on the location of the head collider and the feet collider based on the position of the head and floor area. Additionally, the VR-controllers act as the player's hands, positions of which are tracked by the base stations. The hands have separate colliders for each finger and their collisions with objects are used to interact with most things in the puzzle levels.

Figure 9: A HTC Vive controller showcasing all the buttons used in the puzzles.

The colliders and their dependencies are all created by the VRTK and displayed in Figure 10.

4.2 Puzzle 1: Swimming in VR

This puzzle aims to simulate swimming in VR. The goal for the player in this puzzle is to find a hidden path underwater and eventually locate a key object at the end of the path. Completing the puzzle also involves climbing a swingy rope after emerging from the second water area. The general mechanics include reduced gravity (while in the water) in order to simulate the buoyancy of water and a limited supply of oxygen for the player, which can lead to in-game drowning. The player is guided by a red headlight attached to their imagined forehead in addition to lanterns located in various parts of the puzzle, which light up when approached by the player, making it easier to locate themselves in the otherwise quite dark water, while simultaneously creating a sense of progression. Additionally, the water areas contain bubbles which are moving upwards to help players retain a sense of movement (and direction if they somehow would lose it while standing all the time). The puzzle overview and a few close-ups can be seen in figures 11, 12, and 13.

The puzzle has two variations: the more realistic version A, and the less realistic version B, which utilise different swimming mechanics, explained in more detail later in this chapter. The puzzles themselves are otherwise almost identical, although the hidden path that the player needs to find is altered slightly so that both of the puzzles actually take some effort to solve, and the player cannot just copy the route from the version they test first.

A special interest in this puzzle is to see whether the player's ability to swim in real life has any effect on their enjoyability of the experience and whether players with poor swimming skills are more afraid of in-game drowning than those who rate their swimming skills above average.

This puzzle has the following two different versions which change the swimming mechanic radically for the player.

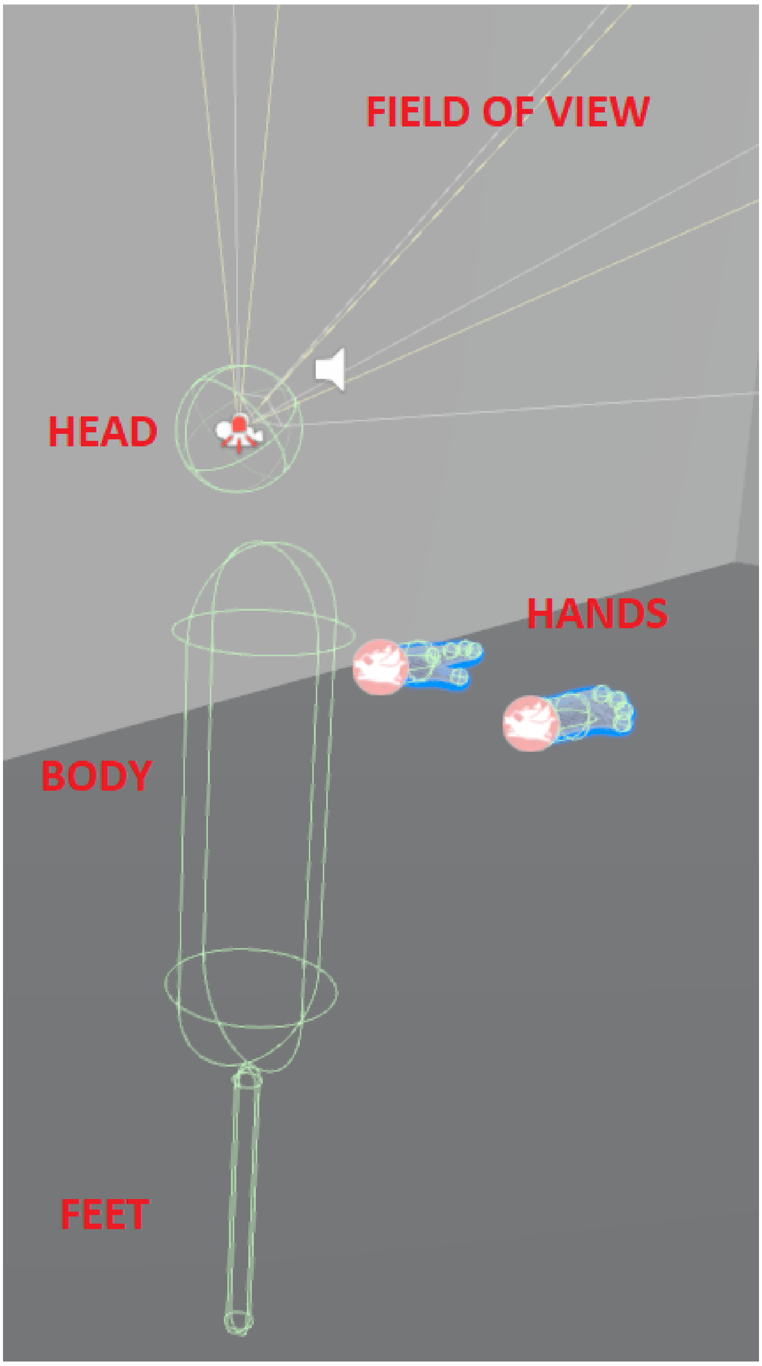

Figure 10: Player's collider composition.

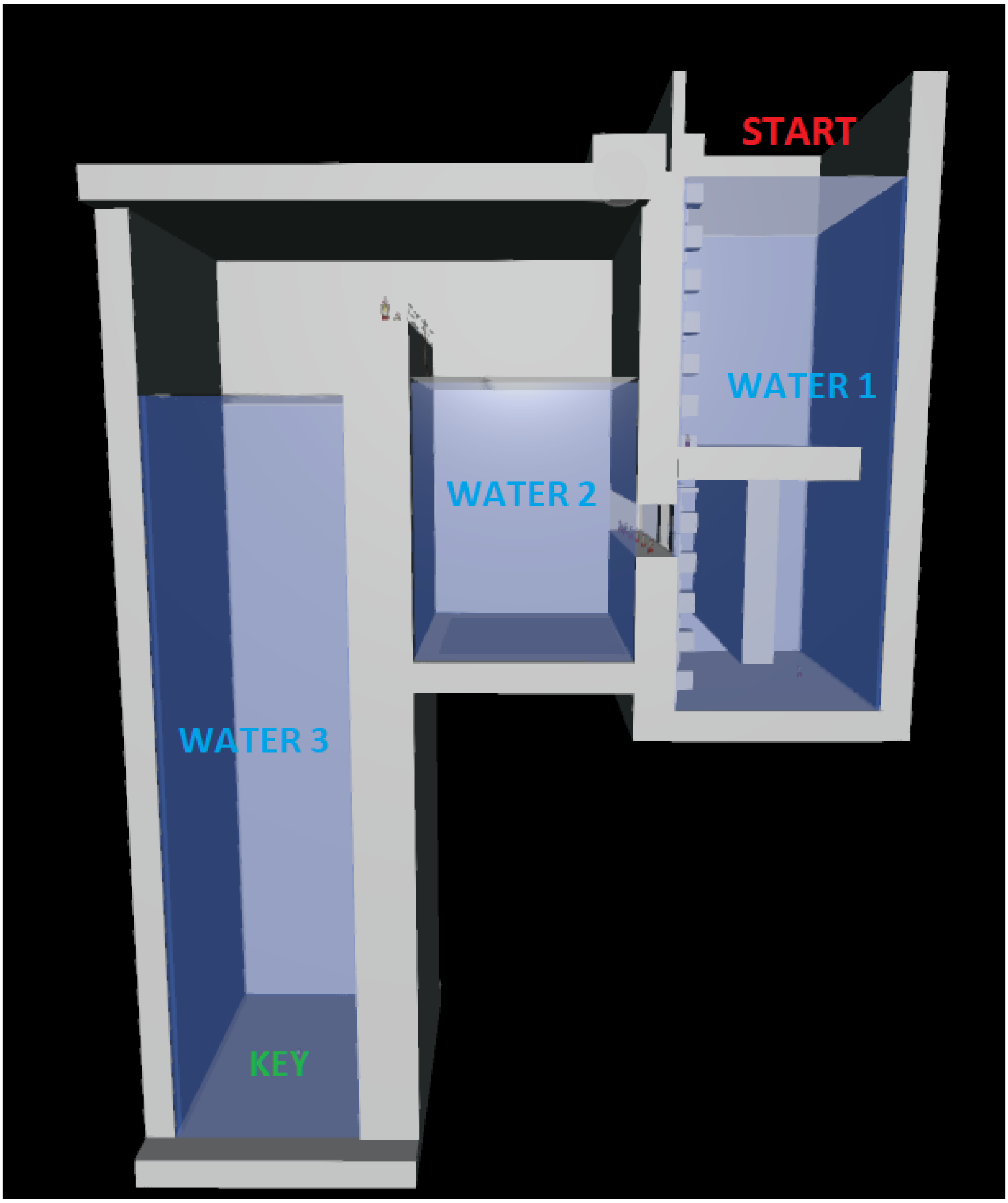

Figure 11: The overview of the Swimming puzzle.

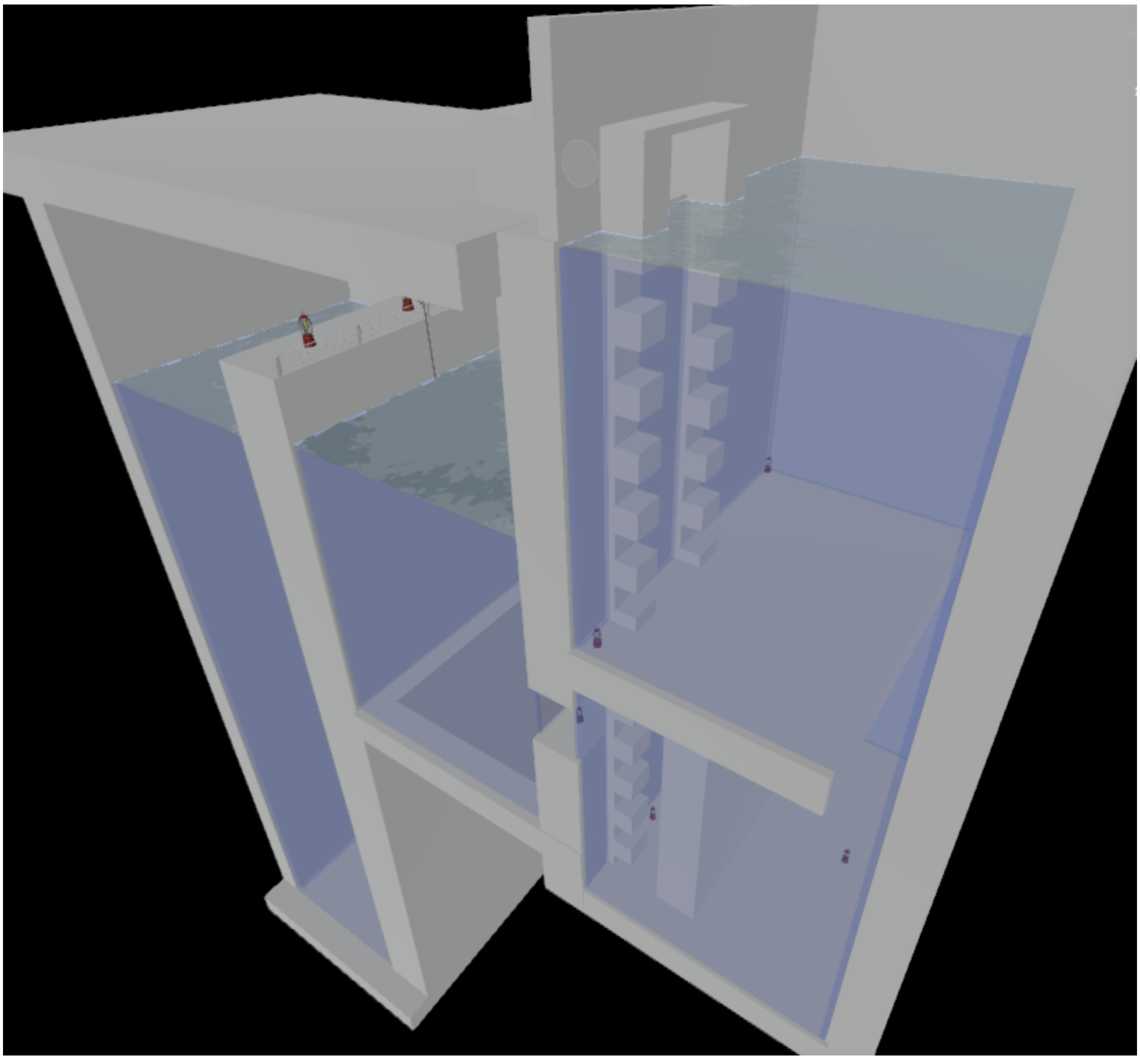

Figure 12: A close-up of the Swimming puzzle.

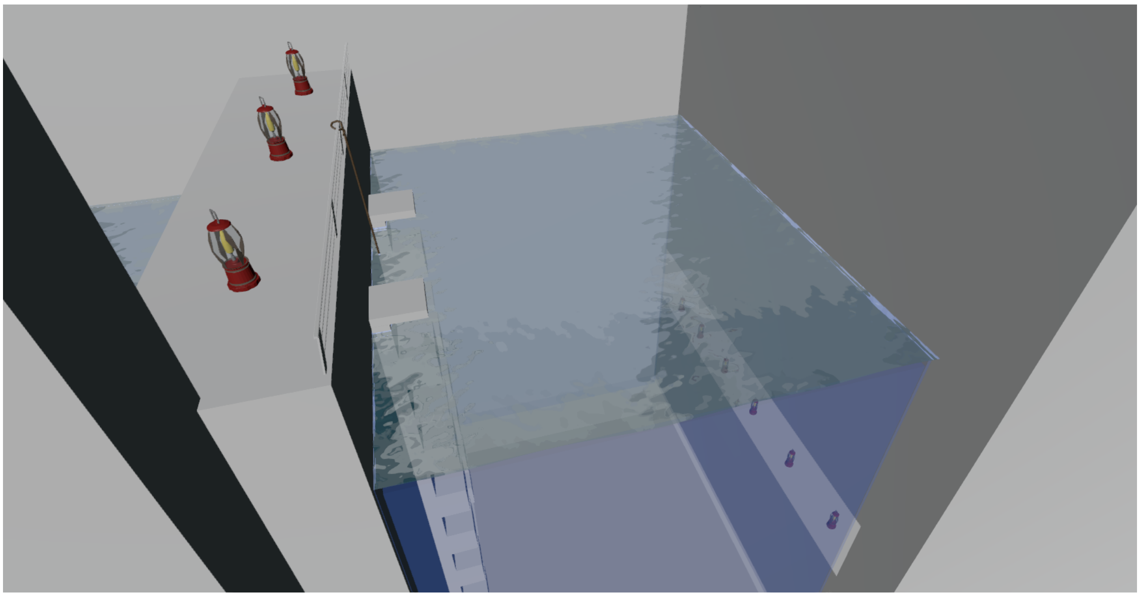

Figure 13: The rope climbing section in the Swimming puzzle.

4.2.1 Swimming puzzle version A (More realistic)

In this version the swimming mechanic requires the player to make swimming-like motion with their hands in order to be able to move in the water. Locomotion (see Section 3.3.2) is enabled in order to move sideways more easily and on the ground but it cannot be utilised when trying to swim upwards, forcing the player to actually physically “swim”.

The way that the swimming actually works in this puzzle is that the entire water area is treated as a climbable surface, and the player can grab hold onto it in order to pull themselves towards a direction. In order to make the pulling feel like swimming, the core mechanics automatically release the player’s grip of the water after about a half a second, forcing them to grab on it again, creating forced swimming movement.

The difference between version B, apart from the swimming mechanic, is that the beginning part of the hidden underwater path (showcased in Figure 14) is slightly different, and also slightly easier to find by design. The design choice to make the path easier to find was made because the more realistic swimming mechanic already poses more challenges for the player as far as exploration is concerned.

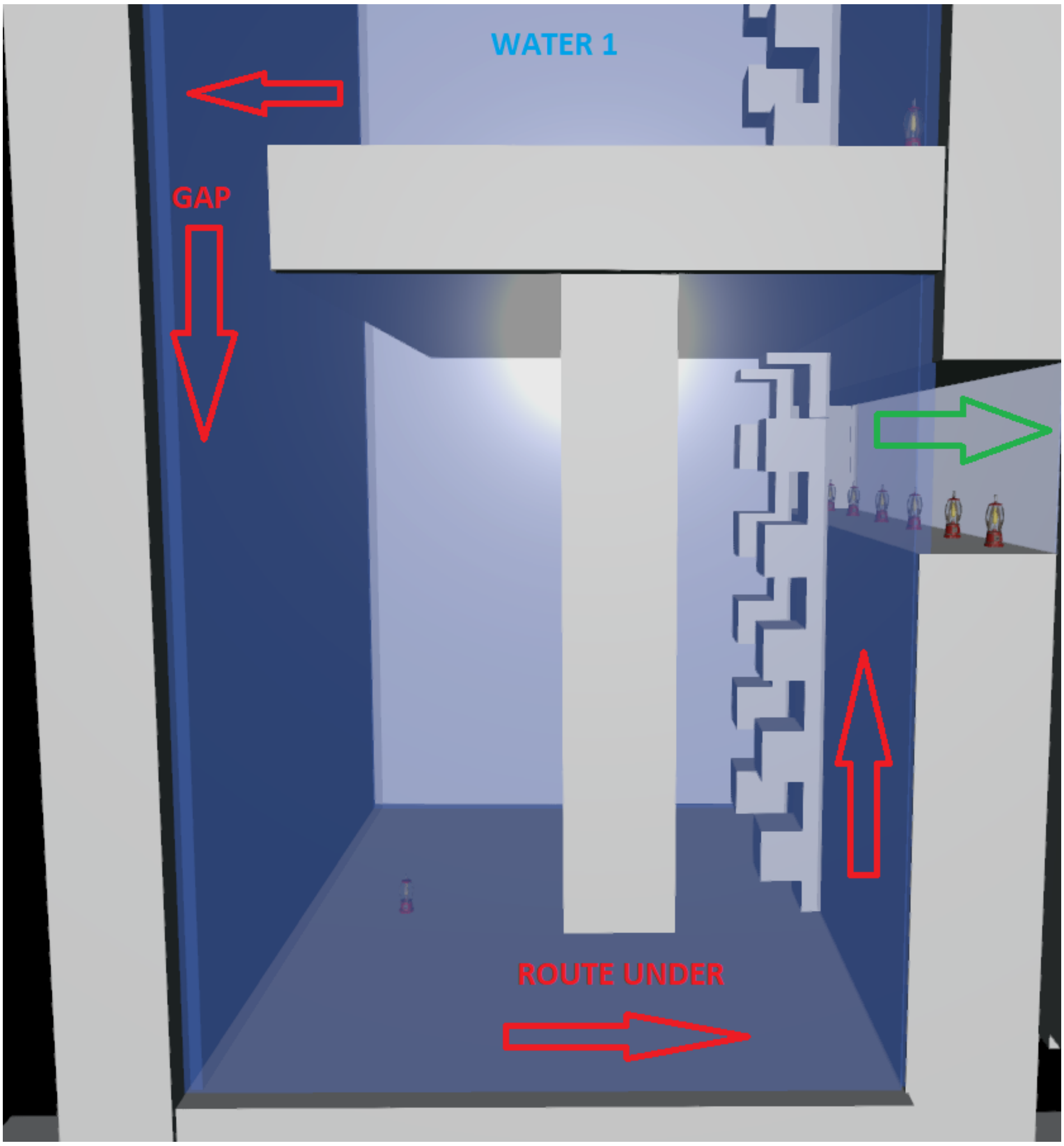

Figure 14: The hidden underwater route from Water 1 to Water 2 in Swimming version A.

4.2.2 Swimming puzzle version B (Less realistic)

In this version the swimming is more automated, meaning that in order for the player to rise to the surface, physical movement is no longer required. In detail, the player can press a button which will start lifting them towards the surface above them, and if reached will allow them to float on top of the surface and move around until another button is pressed. Teleporting is also enabled on horizontal surfaces nearby, making traversing and exploring the underwater floor space significantly faster than in version A. Otherwise the player moves via locomotion similarly to version A.

Because of the easier exploration mechanics, the puzzle itself is slightly more difficult, as the beginning part of the hidden path is less visible and thus harder to find than in version A, seen in Figure 15.

4.3 Puzzle 2: Crawling and climbing

This puzzle aims to experiment with two different ways of how crawling and climbing could be handled in VR. The puzzle consists of alternating tunnels and walls, with three tunnels to choose from each time as seen in Figure 16. Every tunnel has a very low ceiling, forcing the player to duck (crawl) in order to get through, and is labelled with a keyword, which helps the player define their path of choices. Entering a tunnel also produces a distinct sound effect, which relates to the next keyword and the tunnel they should pick.

In order to avoid making the puzzle too complex and long, there are only three choices to be made, creating a total of 27 different possible paths (technically), although the first choice is made to be always correct, so there are three different possible correct paths and practically only 9 actual paths to choose out from. The number of paths is so low also because testing one path takes quite a lot of time and effort, especially for the more realistic crawling and climbing mechanics, and the testers might run out of energy if there were more possible paths.

After the final choice is made, the game will either end with the player

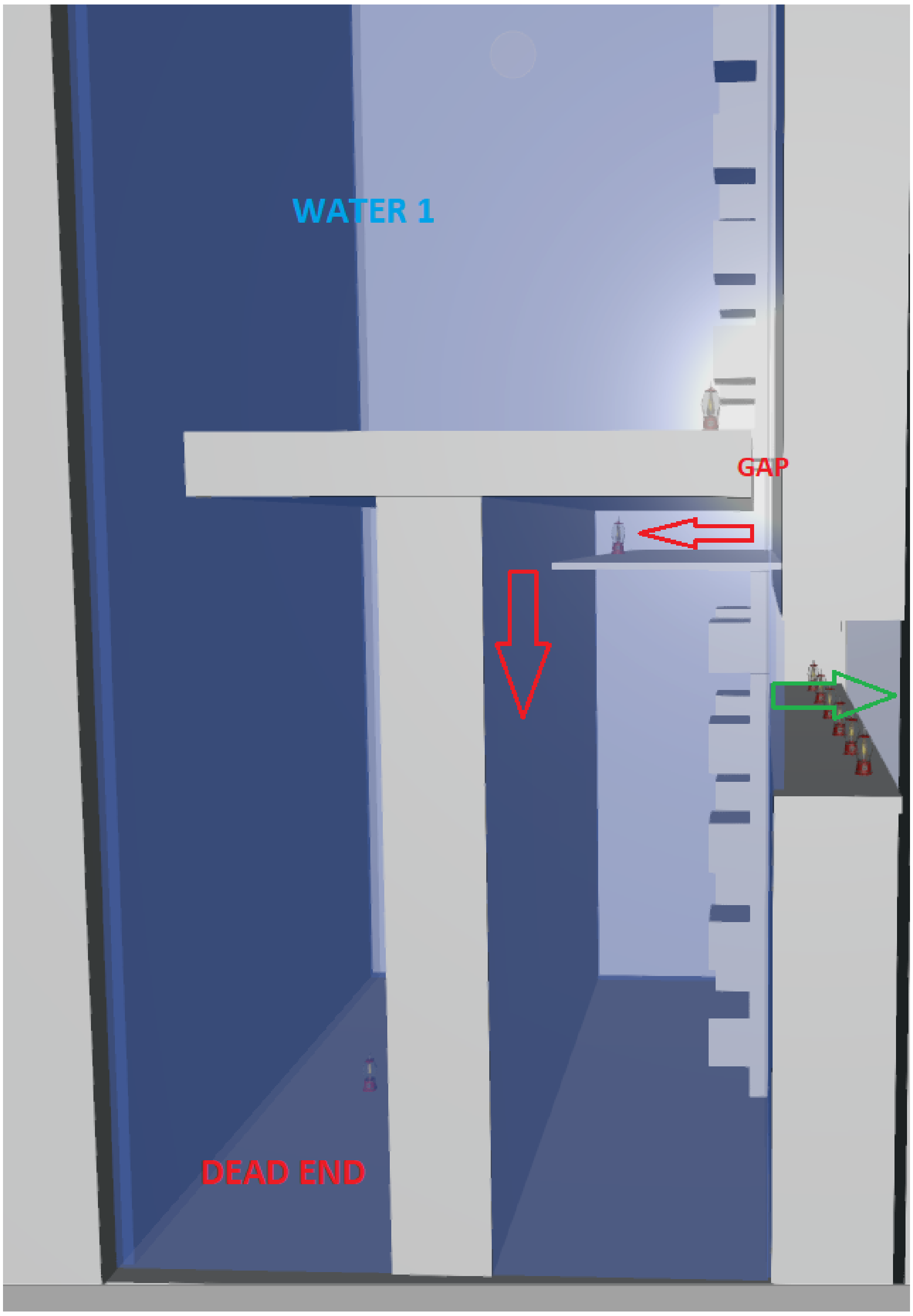

Figure 15: The hidden underwater route from Water 1 to Water 2 in Swimming version B.

Figure 16: One of the path choices the player has to make in the Crawling and climbing puzzle.

emerging victorious, or they will be sent back to the beginning to redo the puzzle. The average number of setbacks will define the perceived difficulty of the puzzle and will be used to determine whether it was too difficult or too easy.

There is also a part in the puzzle which attempts to simulate climbing in a very tight space (showcased in Figure 17), the goal of which is to see if the perceived tightness of space will trigger any claustrophobic reactions.

The differences between the two versions, A the more realistic and B the less realistic, apart from the movement mechanic is that the keywords and sounds and thus the paths they form are completely different.

The main idea with this puzzle was to test especially crawling in VR, as it is a fairly rarely used mechanic because of the challenges it poses, which include:

- how to force the player to crawl in real life, and
- what to do when the player gets up while inside a tunnel.

Overview of the puzzle layout can be seen in Figure 18.

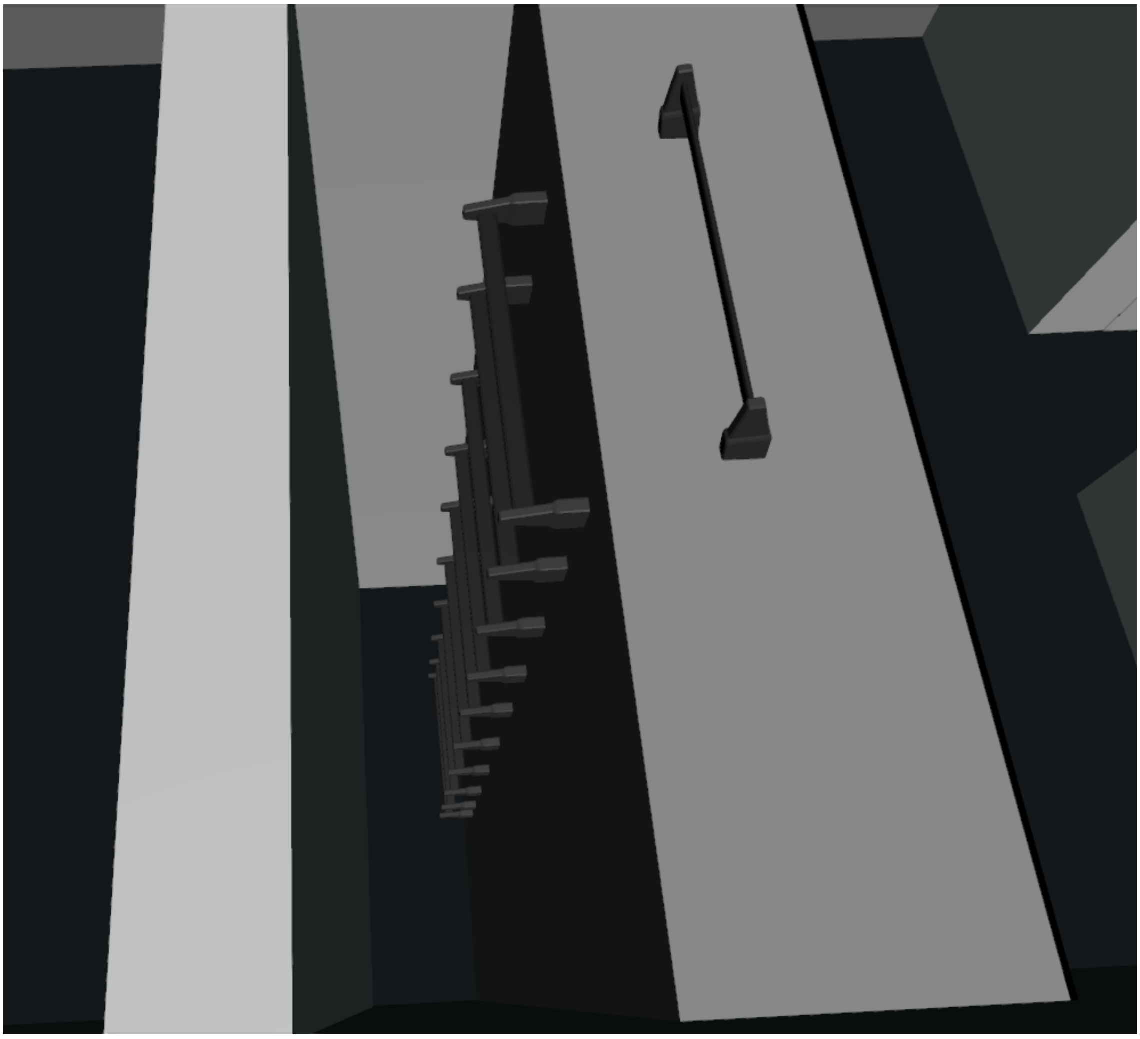

Figure 17: The tight climb part of the Crawling and climbing puzzle.

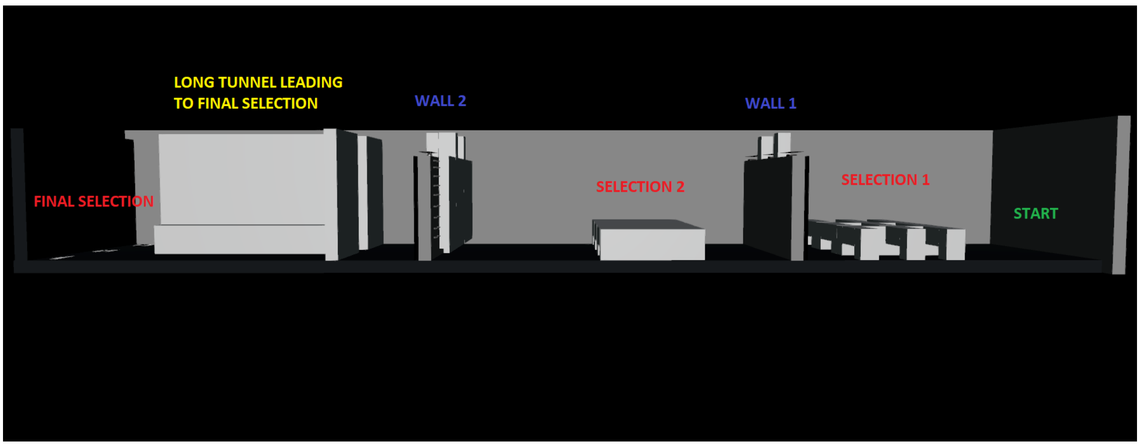

Figure 18: Sideways view of the Crawling and climbing puzzle, highlighting the start area (on the right), path selection locations (in red) and walls.

4.3.1 Crawling and climbing A (More realistic)

In the more realistic version of the puzzle, the whole level is filled with climbable squares (as seen in Figure 19), which are the only way the player can move around apart from room-scale movement, which will not be able to get the player very far. These squares are on the ground level when the player needs to move around on the floor and on the wall when there is a section where the player needs to scale a wall.

The reasoning behind forcing the player to use these squares in order to move, is to try to force the player to crawl in real life while going through the low-ceiling tunnels. Additionally, in order to stop the player from throwing themselves forward, this mechanic also prevents the player from releasing the climbable squares, except when grabbing a hold of another one with their other hand. The vertical distance between the player's headset and the climbable surface on the floor which is currently being held on to is also calculated, and if the distance gets too long, the player will automatically ungrab the surface, in an attempt to prevent unintended ways of movement.

Climbing upwards is quite standard in this version, meaning that the player needs to hold the grab-button down in order to stay attached to a climbable surface. Additionally, the action of pulling oneself on top of a climbable surface is completely manual, requiring the player to imagine their virtual legs as well.

In a case where the player would try to stand up while going through a crawling section, they will be teleported back to the start of their current section, which, while not being the most elegant solution, is quite effective at preventing players from ignoring the low ceilings.

4.3.2 Crawling and climbing B (Less realistic)

In this version, the climbable squares (and sometimes ladder) only exist when climbing the walls up. On other parts of the puzzle, locomotion is used to move (crawl) around, making it possible to duck down and press forward button in order to go through the tunnels. The climbing

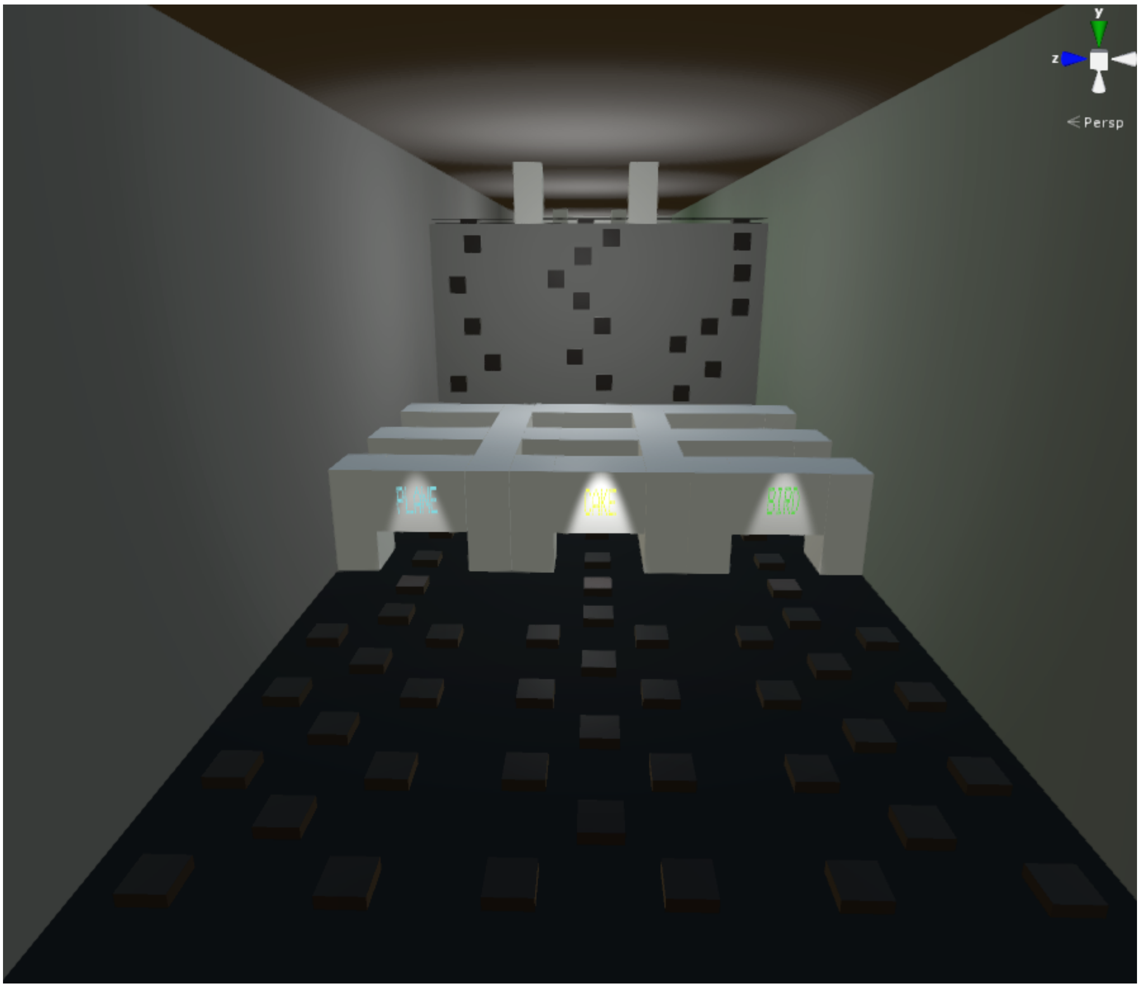

Figure 19: Front view of the Crawling and climbing puzzle version A starting area, highlighting the climbable squares on the floor used for movement.

mechanic also differs in a way that the player is not required to hold down the grab-button while holding on to a climbable surface in order to stay attached, but it rather functions as a toggle. The player does need to pull themselves on top of climbable surfaces manually in this version as well, although the toggle mechanic should make it more difficult for them to fall down while attempting it.

Apart from the mechanical differences, the keywords and sounds are also slightly more difficult to associate with each other by design than in version A, because the movement mechanic offers a faster way to trial and error. Otherwise the layout of the level is exactly the same as in version A.

4.4 Puzzle 3: Seeking and finding

This puzzle tests how difficult locations should objects be hidden in puzzle games and how VR affects the ease of designing these types of puzzles. Basically the puzzle consists of a room which has some random looking block-structures (which are assigned a random colour on runtime) in it, and somewhere among those structures there are two key items hidden, which the player must find. Both of the puzzles also contain hints about the whereabouts of those keys for the player to find, scattered around the rooms.

The main driving force behind this puzzle idea was the fact that as in virtual reality the camera movements do not need to be designed into the gameplay, but rather exist automatically in the player's ability to move their head to any direction and angle they desire (as explained in Section 3.3.1), there is no need to indicate whether entering objects is possible or not to the player, as they simply have to test everything. This creates possibilities for hiding objects within objects without giving any hints about their whereabouts in the form of game mechanics. As a contrast to the previous, for instance, in a first-person non-VR game the player would instantly know whether it would be possible that there are things hidden under a table or not based on would they be able to crouch.

In other words, virtual reality has (basically) complete freedom of movement, as it reflects the player's real-life movements as input. This puzzle aims to figure out whether it is enjoyable to hide things in places where they could be hidden to in the real world (more realistic), but hiding them into a (non-VR) game world in the same way would be either very difficult or impossible altogether.

There are two different versions of the puzzle, one of which the hiding places are ultimately only designed to work in VR conditions, and the other where they are hidden to places which could also be used as hideouts in more standardized games. This means that both of the versions could be considered "realistic", meaning that both kind of hideouts could exist in the real world, so the sensation of realism is not something that can be measured so easily unlike in the other two puzzles. That is why the main focus of the research for this puzzle is related to various levels of enjoyment, and the version which turns out to be more enjoyable can reflect whether the more or less "realistic" hiding mechanic is more enjoyable than the other as well. Additionally, the locomotion movement is enabled for this puzzle and all the objects in the scenes are climbable for maximum exploration possibilities.

4.4.1 Seeking and finding puzzle A (More realistic)

In this version of the puzzle, the two key items are hidden in places which require going inside other objects in the scene. This variation is an attempt to create a near completely explorable area by utilising the advantages of VR. The main challenges related to the design were making sure that the player cannot enter any objects in other ways than the designated ones, while making the designated entrance points function as smoothly as possibly. The overview of the entire level can be seen in Figure 20.

The key in key location 1, illustrated in Figure 21, is located in a hole on top of an object which otherwise has a smooth appearing surface, making it very difficult for the player to actually see the hole until they have climbed on top of the object (or found a hint related to the key). Additionally, the hideout features "going under the floor"-mechanic, meaning that the player actually needs to pull themselves "through" the physical

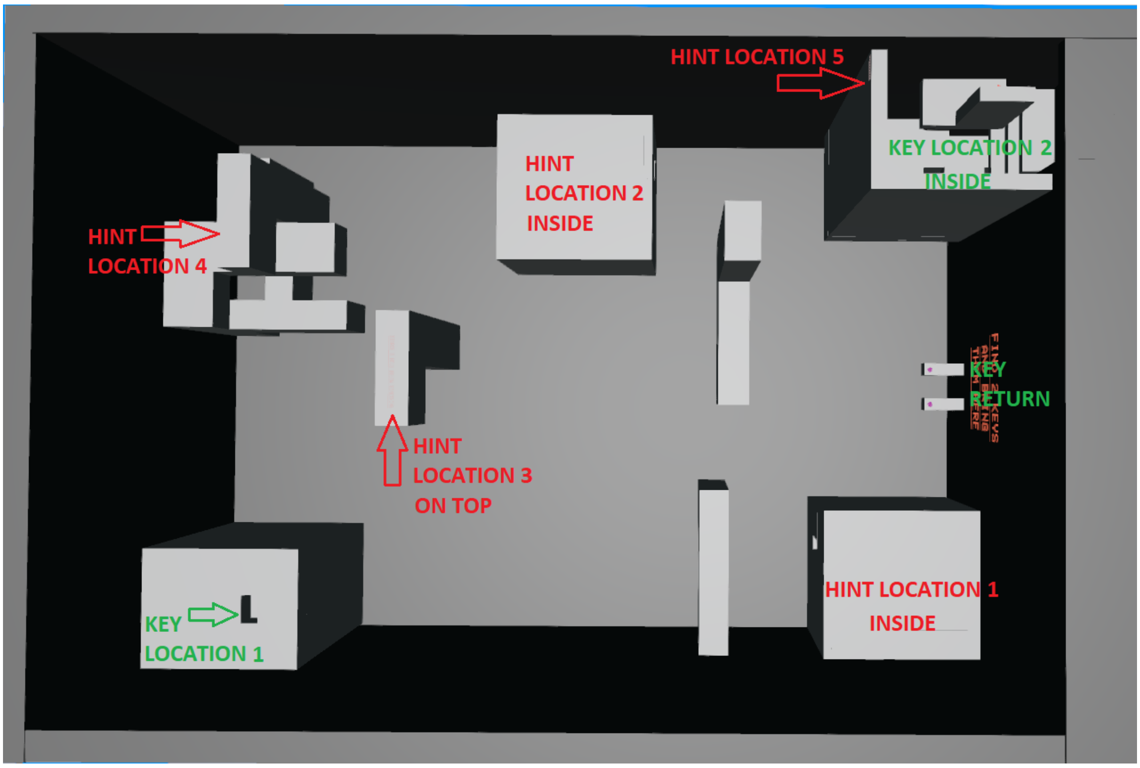

Figure 20: The overview of Seeking and finding puzzle A, highlighting the locations of both of the keys, hints, and the area where the keys need to be taken to when found.

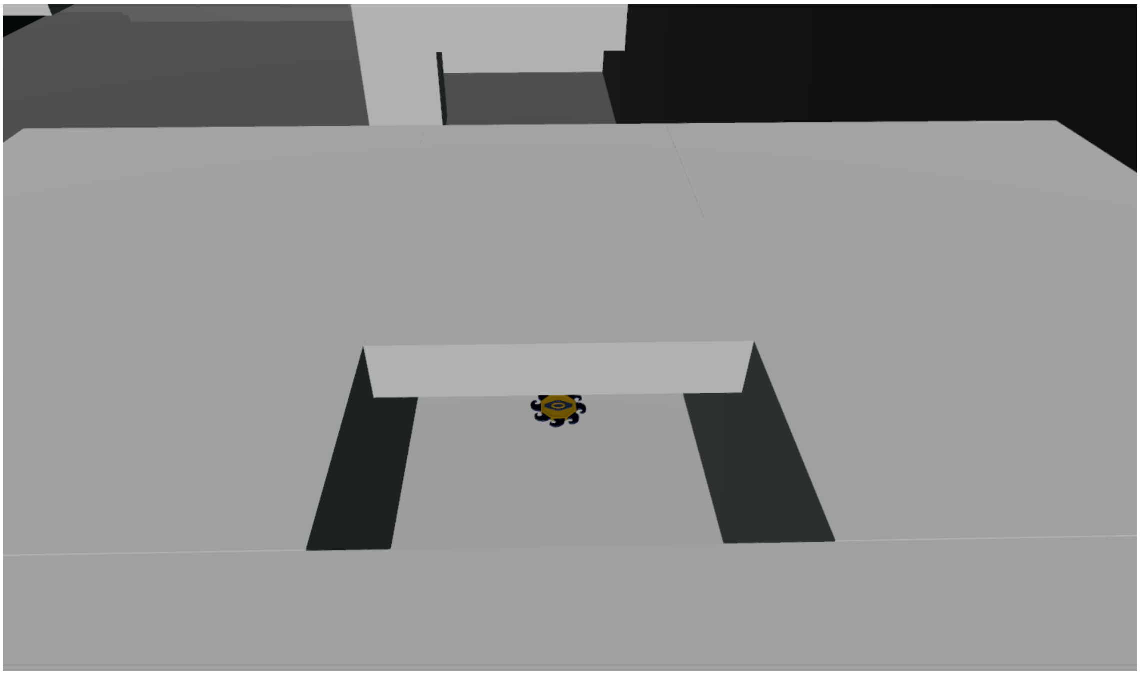

Figure 21: The hole which is on top of the object the player needs to scale in order to find the key 1 (the disc) in Seeking and finding puzzle A.

floor under them in-game, in order to reach the key. In more detail, this mechanic functions so that the player needs to be standing on top of the surface with the hole in it, and then grab one of the sides of the hole and pull their head and hands through the hole, which to them appears to be under the floor they are standing on in real life. An additional requirement for this to work is that the hole (in Figure 21), has to be so small that the player cannot fall into it.

The location containing the key 2 has an entrance located near the floor area of the level, seen in in figures 22 and 23, and it is designed to be much more difficult to find than key 1, as after managing to enter the key location, the player needs to (ideally) twist their head to the right angle in order to be able to see the key.

Additionally, the hideout is designed to attract the player to search the upper areas, which contain nothing more than three question marks in an attempt to taunt the player, while the key itself is hidden near the entrance (highlighted from two different angles in figures 24 and 25). This mechanic is especially something that would be very difficult to perform with standard camera movements, or at least all the possible

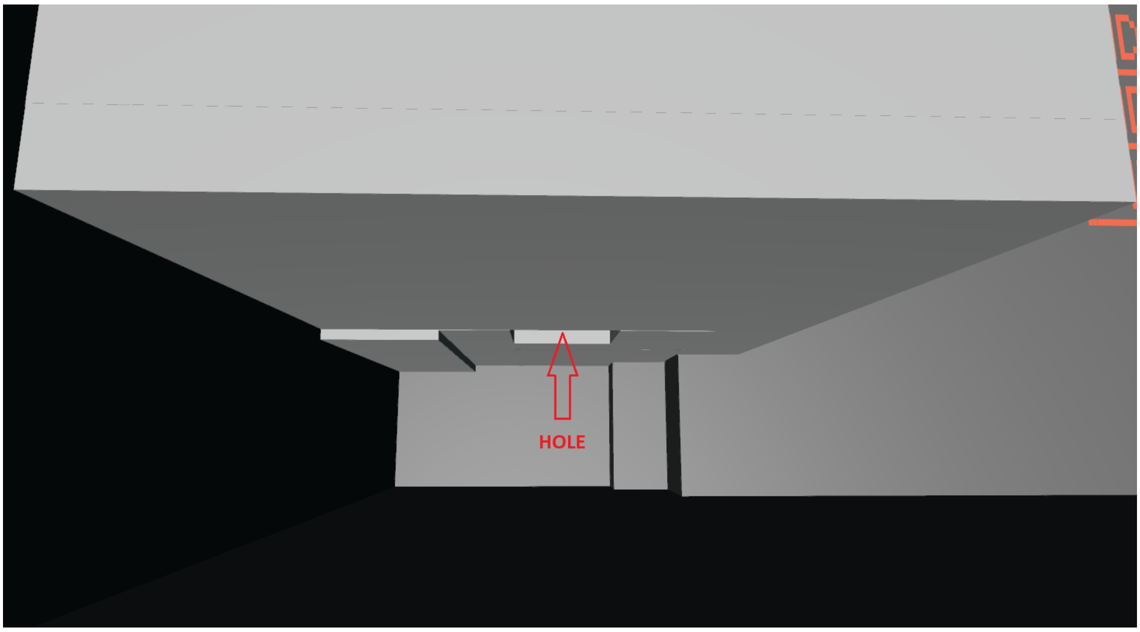

Figure 22: The hole which player needs to enter in order to reach the insides of the key location 2 in Seeking and finding puzzle version A.

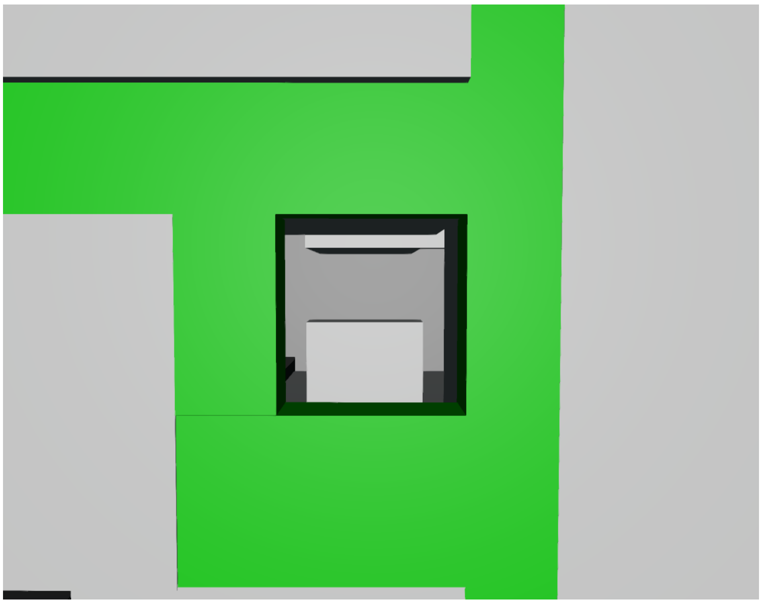

Figure 23: The hole of Figure 22 as a close-up. The green areas will not teleport the player back even if their headset would go through them, in order to make the entering process smoother. Similar designs were used in most of the holes in the puzzle.

locations for the key item would be very obvious to the player in that case.

Figure 24: Key location 2 (version A) in detail from the left side. The ???-part contains three question marks, located at the highest point of the hideout, purpose of which is to taunt the player to figure out that the key is actually below them near the hideout entrance.

4.4.2 Seeking and finding puzzle B (Less realistic)

In this "less realistic" version, the key items are placed in hideouts which would not be too difficult to design for non-VR games either. Another one of them requires a little bit of climbing while the other one needs

Figure 25: Key location 2 (version A) in detail from the right side. The ???-part contains three question marks, located at the highest point of the hideout, purpose of which is to taunt the player to figure out that the key is actually below them near the hideout entrance.

careful exploration to be found. The main idea is to offer a contrast for the A version and to see whether the decrease in difficulty increases enjoyability or not. The only actual "mechanical" difference is that, the version B uses the toggle-method for climbing the objects in the room, similarly to Crawling and climbing puzzle version B (in Section 4.3.2).

Making strict comparisons between the two versions can also be difficult as their room layouts are different, although the size of the rooms is exactly the same. Hopefully the results will show some kind of enjoyability difference which then can be understood to correspond to the level difficulty, which in turn relates to the hiding mechanic. The overview of the version B can be seen in Figure 26.

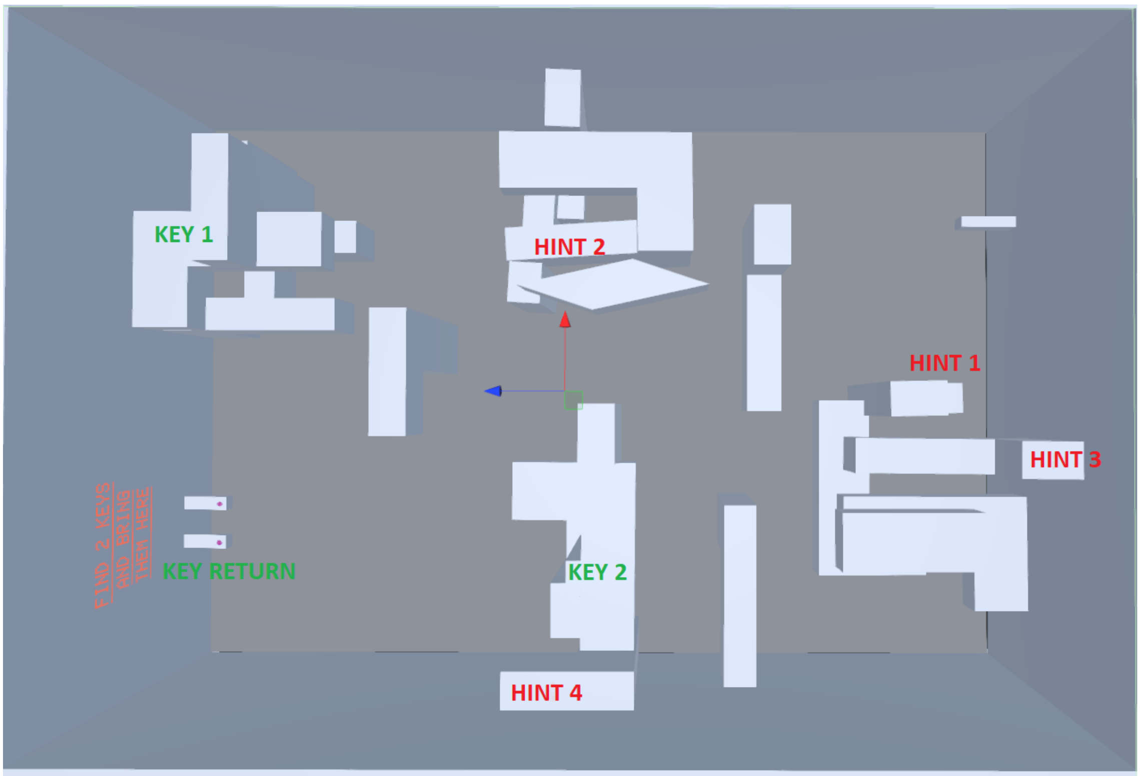

Figure 26: The overview of the Seeking and finding puzzle version B, complete with locations for both keys, hints, and the key return place.

5 Testing methodology

The testing utilised convenience sampling (using testers who are the easiest to access) based on volunteers, which were acquired either through the social channels of University of Turku or by recruiting already familiar people. Additionally Turku Game Hub (HIVE) provided some volunteer testers who had extensive knowledge about games and game or virtual reality development so that the sample would also contain some experts for comparison reasons.

The test subjects were randomly divided into batches of four people where each of the batches would test the puzzles in a different order, in an attempt to minimize the effects of both VR sickness and familiarity affecting the final testing result. Additionally, two testers from each batch were to have exactly the same version order, meaning that they would test either the more or less realistic puzzle version first for all of the three puzzles. Due to lack of volunteers, one of the batches contained only two testers, leaving the version order of that testing order without comparisons.

These volunteers all tested separately at some point during the testing period of around one month, which took place in November-December of 2019. The testing was concluded with a HTC Vive VR-headset (shown in Figure 6) with a personal computer running Windows 10 operating system, Nvidia Geforce GTX 980M graphics card, Intel i7 processor and 8 GB of RAM. The testing environment was a quiet room with around 3×3m sized play area. The testing consisted of first reading written instructions, which explained the goals, fail conditions and inputs of the different versions to the testers, followed by the actual testing and finally filling out the testing questionnaire. The puzzle instructions can be viewed in Appendix A, while the questionnaire will be explained here next.

The testing questionnaire consisted of two parts: the first one (pre-questionnaire) which was filled before testing, containing questions related to the testers, demographics, personal preferences, and abilities, while the second one (post-questionnaire) was filled after testing and contained questions only related to the testing itself. The questionnaire used in the testing mostly utilised Likert scale ([22]) with four response options for questions, which are considered to have equal intervals:

1 = Completely or nearly completely disagree
2 = Disagree to some degree
3 = Agree to some degree
4 = Completely or nearly completely agree

Most of the questions in the questionnaire which required a response on the Likert scale were in the form of "How much would you agree with the following statement" followed by the statement. The idea behind leaving out the "Neither agree or disagree"-option was to force the testers to choose their preference (no matter how slight), while leaving the other options less strict with clauses "to some degree" and "or nearly" was designed to make the choice easier to make. This was done in an attempt to avoid the *central tendency bias* (explained in Section 6.1.1) occurring too much, which would discourage testers from choosing the extreme options (1 and 4).

Further sections in this chapter discuss the picking criteria of the testers in more detail and list the results of the pre-questionnaire, followed by the presentation of the testing hypotheses.

5.1 Test subjects

Although technically anyone was able to join the testing, most testers were members of one of the local universities in Turku, UTU (University of Turku) or TUAS (Turku University of Applied Sciences) and even more frequently the testers had some kind of academic background, a lot due to the convenience sampling method. This means that the results are biased to mostly only apply to groups of people who are or have been in higher education, and especially in the Turku region. Other biases are being a native Finn (20/22 testers) and a male (19/22 testers).

The testers were originally designed to be selected based on exclusion and inclusion criteria, but this idea had to be forfeited due to the difficulty of finding enough volunteers, resulting into the study accepting all available volunteers. The original criteria were as follows:

Exclusion criteria: People with any kind of physical disabilities, the elderly, people younger than 18 years old.

Inclusion criteria: Preferably healthy young adults between 18 and 35, people with previous VR experience.

As far as the exclusion criteria are considered, the voluntary sample fits it perfectly: no tester was a minor, no one had permanent physical disabilities and all testers were either young or middle-aged adults. The inclusion criteria, one of which was to only include testers between 18 and 35 so that the age gap would not cause such a difference in the results, were relaxed, allowing a few testers who were over 40 years old to participate. Additionally, the idea to only include people with previous VR experience came from the need to have testers who could better judge the mechanics they are testing, as there was a chance that newcomers to VR would spend most of the testing getting used to just being in VR, and would not really understand it well enough to judge as objectively. In the end, most testers did not have any or had only little previous VR experience, but as the sample also contained VR veterans, the experience difference can be taken into account when comparing the results, and maybe the middle ground between their possible contrasts will show the true result.

5.1.1 Demographics of the sample

The following tables list the main demographics of the sample considered in this study, which are:

- age
- gender, and
- educational background.

The testers' age distribution, illustrated in Table 1, shows that each of the main categories of interest (18-23, 24-30, and 31-40) had a similar amount of testers in them, meaning that the sample happened to have an almost perfect distribution for testing, despite the convenience sampling method used.

Table 1: Testers' age distribution.

Age	Testers
< 18	-
18-23	5
24-30	7
31-40	6
40+	4

Additionally, as mentioned in Section 5.1, the gender distribution (as seen in Table 2) was heavily in favour of males, although a natural reason for that is the fact that the testing supervisor (myself) was a male, and most likely due to that my search for testers appealed more to the male audience, including some male friends and a family member.

Table 2: Testers' gender distribution.

Male	Female	Other
19	3	-

In addition to age and gender, the questionnaire had a questions related to the testers' educational backgrounds, namely higher education (distribution shown in Table 3). The reasoning behind this was to illustrate the effects of the university channels used in the search for volunteers, while being able to confirm the bias related to being a current or previous university student (mentioned in Section 5.1).

5.1.2 Characteristics and abilities of the sample

One of the main interests in this study was to see whether the personal traits of the testers, especially perceived skills related to the explored

Table 3: Testers' tertiary education background distribution. Some of the testers had both graduated from a university and were currently studying in another degree program as well.
Abbreviations:
UTU = Univesity of Turku,
TUAS = Turku University of Applied Sciences.

Education type	UTU	TUAS	Other in Finland	Other abroad
Studying at	10	4	-	-
Graduated from	3	5	1	3

mechanics, would affect their puzzle enjoyment levels. The chosen skills and abilities were:

- athleticness
- swimming ability
- diving ability
- climbing ability, and
- ability in the English language.

The athleticness distribution (illustrated in Table 4) showed that most of the testers were not very athletic and, additionally, that the sample did not have a very equal distribution. This means that the results related to athleticness will be somewhat unreliable, as the comparisons have to be made with groups of testers with either very small or large sizes.

Table 4: Testers' perceived athleticness distribution.

Perceived athleticness	Testers
Not athletic at all	3
Somewhat athletic	13
Quite a bit athletic	4
Very athletic	2

When it came to the testers' skills (as seen in Table 5), the perceived abilities of all of the skills were centered around the middle options, with only a few testers perceiving to be masters or to have no ability at all, suggesting that the distribution was healthy. Additionally, the English language skill, which was kind of a mandatory requirement for the testing, was a lot higher than the other skills, with no tester rating that they do not understand English at all, which would have led to having to discard their results.

Table 5: Testers' perceived skill distribution.

Skill	No ability	Basic	Average	Above average	Master/Native
Swimming	1	7	6	7	1
Diving	3	10	6	1	2
Climbing	1	8	8	5	-
English	-	1	6	13	2

In addition to the skills, the questionnaire asked about possible phobias of the testers (shown in Table 6), which could affect both performance and enjoyment. The thought behind the idea was that the phobias in question would occur in the Crawling and climbing puzzle, which features tight and high spaces. Unfortunately, the testing sample did not contain many testers suffering from these phobias, so it was quite impossible to determine whether the phobias truly had an effect or not.

Table 6: Testers' phobia distribution, featuring acrophobia (the fear of heights) and claustrophobia (fear of tight enclosed spaces).

Suffers from	Not at all	Slightly sometimes	Quite often	Constantly
Acrophobia	11	9	1	1
Claustrophobia	14	6	1	1

Additionally, there were many questions related to various personal traits of the testers, which included:

- previous VR experience
- previous gaming experience (divided into amount and frequency aspects)
- enjoyment of challenge and puzzles, and
- familiarity with the testing supervisor.

The previous VR experience distribution (as seen in Table 7), showed that most testers had little to no previous experience, while the more experienced testers (which were mostly acquired from the HIVE as mentioned earlier in this chapter) were mostly at the other end of the experience scale (developers). Although the testers were not exactly evenly distributed, the comparisons should be healthy enough in order to possibly see some meaningful results.

Table 7: Testers' previous VR experience distribution (multiple choices for individuals allowed in case they owned a VR system).

VR experience	**Testers**
No experience	6
Some experience	10
Very experienced	-
Owns a VR system	3
Develops for VR	5

The testers gaming experience, as mentioned, was divided into two aspects: the amount of playtime weekly (in hours), and the frequency of weekly play sessions (illustrated in tables 8 and 9). The idea behind the division was to distinguish the "hardcore" and "casual" gamers in the testing sample, and just one of the measurement types might not have necessarily done that, as some games are more time-consuming than others, while others are played more frequently (e.g., games with daily login rewards).

The next studied aspects related to enjoyment, more specifically the enjoyment of challenge and puzzles in general. The challenge aspect could have had more response options in it, although the main focus

Table 8: Testers' weekly gaming amount estimate distribution.

Weekly game time in hours	Testers
0-1	7
2-4	5
5-10	3
11-20	6
20+	1

Table 9: Testers' gaming frequency estimate distribution.

Weekly gaming sessions	Testers
<1	7
1-3	5
4-7	6
7+	4

was just to get a general idea about the sample's preference, and that succeeded, showing that basically all of the testers do enjoy challenge (as highlighted in Table 10). However, challenge cannot be defined objectively, as every person experiences challenge differently in comparison to their own abilities, meaning that everyone was in a way answering a different question. This means that even if there would have been a lot of variance within the results, which there was not, using the statistic would have resulted in unreliable, subjective results.

Table 10: Testers' enjoyment of challenge in games.

Enjoys challenge in games	Testers
Yes	20
No	-
Sometimes	1
N/A	1

The puzzle enjoyment aspect was related to the challenge aspect, as puzzles could be in general perceived to be the "definition of challenge", and the thought behind the question was to make sure that the sample did not have any major bias towards puzzles, as that would have needed to be considered when calculating the results. According to the results, as seen in Table 11, most of the sample either loved puzzles or was at least indifferent about them (with average enjoyment level of 3.86), meaning that the results might be slightly higher than if they would be taken from a larger, more varied population of testers.

Table 11: Testers' enjoyment of puzzles.

Puzzle enjoyment	**Testers**
1	-
2	2
3	7
4	5
5	8

The next measured aspect was testers' familiarity with the testing supervisor (deviation can be seen in Table 12), which is a significant factor in this thesis, because research has shown that people tend to act differently around people they are familiar with when compared to complete strangers, while all tests were performed under the supervision of the same researcher, which only some of the testers knew beforehand. As a hypothesis, it can be expected that testers who were already familiar with the researcher would either be more honest, maybe even brutally honest, because they might feel that they could tell how they really feel to a friend. Alternatively, it could be that the more familiar testers would not dare to rate anything very poorly, as they were aware of how much effort the researcher had spent on making the puzzles and the testing happen, so they maybe subconsciously thought that the researcher would be insulted if they just told they disliked everything. These types of responses are related to the *social desirability bias*, explained in Section 6.1.1.

Additionally, an example of a publication related to this "friends versus strangers" difference in behaviour is *When modesty prevails: Differential favorability of self-presentation to friends and strangers (1995)* by [23],

Table 12: Testers' familiarity with the testing supervisor.

Familiarity	Testers
Complete stranger	12
Vague acquaintance	4
Close acquaintance	5
A family member	1

which although talking about self-presentation, concludes that people tend to be more modest around their friends and more self-enhancing to strangers. This result can be interpreted as that it is quite likely that the familiarity level of the testers did affect the way the testers rated or would have rated if the supervisor would have been someone else.

Finally, the testers were asked about how they feel about the future of VR. This was done both in order to see pre-existing attitudes towards VR more clearly, and as well to see whether the testing process could change the opinions of some of the testers. The change would be evaluated by asking the same question again at the end of the questionnaire when the testing was completed. According to the early results, around 75% of the testers saw VR as a revolutionary tool for the gaming world, so the overall attitudes can be seen as highly positive towards VR.

Table 13: Testers' future vision of VR.

VR will revolutionise the gaming world	**Testers**
Yes	15
No	5
Maybe	1
N/A	1

5.2 Hypotheses

The first hypothesis was that players who suffer from certain phobias will experience virtual reality related to those phobias stronger than those who do not. This means that, for example, people who are afraid of tight spaces are more likely to either be faster than average in clearing the Crawling and climbing puzzles or then a lot slower than average, as it can be assumed that their fear can either cripple them or boost their speed, based on the common reactions people tend to have when confronting their fears. These common reactions are "flight or fight" and the latest addition *freeze*, as highlighted in this article from Psychology Today by Christopher Bergland [24]. Unfortunately, the testing sample did not really suffer from any phobias, so this hypothesis was difficult to verify or discard.

The second hypothesis was that players who are already hardcore gamers and/or into virtual reality, will find all of the puzzles more enjoyable than people who are complete beginners, just because they will not be confused by the controls to the same extent and they have more realistic sets of expectations about what is possible in VR.

The third hypothesis was that players who rate themselves more athletic than average will prefer the more realistic versions of the puzzles because they enjoy the real-life correspondents of those actions already, while the under average rating players will prefer the less realistic versions as they will require less effort. This will be compared in the Swimming and Crawling and climbing puzzles, as they are the ones designed to require physical effort.

The fourth hypothesis was that based on the testing order players will find the puzzles they test first to be more enjoyable than the ones they test last, as their initial shock of trying out something new has probably already faded at that point and there is also chance that they are experiencing some kind of VR sickness. This is why the test subjects will be tested in batches of four, each batch having a unique testing order, consisting of puzzle (Swimming, Crawling and climbing or Seeking and finding) and version (realistic (A) or gamelike (B) versions first) orders.

The fifth hypothesis was that the testers will not rate the enjoyment of any of the puzzle types or variations highly on average, as the lack of proper aesthetics and story affects their enjoyment negatively.

The sixth hypothesis was that the testers who are not familiar with the testing supervisor (myself) will find the puzzles on average less enjoyable than testers who are. This is based on the fact that it is often easier to judge someone's work when there is no emotional connection with the creator, as that connection can skew the judgement, either negatively or positively, although it is suspected that mostly positively in this case.

6 Results and analysis

The main idea in this thesis was to compare the puzzle mechanics in a scale of gamelikeness versus realism and to see whether the realistic or the gamelike aspects (puzzle variations) are found to be more enjoyable by the testing sample. Additionally, the study looked into other aspects which could possibly have had an effect on enjoyment in order to get more verification to the result. In total, the aspects which were compared are as follows:

1. The perceived realism of the movement mechanics and puzzle surroundings, this applies to the Swimming and the Crawling and climbing puzzles.

2. The effects of the skills and traits of the testers, these include previous gaming and VR experience, real-life skills corresponding to the mechanics (e.g., swimming, diving and climbing) and familiarity with the testing supervisor.

3. The effects of the testing order which took place.

4. The perceived difficulty of the puzzles, this consists of the perceived difficulty of using the mechanics and performance.

In addition to displaying and analysing results of these aspects, this chapter presents the recognized error factorials and some more general results unrelated to any specific puzzle alongside with observations.

6.1 Error factorials

The study has multiple possible factors which could reduce the reliability of the results, which consist of response biases, the questionnaire design and some identified flaws in the design of the puzzles, which are presented in this section.

6.1.1 Response biases

The *response bias*, which is a general term for a wide range of tendencies for testers to respond inaccurately or falsely to questionnaire questions, has to be accounted for when considering the reliability of the testers' responses. The response biases which can be identified in this study are:

Social desirability bias, which refers to the tendency of answering questions in such way that the responses are or could be viewed favourably by others. In the case of this study the "others" mainly refers to the testing supervisor, who was the only person actually being able to see the responses being made, sometimes commenting on them in order to make sure the questions are correctly understood. [25]

Acquiescence bias, also known as Agreement bias, refers to the tendency for the testers to select the positive option, despite their true preference. This bias can be guessed to having had an impact in case the results are very positive overall, although it is very difficult to confirm individually. [26]

Extreme responding bias ([27]), which results the testers mostly only choosing the extreme options (in this case 1 and 4) and this can stem from cultural and personal backgrounds which encourage certain types of behaviours (e.g., *machismo*, which refers to a culture of exaggerated masculinity, which could theoretically encourage more "yes or no" type of responses to questions, as "being manly" could be seen as knowing exactly what you want and how you feel).

Question order bias, which is a response bias stemming from the ordering of the questions in the questionnaire and it refers to testers responding differently to the same question based on when in the questionnaire it was introduced. [28]

Demand characteristic, which refers to type of response bias where participants form an interpretation of the experiment's purpose and subconsciously change their behavior to fit that interpretation. [29]

Central tendency bias, which is a tendency to avoid choosing the extreme response options. This can be caused by the fear of being seen as an extremist (a form of social desirability bias) or, it can be caused by the expectation that the respondent will have stronger views in later questions in the questionnaire, so they are in a way "saving" the stronger options for those (even though they may never come up).

Dunning-Kruger effect, which is a *cognitive bias* (a systematic pattern of deviation from norm or rationality in judgment) in which people assess their cognitive ability as greater than it is. It is related to the cognitive bias of *illusory superiority*, a condition of cognitive bias wherein a person overestimates their own qualities and abilities in relation to the same qualities and abilities of other people, and comes from the inability of people to recognize their lack of ability.

As described by social psychologists David Dunning and Justin Kruger in 1999, the cognitive bias of illusory superiority results from an internal illusion in people of low ability and from an external misperception in people of high ability; that is, "the miscalibration of the incompetent stems from an error about the self, whereas the miscalibration of the highly competent stems from an error about others". [30]

The Dunning-Kruger effect is an important error factor in the questionnaire of this thesis, as it means that testers who have lesser talent are more likely to have rated themselves higher when it comes to skills which they mostly lack, while testers with above average abilities might consider themselves more of an average. This means that it is very difficult to trust the face value of the results, and that the rated skills would need to be compared against the actual achieved puzzle completion scores in order to find out how much the bias has possibly affected them.

6.1.2 Possible questionnaire design improvements

The main issue related to the questionnaire design is the response order, which was static even though the testing order of the puzzles was not. This means that most testers had to respond to the questions about the puzzles in a different order than they tested them, which may have caused difficulties in remembering the feelings they were having during the testing because of the confusion caused by the altered order. Additionally, as the questionnaire was filled only after testing all of the puzzles, the testers had some time to forget their initial reactions to the puzzles, which could have been prevented by allowing the testers to fill out the corresponding questions immediately after completing the puzzles. However, this would have required the questionnaire to be designed in such way that the filling order could have been chosen by the tester based on their testing order. This could have been done, although it was not realised during the time when the questionnaire was created and, in hindsight, it would have increased the possible effects of the question order bias (defined in Section 6.1.1).

Other improvements which could have been made are the following:

- Some of the questions in the questionnaire were slightly vague and/or difficult to understand for a lot of the testers, making the answers related to those questions less reliable, as it cannot be certain whether the testers had understood the questions properly, despite the efforts of the testing supervisor to ensure that.
- The answer options could have been slightly altered, as their wordings are too similar to each other in one case as the option "disagree to some degree" can be understood to mean the same as "agree to some degree". Thankfully, according to observations it seemed that most testers perceived them to have a clear difference and there was only one comment about the similarity during the testing.
- There were no questions explicitly asking whether a version was better than another one, but rather just asking to rate puzzles individually, and adding this question could have made the results clearer to see. However, it is possible that it would have been too difficult for the testers to give a definitive answer to that question, so it may have been better that the enjoyment was calculated by the puzzle-specific comparisons instead.

6.1.3 Other design flaws

The puzzle completion time was only recorder for testers who completed the puzzles, but in hindsight it maybe would have been better to also record the time for those who did not succeed, in order to measure how long they endured before giving up, as this probably would have corresponded with their puzzle enjoyment ratings.

In the Seeking and finding puzzle, randomizing the colours for different testers was maybe not the best design choice as it could have reduced comparability of the results, although even if the colours were the same, the testers would have reacted to them differently anyway. A positive of this design is that it makes it almost impossible for the testers to gain any knowledge of the hideouts beforehand, as it would be difficult to describe a hideout location by shapes only, which means that no tester could have been prepared for the Seeking and finding puzzle any better than others. The colours also got some positive feedback as some testers felt that they made the level feel more "ready" and the experience was more enjoyable because of that, compared to, for example, the Crawling and climbing puzzle which was mostly black and white, clearly indicating that it was just a test level.

6.2 Observations

The observations were recorded both during the testing and filling out of the questionnaires, and they consists of feedback received from the testers both prompted and unprompted by the testing supervisor in addition to any kind of behaviour escaping the norm by the testers noticed by the supervisor.

The observations are divided into general, which discuss anything related to the whole testing process, and puzzle-specific, which highlight aspects related to a certain puzzle type.

6.2.1 General observations

In general, a lot of the testers needed to ask clarifying questions related to the vocabulary used in the questionnaire, for example related to words such as "immersion" and "intuitive", and many were not prepared to answer questions related to personal traits and abilities and struggled to decide their perceived skill levels. When it comes to the performance in the puzzles, some testers clearly rated the levels to be too easy in relation to how long it took for them to complete them, indicating the presence of the Dunning-Kruger effect (see Section 6.1.1).

Apart from just clarifying questions, some of the testers were asking for straight out hints on how to complete the puzzle when they felt lost or did not understand something. It is noteworthy that a lot of these questions were already answered in the puzzle instructions (see Appendix A), indicating that maybe some testers did not read them so carefully, understand them well enough or just forgot about them when they got so focused on being inside VR.

6.2.2 Swimming observations

Observation 1:

For some testers of the Swimming puzzle, the drowning did not serve its purpose as a setback which should be avoided, as instead they realised that as the drowning did not cause a serious penalty, it could be utilised in order to speed up the exploration process. This meant that those testers intentionally drowned just in order to be able to explore areas further away faster than the testers who always returned to the surface to get more oxyen. Thankfully, not all testers had this idea, but there were enough to make a small impact on the reliability of the drowning amount results and the overall completion times.

Observation 2:

Some of the testers got confused about the first two water areas, thinking even that they had returned to the beginning when in reality they had made progress and found their way to the next area. This was partly due to some of the underwater structures having being designed to appear the same way in those areas, and caused some testers to lose time and in some cases their sanity as well. Additional source of confusion was the rope the testers needed to climb between water areas two and three, as a lot of testers simply did not realise that it was there even when looking straight at it, forcing the supervisor to sometimes give additional hints related to it.

Observation 3:

The beacons which guided the testers through the murky water were well received, and even though not all testers instantly associated them with progress, the lightsources delighted almost all of the testers when found. Some of the testers even preferred hanging out around them to diving further into the darkness, which indicated that the water served its purpose in creating the underwater-like athmosphere. Then again, some testers also loved the darkness that the water created.

Observation 4:

Some testers experienced a rare bug where they were not registered entering or exiting the water properly, causing them to be able to levitate or unable to swim. This mainly occured when the testers were attempting to climb the rope, and most likely had minor impacts on the enjoyment levels of testers affected.

Observation 5:

The movement mechanic in the version B was reported to be less immersive mainly because of the teleportation used, and additionally there were difficulties in horizontal navigation when utilising the automatic surfacing mechanic, mostly because it was perceived as being too slow and thus not responsive enough.

Observation 6:

Some of the testers reported claustrophobia caused by the darkness while underwater, which was something that had not been foreseen when designing the testing, so there are no official results about it as it was not included as a question related to the Swimming puzzle. This claustrophobia was so strong for the (very few) testers suffering from it, that one of them almost had a panic attack during the testing, and had to skip one of the version of the Swimming puzzle because of it.

6.2.3 Crawling and climbing observations

Observation 1:

As far as the mechanics were concerned, the less realistic climbing mechanic was clearly perceived to be quite strange to use, as testers were very confused about having to toggle their grabbing off when wanting to stop holding on to a climbable surface. This was also noticed because a lot of the testers were holding the GRAB-buttons down in both version of the puzzle when climbing, even though it was not necessary in the B version, resulting some testers having trouble unattaching themselves from the walls. This is why no matter the result, the toggling should most likely not be used for climbing in VR.

Observation 2:

The puzzle version A had to be reset a few times for some testers because of their hands getting stuck on climbable surfaces, reason for this was unclear as it seemed to happen so rarely and was difficult to recreate. Apart from this bug the testing for this puzzle went relatively smoothly.

Observation 3:

Some testers wrongly thought that the colours of the words had a meaning related to the puzzle, and probably lost time because of it. This was emphasised more in the puzzle briefing after the first observation.

Observation 4:

Some of the vocabulary and wordplays used in the puzzle were difficult to understand for some, although the testing supervisor helped when prompted to do so. This could have been avoided by having the puzzles use Finnish instead of English, but then the possible volunteer base would have decreased as well.

6.2.4 Seeking and finding observations

Observation 1:

In version A, the key which was accessed from a hole below (see Section 4.4.1, figures 22, 23, 24 and 25), was dominantly more difficult to discover than the one accessed from a hole above (in Figure 21). This was most likely due to the design of the hideout being at the start of the entered object, causing a lot of the testers to seek from the end area in vain.

Observation 2:

The hints related to compass points were not understood by some testers, as the level did not feature a compass from which to check those, which confused especially testers who were older than the sample average. Then again some testers realised immediately that the compass points were meant to be imagined to start from the points of views of the hint locations, but overall the design of those type of hints could have used revising.

Observation 3:

After the keys had been found, there were some difficulties in delivering them to their designated locations, because of implementation issues. This showed as most testers managing to lose at least one of the keys due to it going inside a wall or some other object during a rapid movement, even though there was a reset system in place to respawn the keys in case they exited the level.

Observation 4:

The problem related to the of the players' innate height (described in Section 3.3.3) caused the first tester not being able to reach one of the keys in the version A after locating it, as the tester was extremely short. This was fixed afterwards by changing the location of the key slightly, causing no further issues related to the problem.

6.3 General results and abbreviations

This section displays and analyses general results which are not tied to a specific puzzle, these include the following:

- puzzle functionality (number of crashes/glitches/bugs experienced)
- hardware performance related issues
- testing exhaustion (both physical and mental)
- and VR sickness caused by the testing.

Abbreviations related to all the later sections in this chapter:

V = Tested version
C/T = Number of completions/testers who attempted the puzzle version
μ = Mean
σ = Standard deviation
σ-% = Relative standard deviation
$\sigma_{\overline{x}}$ = Confidence interval

All confidence intervals were calculated at 68.3% confidence level and any time measurements listed have been rounded to the closest second.

Statistics related to types of enjoyment, realism, and difficulty were rated on a Likert scale (as defined at the start of Chapter 5) from "Completely or nearly completely disagree" (1) to "Completely or nearly completely agree" (4) and their results were converted to a numerical scale from 1 to 4, where 1 indicates that the perceived type of enjoyment, realism, or difficulty was very low while 4 indicates it was very high.

All puzzle completion statistics were calculated only among testers who completed the puzzle (apart from the completion rate statistics), and all calculated results were rounded to the closest second decimal.

Most testers reported having experienced at least some amount of functionality issues in the puzzles which shaped their individual experiences by a varying degree (as illustrated in Table 14). It is very likely that the ones with a lot of performance issues rated their enjoyment lower partly due to those but, as the extreme cases only had three testers combined, the total effect most likely remained low.

Table 14: Puzzle functionality statistics

Number of crashes/glitches/bugs experienced	Testers
Many	2
Some	13
Very few	6
None	1

On the other hand, hardware related issues did not exist for most testers and, for those few (3) who experienced some, they were rated to have from neutral to very minor negative effect on their experience in the worst-case scenario (table omitted here).

When it comes to the physical and mental side-effects caused by the testing (showcased in Table 15), the more realistic A-versions were, as expected, rated to be more physically exhausting than their B-counterparts. The mental exhaustion, on the contrary, was higher in the B-versions for Swimming and Seeking and finding, indicating that they were more difficult to solve as puzzles than their A-versions, while for Crawling and climbing the difference was quite non-existent, indicating that perhaps the balance was better there between versions. VR sickness was very low overall and seems to have beeen divided quite evenly among different puzzles, showing no signs of a single puzzle being more sickness-inducing than others, although a surprising factor was that the Seeking and finding puzzles had any ratings for VR sickness, as their mechanics are quite mild and their environments rather static and stable compared to the other two.

Part of the VR sickness result can also be explained by both Swimming and Crawling and climbing being quite dark for the most parts, which might have been a key factor in tricking the testers' brain not getting sick so easily by the perceived (or in this case not so perceived) move-

Table 15: Exhaustion and VR sickness statistics. Abbreviations:
PE = Physically exhausting,
ME = Mentally exhausting,
CVS = Caused VR sickness.

Puzzle and version	**PE**	**ME**	**CVS**
Swimming A	10	2	3
Swimming B	-	5	2
Crawling and climbing A	11	4	2
Crawling and climbing B	2	3	2
Seeking and finding A	2	7	4
Seeking and finding B	-	2	2

ment. Then again, the Seeking and finding puzzle is well-lit at all times, leaving the brains of the testers able to fully visualise the movement and, additionally, possibly spinning around in circles searching for small hidden objects can definitely cause a headache even in normal games.

Overall the results indicate that most of the sample did not feel much exhaustion apart from physical in Swimming and Crawling and climbing A-versions, and mental in Seeking A. The results for VR sickness have so few ratings that basically most testers did not suffer from VR sickness during most of the testing.

6.4 Results of the Swimming puzzle

The results of the Swimming puzzle are divided into five categories which are:

- general
- perceived realism based
- testers' personal skills and traits related
- testing order based, and
- perceived difficulty related results.

All of the categories compare different aspects of the testing with types of enjoyment of the Swimming puzzle.

6.4.1 General results

The general results section displays the performance, enjoyment, realism, and difficulty statistics across all testers for the Swimming puzzle, so that they can be reflected upon later in the more specific results, while also providing an overview of how the puzzle was received by the sample. The enjoyment levels measured for the Swimming puzzle were general enjoyment and immersion.

Starting off with the completion statistics in Table 16, which indicate that the Swimming B was more difficult for the testers, as it took longer to complete on average than Swimming A and it also had a slightly lower completion rate. The result suggests that maybe the more realistic swimming mechanic is easier for testers to use, or that the hidden path in the Swimming A was just easier to find.

Table 16: Swimming completion statistics.

V	**C/T**	μ	**Min**	**Mode**	**Median**	σ	σ-%	$\sigma_{\overline{x}}$
A	19/22	14 m 7 s	1 m 40 s	28 m 5 s	12 m 8 s	7 m 8 s	56.8%	1 m 38 s
B	17/21	17 m 6 s	6 m 5 s	35 m	15 m 14 s	9 m 42 s	50.1%	2 m 21 s

When it comes to the number of drownings (illustrated in Table 17), the amount was almost the same on average for both puzzle versions, while the drowning itself was perceived to be more realistic in the Swimming A. The slightly (+0.16) higher drowning rate does, however, correlate with the previous result about Swimming B being perceived as more difficult.

The questionnaire also measured the feelings which emerged when a tester drowned, as seen in Table 18, and for both of the versions "Annoyance" came clearly on top, followed by "Confusion". This result shows at least that the drowning was not perceived as a pleasant experience,

Table 17: Swimming drowning statistics, divided into the quantity and perceived realism aspects.

Measured type	μ	Min	Mode	Median	σ	σ-%	$\sigma_{\overline{x}}$
Quantity A	2.41	0	6	2	2	82.7%	0.43
Quantity B	2.57	0	7	2	1.82	70.6%	0.40
Realism A	1.86	1	4	2	0.81	43.7%	0.18
Realism B	1.52	1	3	1	0.66	43.5%	0.15

like fail-conditions usually tend to be, although according to tester feedback, some testers reported that the drowning was absolutely necessary for the swimming to feel realistic at all.

Table 18: Swimming drowning feeling statistics.

Feeling at the moment of drowning	Testers A	Testers B
Anger	3	2
Fear	1	-
Confusion	7	5
Happiness	3	2
Boredom	3	4
Relief	4	2
Annoyance	12	16
Gratefulness	-	-
Disappointment	1	1
Frustration	1	1

When it comes to the puzzle enjoyment, highlighted in Table 19, the enjoyment statistics across all of the testers show that the version A was more enjoyable and immersive than version B by a clear margin. Additionally, for the version A the immersion was mostly rated to stem from the feeling of being deep underwater, possibility of drowning and the soundscape used in the puzzle, while other highly contributing factors were the feeling of swimming and visuals in general. For version B the feeling of being deep underwater and the soundscape were rated highly as well, and while visuals had around the same rating, possibility of drowning was considerably less immersive and feeling of swimming

Table 19: Swimming enjoyment statistics.

Measured type	μ	Min	Mode	Median	σ	σ-%	$\sigma_{\overline{x}}$
General A	3.09	1	4	3	1	32.2%	0.22
General B	2.67	1	4	3	0.94	35.4%	0.21
Immersion A	2.82	1	4	3	0.78	27.5%	0.17
Immersion B	2.24	1	4	2	0.87	38.7%	0.19

almost non-existent, clearly due to the teleportation breaking immersion as many testers commented. The table for the immersion factors is omitted here.

According to Table 20, the movement in the more realistic A version had higher perceived realism (as designed), intuitivity, and control, which when considering the written and verbal feedback received, clearly indicates that the teleportation being enabled was a really large negative factor for realism. Additionally, the difference in the control aspect was most likely caused by the sideways movement when rising up being perceived to be too slow and difficult to use by some testers. The slight difference between perceived intuitivity was most likely caused partly by teleportation and partly because of the fact that testers got to move their hands in the A when swimming, while in B they needed to remember a few button combinations, which caused difficulties to some.

Table 20: Swimming movement statistics, divided into aspects of perceived realism, intuitivity and control of the movement mechanic.

Measured type	μ	Min	Mode	Median	σ	σ-%	$\sigma_{\overline{x}}$
Realism A	2.91	1	4	3	0.60	20.5%	0.13
Realism B	1.62	1	4	1	0.79	48.5%	0.18
Intuitivity A	2.68	1	4	3	0.87	32.6%	0.19
Intuitivity B	2.19	1	4	2	1.01	45.9%	0.22
Control A	3.09	2	4	3	0.51	16.6%	0.11
Control B	2.86	1	4	3	0.89	31.1%	0.20

6.4.2 Results based on perceived realism

These results study the effects of the perceived realism on puzzle enjoyment for the Swimming puzzle. The statistics consist of the comparisons between perceived realism of the swimming mechanic, drowning and the puzzle surroundings.

Starting with the most important one, the perceived realism of the swimming mechanic (the results of which can be found in Table 21), shows that the results were quite similar all around, although unfortunately many categories had few testers in them, meaning the comparisons were not so equal, which makes the results less reliable. However, as a rule, the testers seemed to have around the same levels of enjoyment in both puzzle versions if they had the same ratings for the realism of the swimming mechanic, except for the three testers who rated it to be very realistic (4), for which the enjoyment was a lot higher (+0.75) for the version A. Additionally, the lowest enjoyment levels were found among testers rating lowest perceived realism (1), which is mostly significant for the version B, as it had most of its testers rating 1 or 2, and shows that there was some connection between lesser perceived realism and lesser enjoyment.

Table 21: Swimming enjoyment statistics based on the perceived realism of the swimming mechanic, the general results for that can be seen in Table 20
Abbreviations:
PR = Perceived realism of the swimming mechanic, ranges from 1 (low) to 4 (high),
GE = General enjoyment,
I = Immersion,
GE+I V = Sum of the means divided by the count of means in version V, not an official mathematical variable but used for comparison purposes.

PR	**T A**	**T B**	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
1	1	11	2	2.36	2	1.73	2	2.05
2	2	8	3	3	3	2.75	3	2.88
3	17	1	3.06	3	2.76	3	2.91	3
4	2	1	4	3	3.5	3	3.75	3

All in all, the realism comparison of the swimming mechanic shows that the higher the perceived realism for both of the puzzles was, the higher the average enjoyment level was as well, even if only by a slight margin, indicating that more realism is more enjoyable. The only contradiction to this is the immersion level between testers who rated 2 and 3, for which the immersion of the version A was higher for the ones who perceived the mechanic to be less realistic, although not by a large margin (+0.24).

When it comes to the realism of the surroundings of the puzzle (highlighted in Table 22), which were not designed to be realistic but to appear as a testing environment, apart from the water itself which was part of the mechanic, the idea was to see if they had any major effect on the average enjoyment level. The surroundings did not change based on the version, so the question was a joint one, completed by all testers who tested at least one of the Swimming versions. The results indicate that similarly to Table 21, the more realism the better, although the differences between the ones who rated 1 and 2 contradict the result very slightly for the more realistic version A. Additionally, there was only one tester who rated 4, and while their combined enjoyment level was quite high (3), it still was lower than at rating 3 (3.34) in version A, but the difference (-0.34) cannot be given very much weight in this case.

The idea behind the perceived drowning statistics in Table 23 was to see if just the realism of the drowning sensation in the puzzle had a significant effect on the enjoyment on its own. According to the results it seems that the effect on enjoyment is difficult to measure, as the results show that those who rated 2 enjoyed the puzzle versions more than those who rated 1 or 3, making the result inconsistent. Additionally, almost no one rated 4, which is to be expected as true drowning feeling is extremely difficult to create artificially, although the one tester who did had a high enjoyment level in both puzzle versions.

6.4.3 Results based on personal traits and abilities

These results consider possible traits, skills and other differences which could affect the rated enjoyment result and compares those in order to more easily see what all the factors affecting enjoyability in the Swim-

Table 22: Swimming enjoyment statistics based on the perceived realism of the testers' surroundings inside the puzzle. The surroundings question was a joint question related to both versions, as the surroundings remained the same between versions.
Abbreviations:
PR = Perceived realism of the surroundings, ranges from 1 (low) to 4 (high),
T = The total amount of testers who rated their realism level of the surroundings according to the value, the true amount of testers for each version can vary in case some testers skipped versions, as this question was answered by all testers who completed at least one of the versions, although the variance is at maximum 1-2 tester,
GE = General enjoyment,
I = Immersion,
GE+I V = Sum of the means divided by the count of means in version V, not an official mathematical variable but used for comparison purposes.

PR	T	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
1	2	2.5	2	2.5	1.5	2.5	1.75
2	7	2.29	2.5	2.57	2.17	2.43	2.34
3	12	3.67	2.83	3	2.33	3.34	2.58
4	1	3	3	3	3	3	3

Table 23: Swimming enjoyment statistics based on the perceived realism of the drowning, responding to statement "I felt like I could really drown".
Abbreviations:
PR = Perceived realism of the drowning, ranges from 1 (low) to 4 (high),
T V = The amount of testers with this perceived drowning realism level in version V,
GE = General enjoyment,
I = Immersion,
GE+I V = Sum of the means divided by the count of means in version V, not an official mathematical variable but used for comparison purposes.

PR	**T A**	**T B**	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
1	8	12	3	2.58	2.5	1.75	2.75	2.17
2	10	7	3.2	2.86	3.1	2.86	3.15	2.86
3	3	2	2.67	2.5	2.67	3	2.67	2.75
4	1	-	4	-	3	-	3.5	-

ming puzzle were.

In Table 24, the effects of familiarity are calculated and the results show that the lowest enjoyment level averages were found at the extremities, as was partly hypothesised (in the sixth hypothesis in Section 5.2). The lower results among strangers could stem from not being afraid to tell how they truly felt because they knew they do not ever need to meet the researcher again, while family members probably felt close enough to not worry about the possible negative emotions their real opinion could cause. Then again, as the vague and close acquaintances had more optimistic evaluations, those can be expected to have been caused by the human connection with the researcher (relating to the social desirability bias, explained in Section 6.1.1). Alternatively, it could be that the small sample size just happened to have this type of deviation, but it is quite unlikely for the familiarity aspect to not have any effect on the enjoyment.

Table 24: Average Swimming puzzle enjoyment levels versus the level of familiarity with testing supervisor. The familiarity distribution can be seen in Table 12.
Abbreviations:
CS = Complete stranger,
VA = Vague acquaintance,
CA = Close acquaintance.

Enjoyment type	**CS**	**VA**	**CA**	**Family**
General enjoyment A	3.33	3.5	3.6	2
General enjoyment B	2.58	2.75	3	1
Immersion A	2.92	3.25	3	2
Immersion B	2.08	2.75	2.5	1
Gener. + Imm. A	3.13	3.38	3.3	2
Gener. + Imm. B	2.33	2.75	2.75	1

The next two tables relate to testers' previous experience with video games and virtual reality, as seen in tables 25 and 26.

When the testers' gaming experience was compared with their enjoyability of the Swimming puzzle, the lowest average enjoyment was found in the most inexperienced tester group (2.43), while the apex (4) was reached somewhere between a "mid-core" and a "hardcore" gamer. Additionally, the combined enjoyment average of those testers who rated to play at least 5-10 hours and/or have at least 4-7 gaming sessions a week (12/22 testers) was around 3.33 while testers who rated to play less than 5-10 hours with 1-3 or less sessions (10/22) had a combined average of 2.8. These results mean that the more hardcore half of the sample enjoyed the more realistic puzzle A roughly 19% more than the more casual half, which goes along with the hypothesis about the effects of previous gaming experience in Section 5.2.

And as for the less realistic version B, the enjoyment minimum (2) was located in the most hardcore section of testers and maximum (4) in the middle ground. The combined enjoyment average for both the more hardcore and the more casual halves for B was around 2.67; which means there was no difference between the enjoyment based on previous experience.

When it comes to immersion, the version A peaked (4) again somewhere between the midcore and hardcore gamers, while the lowest point (2) was found at the extreme hardcore end. The more hardcore half had an average immersion level of 2.83 and the more casual of 2.8, showing virtually no difference. For the version B, the peak (3) and lowest point (1.5) cannot really be pinpointed so clearly, as the results had a high variance among similar gaming experience backgrounds, although it is easy to see that version B was a lot less immersive than A. The more hardcore half's immersion level was 2.17 and the more casual's 2.33, indicating possibly that maybe the expectations which stem from previous gaming experience negatively affect the immersion level when they are not being met.

In general, the results indicate that the midcore testers were the most immersed and enjoyed both of the puzzles the most, while the general enjoyment was higher on average for version A for all other groups of testers except for the most casual (0-1 hours and <1 sessions per week), which slightly (+0.24) preferred the B version. The immersion in version A was higher than or equal to B for all except for the most hardcore tester (20+ hours and 7+ session per week), and so was the combined en-

Table 25: Swimming enjoyment statistics based on previous gaming experience.
Abbreviations:
WHP = Weekly hours played, the average amount of hours the testers spend on playing video games every week,
WGS = Weekly gaming sessions, the average amount of gaming sessions the testers have every week,
T = Testers, number of testers with these habits,
GE = General enjoyment,
I = Immersion,
GE+I V = Sum of the means of version V divided by the count of means
HC = The more hardcore gamers of the sample, testers who play at least 5-10 hours or have at least 4-7 play sessions in a week,
C = The more casual gamers of the sample, testers who play less than 5-10 hours and have less than 4-7 play sessions in a week.

WHP	**WGS**	**T**	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
0-1	< 1	7	2.43	2.67	2.86	2.33	2.65	2.5
2-4	1-3	3	3.67	2.67	2.67	2.33	3.17	2.5
2-4	4-7	2	3	2.5	2.5	2.5	2.75	2.5
5-10	1-3	1	3	3	3	3	3	3
5-10	4-7	2	4	3.5	3.5	1.5	3.75	2.5
11-20	1-3	1	4	4	4	3	4	3.5
11-20	4-7	2	4	2	3	2	3.5	2
11-20	7+	3	2.67	2	2.33	1.67	2.5	1.84
20+	7+	1	3	3	2	3	2.5	3
2-4	Any	5	3.4	2.6	2.6	2.4	3	2.5
5-10	Any	3	3.67	3.33	3.33	2	3.5	2.67
11-20	Any	6	3.33	2.2	3	2.2	3.17	2.2
Any	1-3	5	3.6	3	3	2.6	3.3	2.8
Any	4-7	6	3.67	2.67	3	2	3.34	2.34
Any	7+	4	2.75	2.25	2.25	2	2.5	2.13
HC	**HC**	12	3.33	2.67	2.83	2.17	3.08	2.42
C	**C**	10	2.8	2.67	2.8	2.33	2.8	2.5

joyment and immersion as well. Additionally, when looking at the tester groups which had either the same amount of sessions (Any amount of hours and X sessions) or the same amount of hours (X hours and Any amount of session), the combined immersion and enjoyment were higher in version A than B in all cases, and in most cases by a quite significant margin.

The virtual reality experience (shown in Table 26), on the other hand, provided results showing that the testers who were not familiar with VR previously preferred the less realistic version B, perhaps because it was easier for them to understand. Otherwise the testers with even slight previous experience voted overwhelmingly for the version A with differences ranging from 0.8 to 2 in enjoyment and 0.6 to 1 in immersion. The combined score was also in favour of version A for all other testers than the ones with no previous experience, with the largest differences experienced among testers who owned a VR system.

This result further reinforces the previous results that the more realistic swimming mechanic was more enjoyable and, in this case, it seems that it was just a matter of being used to virtual environment that was the dividing factor.

Table 26: Swimming enjoyment statistics based on previous virtual reality experience, the experience distribution can be seen in Table 7.
Abbreviations:
NF = Not familiar,
SF = Slightly familiar and has possibly tested VR a few times,
OV = Owns a VR system and uses it regularly, practical hands-on familiarity,
DV = Develops things for VR and because of that is very familiar with it.

VR experience	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
NF	2.67	3.4	2.83	2.8	2.75	3.1
SF	3.3	2.5	2.9	2.1	3.1	2.3
OV	3.33	2	2.67	1.67	3	1.83
DV	3.4	2.4	2.8	2.2	3.1	2.3
DV + OV	4	2	3	2	3.5	2

Next, as illustrated in Table 27, the testers' perceived abilities to swim and dive were compared with their puzzle completion times and rates, in order to see whether higher skill level increased performance or not. According to the results, it seems that on average testers' rated ability to swim or dive did not affect their completion time in the swimming version A, as the best result is found in "Average ability/Average ability" category, while the other results are really mixed on both sides. For version B, the best result is located in "Swimming master/Diving master" category, which would alone mean that the most skilled tester got the best result, however, the increase in the completion speed is anything but linear and ricochets from very quick to very slow all the time in both versions, which clearly shows that the correlation just does not exist for this sample. Additionally, a lot of the categories have very few testers in them, making the results less reliable.

The next two tables, 28 and 29, compare the swimming and diving skill individually with the Swimming puzzle enjoyment in order to see if there were any clear correlations between them.

The results in Table 28 show that for version A the enjoyment levels stayed mostly the same around the middle swimming skill level categories (basic, average and above average), which had the most testers in them, while the immersion was slightly higher on the lower skill levels, which was the expected result. For version B the enjoyment went down when the skill level increased, which was also to be expected, and the immersion level was quite inconsistent. The extreme responses only had one tester in each of them, so their results are unreliable.

When it comes to the perceived diving skill, the results in Table 29 came out as quite inconsistent for version A, although the general enjoyment level did seem to increase with the diving skill level, when ignoring the "Above average ability"-section with just one tester. For version B, the enjoyment seems to have decreased when the diving skill increased, with immersion level revealing a similar pattern.

Overall, the version A was rated to be more enjoyable among all other swimming and diving skill groups, except for the testers who did not have any swimming or diving skills. The result, although with this very

Table 27: Swimming and diving skills compared with the completion time and rate. The idea was to see if the actual skill corresponds to puzzle performance.
Abbreviations:
C/T V = Completions/Testers for version V.

Swimming skill	**Diving skill**	C/**T A**	C/**T B**	μ A	μ B
No ability	No ability	0/1	0/1	-	-
Basic ability	No ability	1/1	1/1	22 m 53 s	16 m 8 s
Basic ability	Basic ability	6/6	4/6	14 m 25 s	13 m 35 s
Average ability	No ability	0/1	-/-	-	-
Average ability	Basic ability	1/1	1/1	10 m 58 s	10 m 23 s
Average ability	Average ability	2/2	2/2	4 m 46 s	16 m 42 s
Above average ability	Basic ability	1/1	0/1	11 m 40 s	25 m 39 s
Above average ability	Average ability	4/4	4/4	15 m 5 s	11 m 20 s
Above average ability	Above average ability	1/1	1/1	22 m 25 s	35 m
Above average ability	Diving master	0/1	0/1	-	-
Swimming master	Diving master	1/1	1/1	16 m 9 s	7 m 6 s
No ability	Any	0/1	0/1	-	-
Basic ability	Any	7/7	5/7	15 m 37 s	14 m 5 s
Average ability	Any	5/6	5/5	9 m 40 s	21 m 25 s
Above average ability	Any	6/7	6/7	15 m 44 s	17 m 40 s
Swimming master	Any	1/1	1/1	16 m 9 s	7 m 6 s
Any	No ability	1/3	1/2	22 m 53 s	16 m 8 s
Any	Basic ability	10/10	8/10	13 m 42 s	18 m 1 s
Any	Average ability	6/6	6/6	11 m 39 s	13 m 8 s
Any	Above average ability	1/1	1/1	22 m 25 s	35 m
Any	Diving master	1/2	1/2	16 m 9 s	7 m 6 s

Table 28: Swimming skills compared with enjoyment, the skill distribution can be seen in Table 5.
Abbreviations:
SA = Swimming ability,
N/A = No ability,
B = Basic ability,
AA = Average ability,
AAA = Above average ability,
M = Master level ability,
GE = General enjoyment,
I = Immersion,
GE+I = Sum of the means divided by the count of means, not an official mathematical variable but used for comparison purposes.

SA	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
N/A	1	3	3	3	2	3
B	3	2.86	3	2.14	3	2.5
AA	3.33	2.4	2.5	1.8	2.92	2.1
AAA	3.14	2.43	2.71	2.43	2.93	2.43
M	4	4	4	3	4	3.5

small sample size, thus indicates that more realistic swimming mechanic is the more enjoyable one, except for people who are unaware of how swimming functions altogether.

In addition to the swimming and diving skills, the general athleticness was measured and compared against the enjoyability levels (illustrated in Table 30), with a hypothesis that the testers' who rated themselves being more athletic would prefer the version A (declared in Section 5.2). The results show that the testers who rated to be most athletic had (for some reason) the lowest enjoyment and immersion levels across both versions, albeit still preferring version A, and that the least athletic testers preferred version B. Additionally, the results indicate that testers in the "Quite a bit athletic"-category preferred the less realistic version, which does not fit into the hypothesis, suggesting that maybe athleticness does not correlate with the version preference after all. The results suffer from lack of reliability though, as most of the testers (13) rated to be in the "Somewhat athletic"-category, resulting into unbalanced comparisons.

Table 29: Diving skills compared with enjoyment, the skill distribution can be seen in Table 5.
Abbreviations:
DA = Diving ability,
N/A = No ability,
B = Basic ability,
AA = Average ability,
AAA = Above average ability,
M = Master level ability,
GE = General enjoyment,
I = Immersion,
GE+I = Sum of the means divided by the count of means, not an official mathematical variable but used for comparison purposes.

DA	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
N/A	1.67	3	2.67	3	2.17	3
B	3.3	2.8	3.1	2.2	3.2	2.5
AA	3.5	2.5	2.5	2.17	3	2.34
AAA	2	1	2	1	2	1
M	3.5	3	3	2.5	3.25	2.75

Table 30: Testers' perceived athleticness compared with enjoyment of the Swimming puzzle, the athleticness distribution can be seen in Table 4.
Abbreviations:
A = Perceived athleticness level,
N/A = Not athletic,
SA = Somewhat athletic,
QA = Quite a bit athletic,
VA = Very athletic,
GE = General enjoyment,
I = Immersion,
GE+I = Sum of the means divided by the count of means, not an official mathematical variable but used for comparison purposes.

A	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
N/A	2.33	3.33	2.67	2	2.5	2.67
SA	3.54	2.62	2.85	2.23	3.20	2.43
QA	2.75	3.33	3	3	2.88	3.17
VA	2	1	2.5	1.5	2.25	1.25

6.4.4 Results based on testing order

These results compare the testing order (path) and aim to find out whether the testing order affected the enjoyment in addition to whether it affected performance, which in turn is likely to have affected enjoyment.

The completion time statistics based on the testing order, displayed in Table 31, revealed a significant statistical quality: the completion time was lower for both versions when they were tested as second, when compared to another batch which tested the same version as first, while the testing order remained the same, even though the comparison batches consisted of completely different testers. An example to clarify this goes as follows:

Example testing order: Swimming puzzle is tested as first.

Example version order: The more realistic Swimming version A is tested first, before testing Swimming B.

The randomised testing batch, containing 3-4 testers, each of which tests the three puzzles in the same order, tests the Swimming A with a completion time of X.

Now, according to the statistical quality discovered, the other randomised testing batch of 3-4 testers which also tested the Swimming puzzle as first, but which tested the version B first, has a higher completion time than X for the Swimming A. This is true for all the testing orders of the Swimming puzzle (first, second, and third).

The quality showcases that acquiring previous knowledge about the nature of the puzzles helped solving them by a large margin, despite all the individual differences in tester skill, age, previous gaming and VR experience or other qualities. In addition to this very interesting discovery, the completion speed was compared based on the order when the puzzle was tested by the tester (first, second, or third), with following results:

Fastest average completion time was achieved when Swimming version A was tested as second and slowest when tested first, which fits expectations as during the second puzzle the testers can be assumed to already have familiarised themselves with VR, but they have not got VR sickness or fatigue yet, unlike when testing their third puzzle.

Fastest average completion time was achieved when Swimming version B was tested as third and slowest when tested first, which also partly fits expectations, as at the third puzzle the testers are well aware of how VR works so time is not wasted on that. Low amounts of VR sickness experienced by testers (as shown in Table 15) and because the ones who got sick or did not complete not being counted in the completion average can be reasons for why the third puzzle's performance was better than second's.

All in all, the version tested as first was completed slowest, which makes the most sense as the testers are completely unaware of how the testing will be at that point, in addition to in most cases not being used to how VR works at all. However, the results are not completely accurate as they exclude the input from testers who failed to complete the puzzles which could have made a difference, although at least the completion rate does not seem to have any significance based on the tested version order. Additionally, the path-specific results seem to verify the consensus of Table 16 that version B seems to have been more difficult on average, as all of the paths have faster completion averages for A when completed in the same order.

When it comes to enjoyment based on testing order, as highlighted in Table 32, the enjoyment and immersion averages and medians are higher on average in version A when compared to B and the testing order is the same, while the version order does not matter, and higher or equal when the version order matters. This result further reinforces the previous results indicating that version A and thus the more realistic mechanic seems to have been more enjoyable on average among this sample.

In addition, the version which was tested first was found to be more

Table 31: Swimming completion time statistics based on testing order.
Abbreviations:
Order = Refers to the order of testing the Seeking and finding puzzle for testers (first, second or third)
V 1st = Tells which was the first version of the puzzle the testers in that row tested (A or B),
V 2nd = Tells which was the second version of the puzzle the testers in that row tested (A or B),
RO = Results of, tells which version the results displayed on that row are taken from (A or B).

Order	**V 1st**	**V 2nd**	**RO**	C/T	μ	**Min**	**Mode**	**Median**	σ	σ-%	$\sigma_{\bar{x}}$
First	Either	Either	A	7/8	19 m 27 s	7 m 51 s	28 m 5 s	22 m 25 s	6 m 29 s	33.4%	2 m 28 s
First	A	B	A	3/4	21 m 53 s	20 m s	22 m 53 s	22 m 47 s	1 m 20 s	6.1%	47 s
First	B	A	A	4/4	17 m 37 s	7 m 51 s	28 m 5 s	17 m 17 s	8 m 2 s	45.6%	4 m 1 s
First	Either	Either	B	7/8	21 m 45 s	10 m 30 s	35 m 14 s	16 m 10 s	9 m 13 s	42.4%	3 m 29 s
First	A	B	B	3/4	13 m 57 s	10 m 30 s	16 m 8 s	15 m 14 s	2 m 28 s	17.7%	1 m 26 s
First	B	A	B	4/4	27 m 36 s	16 m 10 s	35 m 14 s	29 m 30 s	8 m 1 s	29%	4 m 1 s
Second	Either	Either	A	7/8	9 m 40 s	1 m 40 s	15 m 20 s	10 m 58 s	4 m 28 s	46.2%	1 m 41 s
Second	A	B	A	3/4	13 m 44 s	11 m 50 s	15 m 20 s	14 m 3 s	1 m 27 s	10.5%	51 s
Second	B	A	A	4/4	6 m 36 s	1 m 40 s	10 m 58s	6 m 53 s	3 m 24 s	41.4%	1m 42 s
Second	Either	Either	B	7/7	14 m 13 s	6 m 53 s	32 m 20 s	10 m 23 s	8 m 28 s	59.6%	3 m 12 s
Second	A	B	B	3/3	8 m 52 s	6 m 53 s	10 m 37 s	9 m 7 s	1 m 32 s	17.3%	54 s
Second	B	A	B	4/4	18 m 13 s	9 m 24 s	32 m 20 s	15 m 35 s	9 m 17 s	51%	4 m 39 s
Third	Either	Either	A	5/6	12 m 14 s	4 m 57 s	22 m 45 s	11 m 40 s	6 m 8 s	47.5%	2 m 45 s
Third	A	B	A	3/3	14 m 37 s	4 m 57 s	22 m 45 s	16 m 9 s	7 m 21 s	50.3%	4 m 15 s
Third	B	A	A	2/3	10 m 19 s	8 m 57 s	11 m 40 s	10 m 19 s	1 m 22 s	13.2%	58 s
Third	Either	Either	B	3/6	12 m 57 s	6 m 5 s	25 m 39 s	7 m 6 s	9 m	69.5%	5 m 12 s
Third	A	B	B	2/3	6 m 36 s	6 m 5 s	7 m 6 s	6 m 36 s	31 s	7.7%	22 s
Third	B	A	B	1/3	25 m 39 s	25 m 39 s	25 m 39 s	25 m 39 s	-	-	-

enjoyable in general in 5/6 cases, although two of the cases were equally enjoyable apart from the median, while the immersion was higher on the version tested first in 4/6 cases. This result indicates that the at least some enjoyment is lost when the puzzle is more familiar to the player, which is quite natural as people tend to enjoy many things the most on their first try.

Additionally,the most enjoyable testing orders based on version were the following:

The most enjoyable testing order for version A was when tested as third, while the least enjoyable was as second. The most immersive order was as third, and the least immersive as first. The combined enjoyment and immersion was highest as third, followed by second, then first. According to the result it seems that the slowest order (first) was also the least enjoyable, which can be expected, while testing as third clearly seems to have been the most enjoyable and immersive, albeit it was only the second fastest by completion time.

For the version B, enjoyment order was [3,1,2], immersion order [3,2,1] and combined enjoyment and immersion [3,1,2]. The results were almost identical to the version A, apart from the combined result having the second and third place swapped, giving a clear indication that the puzzle tested as third was the most enjoyable overall for the Swimming puzzle.

Overall, the individual version results indicate that the enjoyment and immersion are related to the testing order in such a way that testing the Swimming puzzle as third seems to have been the most pleasant experience. This could be due to the previous two puzzles having familiarised the testers with VR so well that the swimming felt more natural, while the completion times were also among the fastest, indicating a possible correlation between completion time and enjoyment.

Table 32: Swimming enjoyment statistics based on testing order.
Abbreviations:
Order = means the order of testing the Seeking and finding puzzle for testers (first, second or third),
V 1st = Tells which was the first version of the puzzle the testers in that row tested (A or B),
V 2nd = Tells which was the second version of the puzzle the testers in that row tested (A or B),
RO = Results of, tells which version the results displayed on that row are taken from (A or B),
GE = General enjoyment,
I = Immersion,
GE+I = Sum of the means divided by the count of means, not an official mathematical variable but used for comparison purposes,
GE+I med. = Sum of the medians divided by the count of medians, not an official mathematical variable but used for comparison purposes.

Order	**V 1st**	**V 2nd**	**RO**	**GE** μ	**GE** median	**I** μ	**I** median	**GE+I**	**GE+I** med.
First	Either	Either	A	3	3.5	2.63	3	2.81	3.25
First	A	B	A	3	3.5	3.25	3.5	3.13	3.5
First	B	A	A	3	3	2	2	2.5	2.5
First	Either	Either	B	2.88	3	2	2	2.44	2.5
First	A	B	B	3	3	2.75	3	2.88	3
First	B	A	B	2.75	3	1.25	1	2	2
Second	Either	Either	A	3	3	2.75	3	2.88	3
Second	A	B	A	2.75	3	2.75	2.5	2.75	2.75
Second	B	A	A	3.25	3.5	2.75	3	3	3.25
Second	Either	Either	B	2.43	2	2.14	2	2.29	2
Second	A	B	B	2.67	2	1.67	1	2.17	1.5
Second	B	A	B	2.25	2.5	2.5	2.5	2.38	2.5
Third	Either	Either	A	3.33	3.5	3.17	3	3.25	3.25
Third	A	B	A	3.33	4	3.33	3	3.33	3.5
Third	B	A	A	3.33	3.5	3	3	3.15	3.25
Third	Either	Either	B	2.67	2.5	2.67	2.5	2.67	2.5
Third	A	B	B	2.33	2	2.33	2	2.33	2
Third	B	A	B	3	3	3	3	3	3

6.4.5 Results based on perceived difficulty

The perceived difficulty for the Swimming puzzle consists of the difficulty of the movement mechanic used, which is divided into intuitivity and control, and the number of drownings the testers experienced, as most likely more drownings means higher perceived difficulty. Additionally, the enjoyment is calculated based on the completion time and rate of the puzzle, in order to see whether they increased or decreased the overall enjoyment levels.

First, according to Table 33, it seems that for version A the general enjoyment and immersion were significantly higher with 4 or less drownings than more than 4, while for version B the combined enjoyment was actually highest for the ones who drowned the most. This might be because the effort of swimming upwards was nonexistent in version B, so trying many times was not experienced as negatively as possibly in version A, which required a lot of effort. Additionally, the more realistic A version had a higher combined enjoyment score in all other categories except the 4+ drownings, reinforcing the conclusion that the cost of trying again in A had a greater negative effect than in B.

Table 33: Swimming enjoyment statistics based on the number of drownings, the general drowning statistics can be seen in Table 17.
Abbreviations:
AR = Number of drownings,
T V = The number of testers with this number of drownings in version V,
GE = General enjoyment,
I = Immersion,
GE+I V = Sum of the means divided by the count of means in version V, not an official mathematical variable but used for comparison purposes.

AR	**T A**	**T B**	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
0	5	1	3.6	4	2.4	1	3	2.5
1-2	8	12	3	2.67	3	2.5	3	2.59
3-4	5	4	3.4	2	3.2	1.5	3.3	1.75
4+	4	4	2.25	3	2.5	2.5	2.38	2.75

Next, according to Table 34, higher intuitivity seems to indicate higher levels of enjoyment and immersion, with only exception being in version

B between ratings 2 and 3. The result is very expectable, and it also further confirms that the version A is overall more enjoyable, as it is rated higher in 3/4 of the combined enjoyment level scores.

Table 34: Swimming enjoyment statistics based on the intuitivity of the swimming mechanic, the general statistics for which can be seen in Table 20.
Abbreviations:
INT = Rated intuitivity of the swimming mechanic,
T V = The number of testers with these rated intuitivity and control levels for version V,
GE = General enjoyment,
I = Immersion,
GE+I V = Sum of the means divided by the count of means in version V, not an official mathematical variable but used for comparison purposes.

INT	**T A**	**T B**	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
1	2	7	2	1.71	2	1.71	2	1.71
2	7	5	2.43	3.2	2.71	2.6	2.57	2.9
3	9	7	3.56	2.86	2.89	2.43	3.23	2.65
4	4	2	3.75	4	3.25	2.5	3.5	3.25

In the following table, Table 35, the results comparing sensation of control with enjoyment indicate that higher sensation of control over the swimming mechanic corresponds to higher enjoyment level as well, which is not a surprising result. The mechanic in version A was already rated to be easier to control earlier in Table 20, and as control seems to have a correlation with the sample's enjoyment levels, the version A "scores" another aspect victory over version B on the enjoyability scale.

Next, in Table 36 the results for how much completing the puzzle affected the enjoyment level can be seen. As expected, the vast majority who did complete the puzzle also enjoyed it more, the difference being massive (+1.11 for combined result and +1.75 for just enjoyment) for version A and slight (+0.07 for combined result and +0.51 for just enjoyment)for version B, as for some reason B was not as immersive for the testers who completed the puzzle, but that could just be due to individual preferences among the few who did not complete the puzzle. Overall it does seem that there is a positive correlation between completing the

Table 35: Swimming enjoyment statistics based on the perceived control over the swimming mechanic, the general statistics for which can be seen in Table 20.
Abbreviations:
CTRL = Rated sensation of control over the swimming mechanic,
T V = The amount of testers with these rated intuitivity and control levels for version V,
GE = General enjoyment,
I = Immersion,
GE+I V = Sum of the means divided by the count of means in version V, not an official mathematical variable but used for comparison purposes.

CTRL	**T A**	**T B**	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
1	-	2	-	1.5	-	2	-	1.75
2	2	4	2	1.75	2.5	1.75	2.25	1.75
3	16	10	3.19	3.1	2.75	2.2	2.97	2.65
4	4	5	3.25	3	3.25	2.8	3.25	2.9

puzzle and enjoyment in the Swimming puzzle, as expected.

Table 36: Swimming enjoyment statistics based on puzzle completion, the amount of completions per version can be seen in Table 16.
Abbreviations:
GE = General enjoyment,
I = Immersion,
GE+I = Sum of the means divided by the count of means, not an official mathematical variable but used for comparison purposes,
GE+I med. = Sum of the medians divided by the count of medians, not an official mathematical variable but used for comparison purposes.

Completion	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
Completed	3.32	2.76	2.89	2.13	3.11	2.45
Did not complete	1.67	2.25	2.33	2.5	2	2.38

The last table in this section, Table 37, compared the puzzle enjoyment with completion time for the Swimming puzzle. The results show that for both puzzle version the testers who enjoyed the puzzle the most were found in the second slowest completion time section (μ - $\mu + 0.5\mu$), which had the highest rating for both general enjoyment and immersion. Additionally, the lowest general enjoyment was rated by testers in the slowest section, unsurprisingly, indicating that the frustration of not being able to complete the puzzle turned the experience to less pleasant one than testers who succeeded in a more reasonable time.

The immersion had its lowest point in the second fastest tester group for A with a small -0.17 difference to the slowest testers, while for B the slowest testers were the least immersed as well. This means at least that most of the slow testers did not intentionally delay their completion because they were so immersed and enjoyed spending time within the puzzle level. Additionally, the results contradict the theory about faster completion time correlating with increased enjoyment, which was hypothesised in the results of Table 32, as the fastest completion time did not correspond to the highest enjoyment and the second fastest time had a lot lower enjoyment levels than the fastest and second to slowest.

6.5 Results of the Crawling and climbing puzzle

The results of the Crawling and climbing puzzle are divided into five categories which are:

- general
- perceived realism based
- testers' personal skills and traits related
- testing order based, and
- perceived difficulty related results.

All of the categories compare different aspects of the testing with types of enjoyment of the Crawling and climbing puzzle.

6.5.1 General results

The general results section displays data about the performance, enjoyment, realism, and difficulty statistics across all testers for the Crawling and climbing puzzle, so that they can be reflected upon later in the more

Table 37: Swimming enjoyment statistics based on the completion time, the average statistics for that can be seen in Table 16.
Abbreviations:
CT V = Completion time for version V, the four intervals are drawn from the general completion statistics in the following way: $< 0.5\mu$; 0.5μ - μ; $> \mu$ - $\mu + 0.5\mu$; $> \mu + 0.5\mu$,
T V = The amount of testers who completed the puzzle version V within the given interval,
GE = General enjoyment,
I = Immersion,
GE+I V = Sum of the means divided by the count of means in version V, not an official mathematical variable but used for comparison purposes.

CT A	**CT B**	**T A**	**T B**	**GE** μ A	**GE** μ B	**I** μ A	**I** μ B	**GE+I** A	**GE+I** B
< 7m 4 s	< 8 m 33 s	3	3	3.67	3.33	3	2	3.34	2.67
7 m 4s - 14 m 7 s	8 m 33 s - 17 m 6 s	8	8	3.13	2.63	2.63	2.25	2.88	2.44
14 m 8 s - 21 m 11 s	17 m 7 s - 25 m 39 s	3	3	4	3.33	3.67	2.67	3.84	3
> 21 m 11 s	> 25 m 39 s	5	3	3	2	2.8	1.67	2.9	1.84

specific results, while also providing an overview of how the puzzle was received by the sample.

First we will go through the statistics related to performance, namely completion time, number of resets, and statistics about which ending the testers reached upon completing the puzzle.

When it comes to completion time, as illustrated in Table 38, the version B had a lot faster average completion time, although the maximum completion time and standard deviation were higher than in version A, while the minimum was lower, suggesting that there were more extreme cases in version B than A. The difference in speed was clearly related to the faster crawling mechanic utilised in the version B, as it allowed testers to trial and error faster than in A, possibly making understanding how the puzzle worked less relevant to some testers (conclusion is partly based on testing observations).

Table 38: Crawling and climbing completion time statistics.

V	C/T	μ	**Min**	**Mode**	**Median**	σ	σ-%	$\sigma_{\overline{x}}$
A	17/20	13 m 36 s	4 m 42 s	22 m 51 s	14 m 2 s	5 m 29 s	40.4%	1 m 20 s
B	18/20	9 m 17 s	2 m 49 s	28 m 6 s	7 m 50 s	6 m 19 s	68.0%	1 m 30 s

The second performance-related aspect was the number of resets experienced by the players, as highlighted in Table 39. The results show that the version B had slightly (+0.15) more resets on average, although both versions had the same minimum and maximum amounts, while the deviation was slightly higher for the version B.

Table 39: Crawling and climbing reset statistics

V	μ	**Min**	**Mode**	**Median**	σ	σ-%	$\sigma_{\overline{x}}$
A	1.95	0	7	2	1.69	86.5%	0.43
B	2.1	0	7	1.5	1.9	90.2%	0.43

The final performance-related aspects were the ending keys reached after completing either of the puzzle versions. These ending keys determined what path the tester had taken to complete the level out of the three possible correct paths, which were different for each version. Results in Table 40 indicate that the LAGOON-path was clearly the easiest to complete in version A, as it had got such a high clear rate compared to the other two paths. This is most likely a design issue which should be fixed in the future by adding to the difficulty of LAGOON, while possibly reducing the difficulty of HAZARD-path by a little bit.

Table 40: Crawling and climbing version A reached ending statistics, which highlight which one of the three possible ending "keys", the testers reached upon completing the version (or none if they did not complete the puzzle).

Ending key	**Testers**
LAGOON	12
CREAM	4
HAZARD	1
None	3

For version B, the ending key distribution (displayed in Table 41) shows that the DROWNED-path was the easiest to complete by a longshot, as it was completed by so many more testers than the other two paths. This is most likely a design issue, similar to version A, which should be fixed in the future by adding to the difficulty of DROWNED, although unlike in version A the clear rates for REALITY and WOUNDED -paths seem healthy enough.

Table 41: Crawling and climbing version B reached ending statistics, which highlight which one of the three possible ending "keys", the testers reached upon completing the version (or none if they did not complete the puzzle).

Ending key	**Testers**
DROWNED	11
REALITY	3
WOUNDED	4
None	2

Next, when it comes to the movement mechanics utilised in the puzzle, the results (as seen in Table 42), show virtually no differences between versions, apart from the perceived realism of the mechanics, which was higher for version A (as intended). Additionally, there was a question measuring the perceived elevation level when climbing up the walls in the puzzle. The main idea about this was for it to be featured as a category of perceived realism in the next section, while additionally the question can find out about the realiability of the testers' responses, as the question was duplicated for both puzzle versions, even though the climbing mechanic was almost the same and the layout of the puzzle does not change. The results should have both been the same (and they nearly were, with just 0.05 difference in favour of A), but there was one tester who changed their response by one based on version (from 2 to 1), although there was nothing special about the testers traits or abilities which would have somehow explained this change, so it can be assumed that it was just an honest mistake and conclude that the responses seemed to be realiable, at least in the aspect of duplicate questions.

Table 42: Crawling and climbing movement statistics, here Climbing Elevation means how much the perceived feeling of altitude difference experienced by the testers when climbing up was. The feeling of altitude difference should not really vary based on the tested version as the climbing mechanic is almost identical in both and the level is identical in shape, so the statistics can show whether the testers are reliable when rating their experiences based on the found or not found variance.

Measured type	μ	**Min**	**Mode**	**Median**	σ	σ-%	$\sigma_{\overline{x}}$
General Realism A	3.1	2	4	3	0.7	22.6%	0.16
General Realism B	2.75	2	4	3	0.62	22.6%	0.14
Climbing Elevation A	3.15	2	4	3	0.65	20.8%	0.15
Climbing Elevation B	3.1	1	4	3	0.83	26.8%	0.19
Climbing Control A	3.55	3	4	4	0.50	14%	0.12
Climbing Control B	3.4	2	4	3	0.58	17.1%	0.14
Crawling Control A	3.35	1	4	3	0.73	21.7%	0.17
Crawling Control B	3.35	1	4	3	0.73	21.7%	0.17

Last but not the least, the pure enjoyment statistics, as seen in Table 43. The results show that the more realistic version A was slightly more enjoyable on average and the minimum enjoyment was also higher (2) than in B, which had someone rating it completely unenjoyable (1). Although the difference is small, it can be said that at least in general the more realistic mechanic seemed to be more enjoyable, but this result

Table 43: Crawling and climbing enjoyment statistics

Measured type	μ	Min	Mode	Median	σ	σ-%	$\sigma_{\bar{x}}$
Enjoyment A	3	2	4	3	0.77	25.8%	0.18
Enjoyment B	2.85	1	4	3	0.79	27.8%	0.18

will be later compared with individual enjoyment results to see if similar difference exists there as well.

6.5.2 Results based on the perceived realism

These results study the effects of the perceived realism on puzzle enjoyment for the Crawling and Climbing puzzle. The statistics consist of the comparisons between perceived realism of the movement mechanics, the puzzle surroundings and the perceived altitude change when climbing the walls.

Starting with the perceived realism of the movement mechanics, results of which are displayed in Table 44, show that similarly to the Table 21, the enjoyment levels of the testers seem to increase when the perceived realism does, although there is a slight drop in version A between the realism levels 2 and 3. Additionally, it is interesting that the less realistic B version achieved close to the same or even greater enjoyment results based on the perceived realism, even though it was not designed to be realistic at all. The source for this result could be traced back to the climbing mechanic, which is quite realistic in both of the puzzle version with minor differences, even though the crawling mechanic, which was the main mechanic of study, is not.

Moving on to the effects of the perceived realism of the surroundings, illustrated in Table 45, it seems that, similarly to Table 22, the higher the realism is, the higher is the enjoyment level as well. In addition, the version A was preferred by the ones perceiving higher levels (3 and 4) of realism, while for the lower ones the results were more even or even in favour of version B.

Table 44: Crawling and climbing enjoyment statistics based on the perceived realism of the movement mechanics (crawling and climbing), the general results for that can be seen in Table 42.
Abbreviations:
PR = Perceived realism of the movement mechanics, ranges from 1 (low) to 4 (high),
T V = The amount of testers with these perceived realism levels for version V,
GE = General enjoyment.

PR	**T A**	**T B**	**GE** μ A	**GE** μ B
1	-	-	-	-
2	4	7	2.75	2.29
3	10	11	2.7	3.09
4	6	2	3.67	3.5

Table 45: Crawling and climbing enjoyment statistics based on the perceived realism of the testers' surroundings.
Abbreviations:
PR = Perceived realism of the surroundings as a combined question for both versions, ranges from 1 (low) to 4 (high),
T = The amount of testers with these perceived realism of the surroundings levels,
GE = General enjoyment.

PR	T	**GE** μ A	**GE** μ B
1	2	2	2.5
2	8	2.88	2.88
3	9	3.22	2.89
4	1	4	3

Finally, the results in Table 46 show that the perceived altitude change corresponds strictly to the enjoyment level, meaning the more altitude was perceived to have changed, thus the more realistic the climbing felt, the more the enjoyment increased as well, which contributes to the conclusion that more realism means more enjoyment. Additionally, the lowest perceived altitude category only had one participant, so the results of that cannot be relied upon.

Table 46: Crawling and climbing enjoyment statistics based on the perceived altitude change when climbing the walls.
Abbreviations:
PR = Perceived altitude change when climbing the walls, ranges from 1 (low) to 4 (high),
T V = The amount of testers with these perceived altitude change levels for version V,
GE = General enjoyment.

PR	**T A**	**T B**	**GE** μ A	**GE** μ B
1	-	1	-	4
2	3	3	2.67	2.33
3	11	9	2.91	2.67
4	6	7	3.33	3.14

All in all, the increased perceived realism did increase the puzzle enjoyment for the Crawling and climbing puzzle, with almost no exceptions, and as version A was perceived (as designed) to be more realistic, the results reinforce the conclusion of version A being more enjoyable in general as well.

6.5.3 Results based on personal traits and abilities

The results in this section consist of testers' personal traits and abilities related to the Crawling and climbing puzzle compared with enjoyability. These are the following:

- familiarity with the testing supervisor
- climbing ability

- gaming and VR experience
- athleticness, and
- English language skill.

The comparison category unique to this puzzle is the English language skill, and it is included because the Crawling and climbing puzzle utilised a lot of English vocabulary and wordplays, which were slightly difficult to understand for some testers who were native Finns, based on observations. The comparison aims to figure out whether poorer or more advanced than average English skills were a factor affecting the enjoyability. The statistics related to this are portrayed in Table 47 and the results showed the following:

For version A, the enjoyment stayed the same despite the English skill, except for the native level category, which was considerably lower for some reason. For version B, the basic ability category was the one with the lowest enjoyment level, while others were similar to version A. However, the extreme categories only had 3 testers in them combined, so they cannot really be considered in the conclusion, leaving us with the result that English language skill level did not have a significant effect on the puzzle enjoyment for this sample.

Table 47: Average Crawling and climbing puzzle enjoyment compared with perceived English language skill (the language skill distribution can be seen in Table 5). This was especially a question which was affected by the response bias of illusory superiority (explained in Section 6.1.1 as part of the Dunning-Kruger effect), as it was very easy for the testing supervisor to determine whether some tester's English speaking ability was closer to basic than above average during the two-hour testing session, yet some testers with observed basic ability rated to have above average ability.

English skill	**General enjoyment A**	**General enjoyment B**
Basic	3	2
Average	3	3.2
Above average	3.08	2.83
Native level	2.5	2.5

Up next, in Table 48, the puzzle enjoyment is compared with testers' familiarity with the testing supervisor (myself). Similarly as in Swimming puzzle (in Table 24), the results indicate that the extremities were possibly more honest with their (negative) opinions, as the lowest values can be found within the strangers and family members, while the closer acquaintances have perhaps chosen their ratings more "politically".

Table 48: Average Crawling and climbing puzzle enjoyment compared with level of familiarity with testing supervisor.

Familiarity level	**General enjoyment A**	**General enjoyment B**
Complete stranger	2.7	2.6
Vague acquaintance	3	3.25
Close acquaintance	3.6	3.2
Family member	3	2

In the following table, Table 49, the perceived climbing ability was compared with the puzzle enjoyment. Results show that the puzzle was most enjoyable on average for testers who rated to possess basic climbing abilities, which could indicate that having a moderately good climbing ability decreases or at least does not increase the enjoyment of puzzle levels simulating that skill. This could be because the perceived action is so different than what the more experienced climbers are used to, while having lesser amount of experience can be a positive as the testers are not so aware of how the action should feel. This theory is challenged by the low score in "No ability"-category, although as there was only one tester who belonged to that group, that result should not be considered.

Additionally, the variation could just be due to the small sample size as the difference in enjoyment between the above average and basic climbing ability was not so significant (0.43), at least in the more realistic A version. However, the difference was quite a bit larger in version B (0.69), which could mean that the realism increased the enjoyment of those who already were familiar with the actions in question in real life more than those who are not. Then again, the ones who rated "Average ability" seem to have had no difference in enjoyment based on tested version, which makes the conclusion for this table quite inconclusive.

Table 49: Average Crawling and climbing puzzle enjoyment compared with perceived climbing ability.

Climbing ability	**General enjoyment A**	**General enjoyment B**
No ability	2	2
Basic ability	3.43	3.29
Average ability	2.71	2.71
Above average ability	3	2.6

Next, the other real-life "physical" attribute, perceived athleticness, was compared with the puzzle enjoyment (seen in Table 50). The results indicate that the athleticness did not play a large role in defining puzzle enjoyment for this puzzle, as the statistics are almost identical for all categories, apart from the "Very athletic"-category, which was significantly lower, but it only consisted of the opinions of two testers, which means its results are not reliable enough to be included.

Table 50: Average Crawling and climbing puzzle enjoyment compared with perceived athleticness.

Athleticness	**General enjoyment A**	**General enjoyment B**
Not athletic at all	3	3
Somewhat athletic	3	3
Quite a bit athletic	3.25	2.75
Very athletic	2.5	2

The next two tables are related to the previous gaming and VR experience compared with the puzzle enjoyment.

First, Table 51 compares testers' previous gaming experience with their puzzle enjoyment, and unlike in the Swimming puzzle in Table 25, the gaming experience did not seem to be a differentiating factor for the Crawling and climbing puzzle. The enjoyment values ranged from 2.33 to 4 for version A and from 2.33 to 3.66 for B, with the worst values found among the most and the least hardcore gamers, meaning that the midcore seems to have enjoyed this puzzle the most, although not by a large margin.

When zooming into the more hardcore and more casual halves of the sample, the more hardcore half had an average of 3.08 enjoyment for A and the more casual one 2.88, while for version B the hardcore netted 2.83 and casual 2.88. While these differences are not great, the trend can be seen for this puzzle as well: the more experienced gamers enjoyed the version A slightly more than the ones with next to no experience, while the B version was very slightly preferred by the less experienced ones, indicating that the easier mechanic is more enjoyable to the inexperienced, although in this case only by a marginal amount.

The main takeway from this table is again that the more realistic puzzle seems to have been more enjoyable all around, albeit only very slightly, as the "Any/X and X/Any"-results show that the versions were equal in 4/6 cases while the rest were in favour of version A.

Second, the VR experience linking to puzzle enjoyment in Table 52 shows that the version B was more enjoyable among 3/5 experience categories, only ones being in favour of version A were those with slight previous VR experience, with one draw. The differences overall between version are quite slight and as both the experts as well as complete beginners seemed to slightly prefer the less realistic mechanic, the only conclusion to be drawn out of the table is that previous VR experience does not seem to have a significant effect on the enjoyment of the Crawling and climbing puzzle.

6.5.4 Results based on testing order

These results aim to estimate the effects of the testing order that took place during testing by comparing it with puzzle enjoyment and completion time.

When it comes to the enjoyment comparisons, results (illustrated in Table 53) show that in 4/6 cases the version which was tested first achieved higher average enjoyment score with one tie between the versions, which

Table 51: Crawling and climbing statistics based on previous gaming experience. Abbreviations:
WHP = weekly hours played, the average amount of hours the testers spend on playing video games every week,
WGS = weekly gaming sessions, the average amount of gaming sessions the testers have every week,
T = testers, number of testers with these habits,
GE = General enjoyment,
HC = the more hardcore gamers of the sample, testers who play at least 5-10 hours or have at least 4-7 play sessions in a week,
C = the more casual gamers of the sample, testers who play less than 5-10 hours and have less than 4-7 play sessions in a week.

WHP	**WGS**	**T**	**GE** μ A	**GE** μ B
0-1	< 1	7	2.6	2.4
2-4	1-3	3	3.33	3.66
2-4	4-7	2	3.5	3
5-10	1-3	1	3	3
5-10	4-7	2	3.5	2.5
11-20	1-3	1	4	3
11-20	4-7	2	3	3.5
11-20	7+	3	2.33	2.33
20+	7+	1	3	3
2-4	Any	5	3.4	3.4
5-10	Any	3	3.33	2.67
11-20	Any	6	2.83	2.83
Any	1-3	5	3.4	3.4
Any	4-7	6	3.33	3
Any	7+	4	2.5	2.5
HC	**HC**	12	3.08	2.83
C	**C**	10	2.88	2.88

Table 52: Crawling and climbing enjoyment statistics based on previous virtual reality experience; the experience distribution can be seen in Table 7.
Abbreviations:
NF = Not familiar,
SF = Slightly familiar and has possibly tested VR a few times,
OV = Owns a VR system and uses it regularly, practical hands-on familiarity,
DV = Develops things for VR and because of that is very familiar with it.

VR experience	GE μ A	GE μ B
NF	3.2	3
SF	3.22	2.89
OV	2.67	2.67
DV	2.6	3
DV + OV	3	3.5

is in accordance with the results from the Swimming puzzle in Table 32, further reinforcing the theory that the first version might be more enjoyable to test, partly because the mechanic, puzzle and the setting are brand new to the tester.

The most enjoyable testing order for version A was as second, followed by first then third. For the version B, enjoyment order was [1,2,3]. The enjoyment order contradicts the supposed conclusion from Table 32 that the puzzle which was tested as last would be the most enjoyable, and makes it difficult to draw any further conclusions about the subject.

As far as the enjoyability goes for the versions themselves, the more realistic A version had higher enjoyment average in 2/3 cases, losing only when tested as first with 0.17 lower average score. The correlation is not as strong as in the Swimming puzzle, but it seems that the more realistic mechanic was more enjoyable for Crawling and climbing as well.

The completion statistics of the testers based on the testing order are displayed in Table 54. Fastest average completion time was achieved when Crawling and climbing version A was tested as third and slowest when tested as first. Results of tested as third were also significantly better than as first or second, but that could just be related to the sample's differences.

Table 53: Crawling and climbing enjoyment statistics based on testing order. Abbreviations:
Order = Refers to the order of testing the Seeking and finding puzzle for testers (first, second or third)
V 1st = Tells which was the first version of the puzzle the testers in that row tested (A or B),
V 2nd = Tells which was the second version of the puzzle the testers in that row tested (A or B),
RO = Results of, tells which version the results displayed on that row are taken from (A or B),

Order	**V 1st**	**V 2nd**	**RO**	**Enjoyment μ**	**Enjoyment median**
First	Either	Either	A	3	3
First	A	B	A	3	3
First	B	A	A	3	3
First	Either	Either	B	3.17	3
First	A	B	B	2.75	3
First	B	A	B	3.67	4
Second	Either	Either	A	3.125	3
Second	A	B	A	3.25	3.5
Second	B	A	A	3	3
Second	Either	Either	B	2.875	3
Second	A	B	B	2.75	3
Second	B	A	B	3	3
Third	Either	Either	A	2.83	3
Third	A	B	A	2.5	2.5
Third	B	A	A	3	3
Third	Either	Either	B	2.5	2.5
Third	A	B	B	1.5	1.5
Third	B	A	B	3	3

For version B, the fastest order was as third and slowest when tested as first, which fits into the theory about performance increasing with VR experience. Low amounts of VR sickness experienced by testers (as can be seen in Table 15) and the ones who got sick or did not complete not being counted in the completion average can be reasons for why the third puzzle's performance was better than the second's. Results of third were also significantly better than first or second, but that could be related to the sample's differences, which are analysed further into this chapter.

The significant statistical quality that the sample had is that all of the randomised testing batches had a lower completion time in a version they tested as second, when compared to another batch which tested the same version as first, when their testing order of the puzzle was the same, even though the comparison batches consisted of completely different testers, similarly to the testing order results analysed earlier in the Swimming puzzle (in Section 6.4.4 which contains a clarifying example) in Table 31. This further reinforces that acquiring previous knowledge about the nature of the puzzles helped solving them by a large margin, despite the individual differences in tester skill, age, previous gaming and VR experience or other qualities.

Although the results are not completely accurate as they do not include any input from testers who failed to complete puzzles, the completion rate was also higher or equal in the cases where a version was tested second rather than first, apart from when Crawling and climbing was tested as second for the version A, where testing A first got 4/4 completion rate and second only 2/4.

6.5.5 Results based on the perceived difficulty

Perceived difficulty includes the number of resets, movement control and difficulty in choosing the correct path compared with enjoyability, while the perceived realism is related to the experienced realism of the movement mechanics.

First off, in Table 55 the number of resets is compared with the enjoy-

Table 54: Crawling and climbing completion statistics based on testing order. The possible endings for both versions are displayed in tables 40 and 41 Abbreviations:
Order = Refers to the order of testing the Seeking and finding puzzle for testers (first, second or third)
V 1st = Tells which was the first version of the puzzle the testers in that row tested (A or B),
V 2nd = Tells which was the second version of the puzzle the testers in that row tested (A or B),
RO = Results of, tells which version the results displayed on that row are taken from (A or B),
MCE = most common ending, tells which one of the three possible endings was achieved the most by testers
Ending types: **LG** = LAGOON, **DD** = DROWNED, **RY** = REALITY, **CM** = CREAM, **AEQ** = all ending types tied

Order	**V 1st**	**V 2nd**	**RO**	C/T	μ	**Min**	**Mode**	**Median**	MCE	σ	σ-%	$\sigma_{\overline{x}}$
First	Either	Either	A	5/6	16 m 59 s	13 m 24 s	22 m 51 s	16 m 46 s	LG	3 m 22 s	19.8%	1 m 31 s
First	A	B	A	2/3	19 m 49 s	16 m 46 s	22 m 51 s	19 m 49 s	LG	3 m 3 s	15.4%	2 m 10 s
First	B	A	A	3/3	15 m 6 s	13 m 24 s	17 m 51 s	14 m 2 s	LG	1 m 58 s	13%	1 m 9 s
First	Either	Either	B	6/6	11 m 29 s	2 m 49 s	28 m 6 s	8 m 56 s	DD	8 m 34 s	74.5%	3 m 30 s
First	A	B	B	3/3	7 m 40 s	2 m 49 s	16 m 1 s	4 m 9 s	DD	5 m 56 s	77.4%	3 m 26 s
First	B	A	B	3/3	15 m 19 s	8 m 12 s	28 m 6 s	9 m 39 s	RY	9 m 4 s	59.1%	5 m 14 s
Second	Either	Either	A	6/8	14 m 53 s	7 m 24 s	21 m 7 s	17 m	LG	5 m 21 s	36%	2 m 12 s
Second	A	B	A	4/4	15 m 12 s	8 m 3 s	18 m 47 s	17 m	LG	4 m 23 s	28.8%	2 m 12 s
Second	B	A	A	2/4	14 m 16 s	7 m 24 s	21 m 7 s	14 m 16 s	LG/CM	6 m 52 s	48.1%	4 m 51 s
Second	Either	Either	B	6/8	10 m 15 s	4 m 12 s	18 m 41 s	9 m 44 s	DD	4 m 59 s	48.6%	2 m 3 s
Second	A	B	B	3/4	5 m 45 s	4 m 12 s	6 m 57 s	6 m 6 s	AEQ	1 m 9 s	20%	40 s
Second	B	A	B	3/4	14 m 46 s	12 m 31 s	18 m 41 s	13 m 5 s	DD	2 m 47 s	18.9%	1m 37 s
Third	Either	Either	A	6/6	9 m 29 s	4 m 42 s	16 m 24 s	8 m 36 s	LG	4 m 22 s	46%	1 m 47 s
Third	A	B	A	2/2	12 m 6 s	11 m	13 m 12 s	12 m 6 s	LG	1 m 6 s	9.1%	47 s
Third	B	A	A	4/4	8 m 11 s	4 m 42 s	16 m 24 s	5 m 48 s	LG/CM	4 m 47 s	58.4%	2 m 24 s
Third	Either	Either	B	6/6	6 m 7 s	3 m 7 s	9 m 24 s	6 m 26 s	DD	2 m 23 s	38.9%	59 s
Third	A	B	B	2/2	5 m 39 s	3 m 28 s	7 m 50 s	5 m 39 s	DD	2 m 11 s	38.6%	1 m 33 s
Third	B	A	B	4/4	6 m 21 s	3 m 7 s	9 m 24 s	6 m 26 s	DD	2 m 26 s	38.4%	1 m 14 s

ability level, and the main finding is that the enjoyment seems to stay around the same despite the reset amount. Additionally, in version B the enjoyment level is highest among the testers with most resets, similarly to the result about the effects of drowning amount in the Swimming puzzle version B in Table 33. Even though for the Crawling and climbing puzzle the differences are very small in comparison to the Swimming puzzle, it is quite interesting that the B version was more enjoyable for both with many resets or drownings. Perhaps this means that the ones who drowned or were reset the most also got to experience the most and that is why they got emotionally attached to the puzzle version and thus gave it a higher rating? More plausible scenario is that the testers' preferences just happened to be in favour of that version overall, as both puzzles had so few testers with more than four resets/drownings that the result cannot be relied on too much.

Table 55: Crawling and climbing enjoyment statistics based on the number of resets. Abbreviations:
AR = number of resets per tester,
T V = The amount of testers with this amount of resets for version V,
GE = General enjoyment.

AR	**T A**	**T B**	**GE** μ A	**GE** μ B
0	4	4	3.25	2.75
1-2	10	9	2.8	2.89
3-4	4	4	3.25	2.75
4+	1	3	3	3

Next up, results in Table 56 clearly showcase that the higher the perceived control over the crawling mechanic, and thus lower the difficulty of controlling, the higher the enjoyment level as well. The level of control was very high overall so the only difference that could be measured was between those who felt either near complete or quite good control over the mechanic, and between those two the result is in favour of the higher level of control by a significant margin. Additionally, the results add to the result that the more realistic version would be more enjoyable overall for the Crawling and climbing puzzle.

Table 56: Crawling and climbing enjoyment statistics based on perceived control over the crawling mechanic.
Abbreviations:
CTRL = Perceived control over the crawling mechanic, ranges from 1 (low) to 4 (high),
T V = The amount of testers with these perceived control ratings for version V,
GE = General enjoyment.

CTRL	**T A**	**T B**	**GE** μ A	**GE** μ B
1	1	1	2	2
2	-	-	-	-
3	10	9	2.8	2.67
4	9	10	3.33	3.1

Moving on to the results in Table 57, which mirror the result of Table 56, confirming that higher perceived control over the movement mechanics corresponds to increased enjoyment, which is expectable. For the climbing control, the results were, similarly to crawling control, very high, indicating that most testers did not have problems with controlling their actions in this puzzle.

Table 57: Crawling and climbing enjoyment statistics based on perceived control over the climbing mechanic.
Abbreviations:
CTRL = Perceived control over the climbing mechanic, ranges from 1 (low) to 4 (high),
T V = The amount of testers with these perceived control ratings for version V,
GE = General enjoyment.

CTRL	**T A**	**T B**	**GE** μ A	**GE** μ B
1	-	-	-	-
2	-	1	-	3
3	9	10	2.56	2.4
4	11	9	3.36	3.33

Next, according to the results displayed in Table 58, the enjoyment seems to have peaked when the difficulty proved challenging (3), but went down again when it became overbearing (4) in both puzzle versions. The result is a very expected one, especially according to the previous statistics about the testers' enjoyment of challenge in Table 10, in which most testers rated to enjoy challenge. It is most likely the case that when the challenge becomes too much to take the enjoyment suffers, then again it is difficult to say which one is better according to the results: too much or too little challenge?

Table 58: Crawling and climbing enjoyment statistics based on perceived difficulty of figuring out the next correct path.
Abbreviations:
PD = Perceived difficulty of figuring out the next correct path, ranges from 1 (low) to 4 (high),
T V = The amount of testers with these perceived difficulty ratings for version V,
GE = General enjoyment.

PD	**T A**	**T B**	**GE** μ A	**GE** μ B
1	-	-	-	-
2	5	6	2.6	2.67
3	8	5	3.25	3.2
4	7	7	3	2.57

The next two tables (59 and 60) compare the completion times and rates to the perceived enjoyability, in order to see how the effect of the different performance levels.

According to Table 59, higher completion time seems to correlate with greater enjoyment level within the studied sample. Possible reason for this could be that after trying to find the correct path for so long and spending so much effort crawling and climbing through the level, eventually finishing it can feel like such a great achievement that it corresponds to the overall enjoyment as a positive factor rather than a negative one. For the more and less realistic mechanics, it seems that within the completion time measurement, their popularity is quite even with both versions having higher ratings in 2/4 of the completion time sections.

The completion rate, in Table 60, has higher enjoyment level for the testers who did complete the puzzle in version A, and for the testers who did not in version B. The latter result is probably due to the small

Table 59: Crawling and climbing enjoyment statistics based on the completion time, the average statistics for that can be seen in Table 16.
Abbreviations:
CT V = Completion time for version V, the four intervals are drawn from the general completion statistics in the following way: $< 0.5\mu$; 0.5μ - μ; $> \mu$ - $\mu + 0.5\mu$; $> \mu + 0.5\mu$,
T V = The amount of testers who completed the puzzle version V within the given interval,
GE = General enjoyment.

CT A	**CT B**	**T A**	**T B**	**GE** μ A	**GE** μ B
$<$ 6m 48 s	$<$ 4 m 39 s	3	5	3	2
6 m 48s - 13 m 36 s	4 m 39 s - 9 m 17 s	5	7	2.6	3
13 m 37 s - 20 m 24 s	9 m 18 s - 13 m 56 s	7	3	3.29	3.67
$>$ 20 m 24 s	$>$ 13 m 56 s	2	3	3.5	3

number of testers who did not complete the version B (2), as it is quite unlikely that not completing a puzzle is more enjoyable than completing it.

Table 60: Crawling and climbing enjoyment statistics based on puzzle completion, the amount of completions per version can be seen in Table 38.
Abbreviations:
GE = General enjoyment.

Completion	**GE** μ A	**GE** μ B
Completed	3.06	2.83
Did not complete	2.67	3

6.6 Results of the Seeking and finding puzzle

For the Seeking and finding puzzle, the results are mainly focused on the enjoyment of searching and finding the keys, while the mechanic itself exists in a way “within” the design of the hideouts. The idea is that the hideouts (the hiding *mechanic*) in the “more realistic” version are more difficult to implement in traditional games, thus they are less “gamelike”, which creates the difference in the “realism” of the two puzzle versions.

Due to the previous paragraph, the Seeking and finding puzzle does not have any specific measurements for just realism, as both of the puzzle versions can be considered " realistic", meaning that both of the ways of hiding objects can occur in the real world. Additionally, the level itself is not designed to look realistic, so it would not have made sense to ask questions related to it.

As a contrast, partly due to the missing measurable realism aspect, the Seeking and finding puzzle has the most measurable levels of enjoyability: apart from the general puzzle enjoyment, the enjoyment is divided into the key searching enjoyment and the surprise factor of finding the keys. The surprise factor exists to measure the challenge of searching the keys apart from a plain difficulty measurement: a hidden object can be difficult to find even if the player would have narrowed its location down to a small radius, thus finding it there eventually might not be surprising, but rather frustrating. The surprise factor measures the design of the hideouts: if the surprise factor was high then most likely the hideout was designed well, meaning it fulfilled its purpose.

6.6.1 General results

The general results section displays the performance and enjoyment statistics across all testers for the Seeking and finding puzzle, so that they can be reflected upon later in the more specific results, while also providing an overview of how the puzzle was received by the sample.

When it comes to completion, as illustrated in Table 61, the version B was, unsurprisingly, way easier to complete and it had a lot (50%) lower average completion time as well. This result confirms that the design of the versions functioned as intended.

Table 61: Seeking and finding completion statistics.

V	C/T	μ	**Min**	**Mode**	**Median**	σ	σ-%	$\sigma_{\overline{x}}$
A	15/22	14 m 2 s	4 m 19 s	22 m	11 m 30 s	5 m 58 s	42.5%	1 m 33 s
B	19/20	7 m 24 s	1 m 1 s	20 m	4 m 55 s	4 m 55 s	66.3%	1 m 8 s

The general enjoyment statistics (highlighted in Table 62) seem to indicate that there was not much difference in the enjoyment level between versions, as while the general enjoyment is slightly higher in A and the key searching enjoyment slightly higher in B, overall the results are almost identical, as even the standard deviations are all almost the same. However, the surprise factor has a considerable difference in favor of version A, and version B seems to have quite mixed results between different testers in that section. The puzzle was designed so that the surprise factor should be lower in version B, so that seems to have worked as intended, although there is anything but consensus among testers about that with 45.6% relative standard deviation and it is surprising that the difference in surprise factor does not seem to reflect on the overall puzzle enjoyment that much, as according to personal experience finding hidden items in games feels more enjoyable when the hideout is not very obvious.

Table 62: Seeking and finding enjoyment statistics, surprise factor refers to whether the testers were surprised when they found one or both of the keys in a puzzle version, 1 = not surprised, 4 = very surprised.

Measured type	μ	**Min**	**Mode**	**Median**	σ	σ-%	$\sigma_{\overline{x}}$
General enjoyment A	3.36	1	4	3.5	0.77	22.9%	0.17
General enjoyment B	3.25	2	4	3	0.77	23.6%	0.18
Key searching enjoyment A	3.18	2	4	3	0.72	22.5%	0.16
Key searching enjoyment B	3.3	1	4	3	0.78	23.7%	0.18
Surprise factor A	3	2	4	3	0.63	21.1%	0.15
Surprise factor B	2.45	1	4	3	1.12	45.6%	0.25

6.6.2 Results based on personal traits and abilities

These results take into account possible traits, skills and other differences which could affect the rated enjoyment and compares those in order to more easily see what the real result could be. For the Seekind and finding puzzle, these are:

- familiarity with the testing supervisor,
- previous gaming experience, and

- previous VR experience.

Table 63 showcases seeking puzzle's enjoyment level averages versus familiarity with the testing supervisor, and the results again show that the lowest enjoyment levels were found in the "Complete stranger" category followed by "Family", similarly to the other two puzzles in tables 24 and 48. The result seems consistent enough across all of the three puzzles to confirm the theory that familiarity with the testing supervisor did have an effect on the enjoyment, the effect being that strangers enjoyed the puzzles less than testers with some level of acquintance, ignoring the family member category as it consisted of only one tester.

Table 63: Seeking and finding enjoyment averages versus level of familiarity with testing supervisor.
Abbreviations:
CS = Complete stranger,
VA = **VA**gue acquaintance,
CA = Close acquaintance.

Enjoyment type	**CS**	**VA**	**CA**	**Family**
General enjoyment A	3.5	3.75	3.4	3
General enjoyment B	3	3.75	3.4	3
Key searching enjoyment A	3	3.5	3.4	3
Key searching enjoyment B	2.9	3.75	3.8	3
Surprise factor A	2.8	3.25	3.2	3
Surprise factor B	2.2	3.25	2.2	3
Combined enjoyment A	3.1	3.5	3.33	3
Combined enjoyment B	2.7	3.58	3.13	3

In Table 64 the Seeking and finding puzzle's enjoyment levels were compared with the testers' previous gaming experience with the following results:

As could be expected, for both puzzles the general and key searching enjoyment levels were overall higher among testers with more previous gaming experience in the hardcore half, while the surprise factor was higher among the more casual half of the testers. The surprise factor was higher for the version A, apart from testers with 11-20 WHP and 4-7 WGS, who for some reason were more suprised by version B's hideouts, most likely just due to individual tester differences. Overall, the less gamelike hideouts in version A had higher combined enjoyment levels in 13/17 measured tester categories, with two ties, indicating a very clear preference towards version A, except for the most casual testers.

When it comes to previous VR experience, as illustrated in Table 65, the general puzzle enjoyment and surprise factor were in most cases rated higher for the more realistic A version, but it was more enjoyable to search for the hidden keys in the B version. The combined (C) results favor A in 3/5 of the categories while the complete VR-beginners remain undecided and surprisingly the developers with headsets preferred version B. The result is more difficult to interpret to be caused by realism or the lack of it than the other puzzle types because both of the versions were quite realistic in a way that the hideouts in both versions would be possible in real life, just the more difficult one (A) would not be so easily possible to implement in traditional games. Overall the puzzle enjoyment was extremely high, all combined results for version A were higher than 3 while for B only two were not and even those were within a 0.15 range of it.

6.6.3 Results based on the testing order

The testing order based results consists of the completion time and enjoyment comparisons related to the testing order which took place.

The completion time statistics are displayed in Table 66, with the fastest average completion time being achieved when Seeking and finding version A was tested as third and slowest when tested first. Results of testing as third were also significantly better than as first or as second, but that could be related to the sample's individual differences.

For the version B, fastest average completion time was achieved when tested as second, although the difference in the average to when tested as third was only 3 seconds, which is not statistically significant, while testing as first was around 2 minutes slower on average than the other two.

Table 64: Seeking and finding enjoyment statistics based on previous gaming experience. Abbreviations:
WHP = weekly hours played, the average amount of hours the testers spend on playing video games every week,
WGS = weekly gaming sessions, the average amount of gaming sessions the testers have every week,
T V = testers, amount of testers with these habits in version V,
GE = General enjoyment,
KS = Key searching enjoyment
SF = Surprise factor, meaning how surprised the player was when finding one or both of the keys
Comb. V = Sum of the means of version V divided by the count of means, not an official mathematical variable but used for comparison purposes,
HC = the more hardcore gamers of the sample, testers who play at least 5-10 hours or have at least 4-7 play sessions in a week,
C = the more casual gamers of the sample, testers who play less than 5-10 hours and have less than 4-7 play sessions in a week.

WHP	**WGS**	**T**	**GE** μ A	**GE** μ B	**KS** μ A	**KS** μ B	**SF** A	**SF** B	**Comb.** A	**Comb.** B
0-1	< 1	7	3.29	3	2.57	2.67	3.14	2.5	3	2.72
2-4	1-3	3	3	3	3.67	3.67	3	2.67	3.22	3.11
2-4	4-7	2	3	3	3	3.5	3	1	3	2.5
5-10	1-3	1	4	4	4	4	3	3	3.67	3.67
5-10	4-7	2	3.5	3	3.5	3	4	4	3.67	3.33
11-20	1-3	1	4	4	4	4	4	4	4	4
11-20	4-7	2	4	4	3.5	4	2	3	3.17	3.67
11-20	7+	3	3.33	3	3.33	3.33	2.33	1.67	3	2.67
20+	7+	1	4	4	3	3	3	2	3.33	3
2-4	Any	5	3	3	3.4	3.6	3	2	3.13	2.87
5-10	Any	3	3.67	3.5	3.67	3.5	3.5	3.5	3.61	3.5
11-20	Any	6	3.67	3.5	3.5	3.67	2.6	2.67	3.26	3.28
Any	1-3	5	3.4	3.4	3.8	3.8	3.2	3	3.47	3.4
Any	4-7	6	3.5	3.4	3.33	3.6	3	2.4	3.28	3.13
Any	7+	4	3.5	3.25	3.25	3.25	2.5	1.75	3.08	2.75
HC	HC	12	3.58	3.45	3.42	3.55	2.9	2.36	3.3	3.12
C	C	10	3.2	3	2.9	3	3.1	2.56	3.07	2.85

Table 65: Seeking and finding enjoyment statistics based on previous virtual reality experience, the experience distribution can be seen in Table 7.
Abbreviations:
NF = Not familiar,
SF = Slightly familiar and has possibly tested VR a few times,
OV = Owns a VR system and uses it regularly, practical hands-on familiarity,
DV = Develops things for VR and because of that is very familiar with it,
GE = General enjoyment,
KS = Key searching enjoyment
SF = Surprise factor, meaning how surprised the player was when finding one or both of the keys
C V = Combined sum of the means of version V divided by the count of means, not an official mathematical variable but used for comparison purposes.

VR Exp	**GE** μ A	**GE** μ B	**KS** μ A	**KS** μ B	**SF** A	**SF** B	**C** A	**C** B
NF	3.83	3.8	3	3.2	3.33	3.2	3.39	3.4
SF	3	2.89	3.2	3.33	3.11	2.33	3.10	2.85
OV	3.67	3.33	3	3	2.5	2.33	3.06	2.89
DV	3.8	3.6	3.6	3.8	2.25	2.2	3.22	3.2
DV + OV	4	4	3.5	4	2	3	3.17	3.67

The significant statistical quality that the sample had is that all of the randomised testing batches had a lower completion time in a version they tested as second, when compared to another batch which tested the same version as first, when their testing order of the puzzle was the same, even though the comparison batches consisted of completely different testers, similarly to results in the Swimming and Crawling and climbing puzzles in tables 31 and 54. This result mostly confirms that acquiring previous knowledge about the nature of the puzzles helps solving them by a large margin, despite the individual differences in tester skill, age, previous gaming and VR experience or other qualities.

However, the results are not completely accurate as they do not include any input from testers who failed to complete puzzles, although even the completion rate was higher or equal in the cases where a version was tested second rather than first.

The statistics related to enjoyment based on testing order (displayed in Table 67) show that, in 4/6 cases, the version which was tested as

Table 66: Seeking and finding completion time statistics based on testing order. Abbreviations:
Order = Refers to the order of testing the Seeking and finding puzzle for testers (first, second or third)
V 1st = Tells which was the first version of the puzzle the testers in that row tested (A or B),
V 2nd = Tells which was the second version of the puzzle the testers in that row tested (A or B),
RO = Results of, tells which version the results displayed on that row are taken from (A or B),

Order	**V 1st**	**V 2nd**	**RO**	C/T	μ	**Min**	**Mode**	**Median**	σ	σ-%	$\sigma_{\overline{x}}$
First	Either	Either	A	4/8	17 m 51 s	10 m 50 s	22 m	19 m 18 s	4 m 14 s	23.7%	2 m 7 s
First	A	B	A	2/4	19 m 18 s	18 m 55 s	20 m	19 m 18 s	43 s	3.7%	31 s
First	B	A	A	2/4	16 m 25 s	10 m 50 s	22 m	16 m 25 s	5 m 35 s	34%	3 m 57 s
First	Either	Either	B	8/8	8 m 56 s	3 m 30 s	20 m	7 m 30 s	5 m 22 s	60%	1 m 54 s
First	A	B	B	4/4	5 m 35 s	3 m 30 s	10 m	4 m 25 s	2 m 35 s	46.2%	1 m 18 s
First	B	A	B	4/4	12 m 18 s	6 m	20 m	11 m 35 s	5 m 19 s	43.3%	2 m 40 s
Second	Either	Either	A	5/6	13 m 2 s	4 m 19 s	20 m 12 s	11 m 30 s	6 m 13 s	47.7%	2 m 47 s
Second	A	B	A	2/3	14 m 35 s	9 m 9 s	20 m	14 m 35 s	5 m 26 s	37.2%	3 m 51 s
Second	B	A	A	3/3	12 m	4 m 19 s	20 m 12 s	11 m 30 s	6 m 30 s	54.1%	3 m 45 s
Second	Either	Either	B	4/5	6 m 15 s	1 m 1 s	11 m 13 s	6 m 23 s	3 m 45 s	60.1%	1 m 53 s
Second	A	B	B	1/2	1 m 1 s	1 m 1 s	1 m 1 s	1 m 1 s	-	-	-
Second	B	A	B	3/3	7 m 59 s	4 m 55 s	11 m 13 s	7 m 50 s	2 m 34 s	32.2%	1 m 30 s
Third	Either	Either	A	6/8	12 m 20 s	6 m	21 m	10 m	5 m 36 s	45.5%	2 m 18 s
Third	A	B	A	3/4	12 m 59 s	10 m	18 m 57 s	10 m	4 m 13 s	32.5%	2 m 27 s
Third	B	A	A	3/4	11 m 40 s	6 m	21 m	8 m	6 m 39 s	57%	3 m 51 s
Third	Either	Either	B	7/7	6 m 18 s	3 m 25 s	16 m	3 m 58 s	4 m 27 s	70.6%	1 m 42 s
Third	A	B	B	3/3	3 m 42 s	3 m 33 s	3 m 58 s	3 m 26 s	11 s	5%	7 s
Third	B	A	B	4/4	8 m 16 s	3 m 25 s	16 m	6 m 49 s	5 m 5 s	61.6%	2 m 33 s

second achieved a higher average combined enjoyment score, although there were ties between the ratings in the general and key searching enjoyment levels, while the surprise factor's result was 4/6 in favor of the version tested first with one tie. This indicates that the version order did not play a large role in the Seeking and finding puzzle's enjoyment.

The most enjoyable testing order for both versions was as second, followed by first then third. The enjoyment order confirms the contradiction of Table 53 from Table 32 and the conclusion seems to be that the testing order did not have a significant effect on the overall puzzle enjoyment, as the results are not very unified across the three puzzles.

When comparing the enjoyment levels based on different enjoyability types, the more realistic A version has higher general enjoyment average in 2/3 cases with one tie, but loses in all cases of key searching enjoyment to version B, while winning in the suprise factor 2-1. The differences in the surprise factor, which were to be expected because of the nature of the different hideouts, is so significant that it pushes the combined score to be in favor of A 2-1. So even the Seeking and finding puzzle would suggest that the more realistic mechanic is more enjoyable, although the difference is tiny and based on the types of enjoyability measured could go either way.

6.6.4 Results based on the perceived difficulty

For the Seeking and finding puzzle the main aspect of perceived difficulty was the difficulty of finding the keys in comparison with the factors which reduce difficulty, which were the hints and the help gained out of them. Additionally, the completion time and rate were compared with enjoyability in order to see if they had any impact.

Starting with the main aspect of the difficulty of finding the keys, as showcased in Table 68, it seems to have had slightly inconsistent effects on the testers. For version A, the general enjoyment seems to have been higher for testers rating 2 and 3 than testers rating 1 and 4, while the key searching enjoyment has next to no differences between different ratings. The surprise factor then again seems to increase the higher the perceived

Table 67: Seeking and finding enjoyment statistics based on testing order. Abbreviations:
Order = Refers to the order of testing the Seeking and finding puzzle for testers (first, second or third)
V 1st = Tells which was the first version of the puzzle the testers in that row tested (A or B),
V 2nd = Tells which was the second version of the puzzle the testers in that row tested (A or B),
RO = Results of, tells which version the results displayed on that row are taken from (A or B),
GE = General enjoyment of the puzzle
KS = Key searching enjoyment
SF = Surprise factor, meaning how surprised the player was when finding one or both of the keys
Comb. = Sum of the means divided by the count of means, not an official mathematical variable but used for comparison purposes
Comb. med. = Sum of the medians divided by the count of medians, not an official mathematical variable but used for comparison purposes

Order	**V 1st**	**V 2nd**	**RO**	**GE** μ	**GE median**	**KS** μ	**KS median**	**SF** μ	**SF median**	**Comb.** μ	**Comb.** med.
First	Either	Either	A	3.25	3.5	2.875	3	3.375	3	3.17	3.17
First	A	B	A	3	3.5	2.75	2.5	3.5	3.5	3.08	3.17
First	B	A	A	3.5	3.5	3	3	3.25	3	3.25	3.17
First	Either	Either	B	3.25	3.5	3.125	3	2.375	2.5	2.92	3
First	A	B	B	3	3	3	3.5	2.25	2	2.75	2.83
First	B	A	B	3.5	3.5	3.25	3	2.5	2.5	3.08	3
Second	Either	Either	A	3.5	3.5	3.5	3.5	2.8	3	3.27	3.33
Second	A	B	A	3.67	4	3.33	3	2.5	2.5	3.17	3.17
Second	B	A	A	3.33	3	3.67	4	3	3	3.33	3.33
Second	Either	Either	B	3.4	3	3.6	4	3.2	3	3.4	3.33
Second	A	B	B	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5
Second	B	A	B	3.33	3	3.67	4	3	3	3.33	3.33
Third	Either	Either	A	3.375	3.5	3.25	3	2.71	3	3.11	3.17
Third	A	B	A	3.5	3.5	3.5	3.5	3	3	3.33	3.33
Third	B	A	A	3.25	3.5	3	3	2.5	2.5	2.92	3
Third	Either	Either	B	3.14	3	3.29	3	2.25	2	2.89	2.67
Third	A	B	B	3.33	3	3.67	4	2	1	3	2.67
Third	B	A	B	3	3	3	3	2	2	2.67	2.67

difficulty is, when the one tester who rated 1 is excluded from the result.

For version B, the general enjoyment stays around the same at all ratings, while the key searching enjoyment seems to have been slightly more enjoyable for the ones who perceived difficulty to be quite easy (2). The surprise factor has highly inconsistent results, as the ones who rated 2 rated it a lot lower than the ones who rated 1 or 3, which would mean that the surprise factor is higher when the hideout is more difficult, but also higher when it is less difficult, which is incoherent as the rating 2 already means quite easy.

Overall, the less gamelike hideouts in version A had higher enjoyment levels in the combined (C), general, and key searching enjoyment sections, while surprise factor was closer to a tie. In conclusion, according to the combined results, the perceived difficulty did not have a major effect on the enjoyability levels for the Seeking and finding puzzle, which is quite a surprising result.

Table 68: Seeking and finding enjoyment statistics based on perceived difficulty of finding the keys.
Abbreviations:
PD = Perceived difficulty of finding one or both of the keys, ranges from low (1) to high (4),
T V = Testers, amount of testers with these perceived difficulty ratings in version V,
GE = General enjoyment,
KS = Key searching enjoyment
SF = Surprise factor, meaning how surprised the player was when finding one or both of the keys
C V = Combined sum of the means of version V divided by the count of means, not an official mathematical variable but used for comparison purposes.

PD	**T A**	**T B**	**GE** μ A	**GE** μ B	**KS** μ A	**KS** μ B	**SF** A	**SF** B	**C** A	**C** B
1	1	4	3	3	4	3	3	3	3.33	3
2	4	10	4	3.3	3.75	3.5	2.5	1.8	3.42	2.87
3	6	6	3.67	3.33	3.5	3.17	2.8	3.17	3.32	3.22
4	10	-	3.1	-	3.5	-	3.3	-	3.3	-

Another difficulty related aspect are the factors that reduce it, in this case the in-game hints. The effects of the perceived help gained from the in-game hints are listed in Table 69, and the following conclusions can be drawn out of them:

For version A, the higher level of help gained seems to have meant higher general and key searching enjoyment levels as well, at least when ignoring the categories with very few testers in them. Additionally, as can be expected, the surprise factor was slightly lower when the help gained was higher, although the differences are quite menial.

For version B, the general and key searching enjoyment levels seem to have stayed around the same across all levels of help gained, while the surprise factor was inconsistently low among the ones who rated 3, similarly to the result of the ones who rated 2 in Table 68, suggesting that the surprise factor aspect for the version B does not correlate with any type of difficulty related aspect.

Table 69: Seeking and finding enjoyment statistics based on perceived help gained from in-game hints.
Abbreviations:
PH = Perceived help gained from in-game hints, ranges from low (1) to high (4),
T V = Testers, amount of testers with these perceived gained help ratings in version V,
GE = General enjoyment,
KS = Key searching enjoyment
SF = Surprise factor, meaning how surprised the player was when finding one or both of the keys
C V = Combined sum of the means of version V divided by the count of means, not an official mathematical variable but used for comparison purposes.

PH	**T A**	**T B**	**GE** μ A	**GE** μ B	**KS** μ A	**KS** μ B	**SF** A	**SF** B	**C** A	**C** B
1	1	2	4	3.5	3	4	4	3.5	3.67	3.67
2	2	3	3.5	3.33	3	3.33	3	2.67	3.17	3.11
3	10	7	3.1	3	3	3.14	3	1.86	3.03	2.67
4	9	8	3.67	3.38	3.44	3.25	2.88	2.63	3.33	3.09

In the next two tables, 70 and 71, the completion rate and time are compared with enjoyability in order to see the possible differences.

For the completion rate statistic, the surprise factor was omitted, as it

Table 70: Seeking and finding enjoyment statistics based on puzzle completion, the amount of completions per version can be seen in Table 38.
Abbreviations:
GE = General enjoyment,
KS = Key searching enjoyment
C V = Combined sum of the means of version V divided by the count of means, not an official mathematical variable but used for comparison purposes.

Completion	**GE** μ A	**GE** μ B	**KS** μ A	**KS** μ B	**C** A	**C** B
Completed	3,8	3,42	3,27	3,32	3,54	3,37
Did not complete	3,29	3	3	3	3,15	3

would not really make sense to compare surprise factor of finding keys for testers who did not necessarily find any keys. For version B, the completion rate cannot really be compared as there was only one tester who did not complete the version, however for the version A, the ones who did complete the puzzle had a considerably higher perceived enjoyment both combined and individually. Additionally, among the ones who did complete the versions, the version A was more enjoyable in general and when combined, while version B had narrowly higher perceived key searching enjoyment level.

According to results in Table 71, for version A the general and key searching enjoyment levels were higher for testers who completed the puzzle in less than average time than those who took longer, while the surprise factor increased more the more time the completion took. For version B, the general enjoyment reached highest ratings among testers in the extremities of the completion time scale, while the enjoyment of searching for the keys grew lesser the longer it took to search. The surprise factor surged when taking longer than the fastest testers, but remained around the same for all the other completion speed categories.

Overall, the combined enjoyment level seems to have gone lower the longer the completion took for version A, while for version B the enjoyment stayed around the same across all completion times. Additionally, the version A was rated to be slightly ($<$11%) more enjoyable in 3/4 of the completion time categories, although a lot of it was due to the surprise factor category, which was not the fairest of comparisons as the

Table 71: Seeking and finding enjoyment statistics based on the completion time, the average statistics for that can be seen in Table 61.
Abbreviations:
CT V = Completion time for version V, the four intervals are drawn from the general completion statistics in the following way: $<$ 0,5μ; 0,5μ - μ; $>$ μ - μ + 0,5μ; $>$ μ + 0,5μ,
T V = The amount of testers who completed the puzzle version V within the given interval,
GE = General enjoyment,
KS = Key searching enjoyment
SF = Surprise factor, meaning how surprised the player was when finding one or both of the keys
C V = Combined sum of the means of version V divided by the count of means, not an official mathematical variable but used for comparison purposes.

CT A	**CT B**	**T A**	**T B**	**GE** μ A	**GE** μ B	**KS** μ A	**KS** μ B	**SF** A	**SF** B	**C** A	**C** B
$<$ 7m 4 s	$<$ 8 m 33 s	2	5	3,5	4	4	3,4	2,5	1,8	3,33	3,07
7 m 4s - 14 m 7 s	8 m 33 s - 17 m 6 s	6	6	3,83	3	3,5	3,5	2,8	2,67	3,38	3,06
14 m 8 s - 21 m 11 s	17 m 7 s - 25 m 39 s	6	4	3,17	3,25	3	3,25	3,17	2,5	3,11	3
$>$ 21 m 11 s	$>$ 25 m 39 s	1	4	3	3,5	2	3	4	2,75	3	3,08

hideouts had such different difficulties depending on the version.

6.7 Combined results

This section sums up all the enjoyment comparisons together in order to see the final result about which mechanic was perceived to be more enjoyable. The comparisons are calculated based on all the different result categories, apart from the general results considering the sample as a whole, which are analysed separately.

The calculations are divided into two different types, individual and combined enjoyment, in order to see the contrasts between them, which are partly caused by the selection of the enjoyment levels. The individual comparisons calculate the sum of all the enjoyment level comparisons between A and B-versions in enjoyment types, such as immersion (Swimming) and Surprise factor (Seeking and finding), across all the result tables. For each comparison, the version which had a higher enjoyment level gets a point for that comparison, while ties give points to neither and are calculated on their own. For the combined comparisons the calculations work the same way, but this time only combined results, such as "immersion + general enjoyment" (Swimming) are taken into account. The only exception being the Crawling and climbing puzzle, which only had one type of enjoyment, which will be included in both results in the same way.

The most enjoyable mechanic

The most enjoyable mechanic seems to have been the more realistic one (version A), as it got higher combined and individual average enjoyability results in all of the three puzzles in each comparison category, as displayed in tables 72 and 73.

It is also worth noticing, that the differences between the individual and combined ratings are in some cases quite vast, making the combined results less reliable, as just focusing on them can make the enjoyment differences appear a lot more or less radical than they actually were.

All in all, the total amount of preferred categories for the individual comparisons for version A were 232/372, which is around 62.4% of the total, while version B only received a preference level of 70/372 or 18.8%, which leaves 70 or 18.8% ties. For the combined comparisons, the numbers reflected even higher enjoyment difference, with version A gaining 135/197 or around 68.5% of the total preference, leaving B with just 36 or 18.3% in addition to 26 or 13.2% ties.

The highest enjoyment differences between versions were found in the Swimming puzzle, with version B only having a 11.7% preference for the individual and 16.4% for the combined comparisons, compared to version A's 80 and 80.8 percentages.

As a contrast, the Seeking and finding puzzle had the closest enjoyment ratings between versions, 47.9% (A) to 26.1% (B) in the individual comparisons, although the combined ones were less close (72.9% to 18.8%), highlighting the unreliability of combining results. Another interesting fact related to the Seeking and finding results is, that one of the categories was almost tied, the tester based category in the individual results, which scored a huge number of 27/78 ties, which is more than a third of the total amount of individual comparisons in the category. This result indicates that maybe the correlation between the testers' measured traits and preferred version is not very significant or at least not very clear in the Seeking and finding puzzle.

Additionally, the general enjoyment statistics across all testers (calculated as combined averages from tables 19, 43, and 62) for every puzzle showed that the Seeking and finding puzzle version A was the most enjoyable (3.18), followed by Crawling and climbing A (3), and Swimming A (2.96). The B-versions had identical order, which was numberwise 3, 2.85, and 2.46, and when the versions were combined, the order naturally stayed the same, resulting into 3.09, 2.93, and 2.71 respectively. This result means that the sample as a whole preferred A-versions to Bs and, additionally, that the Seeking and finding puzzle was the most enjoyable one, even sharing the second most enjoyable spot with its B-version, while the Swimming puzzle was the least enjoyable. This result acts as an example to show that while the results among all testers can tell a lot about the sample, they do not necessarily reveal anything about the puzzles individually, as the other results showed us how much the different qualities of the testing, testing order and the testers themselves affect the results in a more individual scale. However, this result does confirm the fifth hypothesis about testers not rating any of the puzzles highly on average, as most puzzle version had a higher enjoyment level than the half-way point (2.5).

Table 72: The individual result categories for all puzzles, displays the total numbers for how many times one version was preferred over the other in terms of enjoyability, but only considers the individual levels of enjoyment like immersion for the Swimming puzzle, leaving out combined enjoyment results. Each comparison which had at least 1 tester for both versions is considered in these results.
Abbreviations:
SP = Swimming puzzle,
CC = Crawling and climbing puzzle,
SEF= Seeking and finding puzzle,
RC = Result category, tells which result category the results are from
RB = Realism based category, has results which compare perceived realism with enjoyability,
OB = Order based category, has results which compare the testing order and enjoyability,
DB = Difficulty based category, has results which compare perceived difficulty with enjoyability,
TB = Tester based category, has results which compare the testers' personal traits and abilities with enjoyability,
ALL = All categories combined,
P/T V = Preferred/Total, highlights how many times the specified version V was preferred out of all the individual enjoyment level comparisons in this category.

Puzzle	**RC**	**P/T A**	**P/T B**	**Ties/T**
SP	RB	15/22	4/22	3/22
SP	TB	64/80	8/80	8/80
SP	OB	17/18	0/18	1/18
SP	DB	27/34	6/34	1/34
SP	ALL	123/154	18/154	13/154
CC	RB	7/10	2/10	1/10
CC	TB	18/39	6/39	15/39
CC	OB	5/9	2/9	2/9
CC	DB	11/18	5/18	2/18
CC	ALL	41/76	15/76	20/76
SEF	TB	33/78	18/78	27/78
SEF	OB	12/27	9/27	6/27
SEF	DB	23/37	10/37	4/37
SEF	ALL	68/142	37/142	37/142
ALL	**ALL**	232/372	70/372	70/372

Table 73: The combined result categories for all puzzles, displays the total numbers for how many times one version was preferred over the other in terms of enjoyability, but only considers the combined levels of enjoyment (e.g., immersion + general enjoyment for the Swimming puzzle), leaving out the individual enjoyment results. For the crawling puzzle there was only one enjoyment level, so the results are identical to Table 72. Each comparison which had at least 1 tester for both versions is considered in these results.
Abbreviations:
SP = Swimming puzzle,
SEF= Seeking and finding puzzle,
RC = Result category, tells which result category the results are from
RB = Realism based category, has results which compare perceived realism with enjoyability,
OB = Order based category, has results which compare the testing order and enjoyability,
DB = Difficulty based category, has results which compare perceived difficulty with enjoyability,
TB = Tester based category, has results which compare the testers' personal traits and abilities with enjoyability,
ALL = All categories combined,
P/T V = Preferred/Total, highlights how many times the specified version V was preferred out of all the individual enjoyment level comparisons in this category.

Puzzle	**RC**	**P/T A**	**P/T B**	**Ties/T**
SP	RB	4/7	3/7	0/7
SP	TB	32/40	6/40	2/40
SP	OB	9/9	0/9	0/9
SP	DB	14/17	3/17	0/17
SP	ALL	59/73	12/73	2/73
CC	RB	7/10	2/10	1/10
CC	TB	18/39	6/39	15/39
CC	OB	5/9	2/9	2/9
CC	DB	11/18	5/18	2/18
CC	ALL	41/76	15/76	20/76
SEF	TB	18/26	5/26	3/26
SEF	OB	6/9	3/9	0/9
SEF	DB	11/13	1/13	1/13
SEF	ALL	35/48	9/48	4/48
ALL	**ALL**	135/197	36/197	26/197

7 Conclusion

According to all the results considered in all of the puzzles combined, and when taking into account observations and even some of the response biases, the more realistic mechanics came out ahead by a landslide as concluded in Section 6.7, with total preference ratings for the individual enjoyment comparisons being (illustrated in Table 72) 62.4% for A-versions with just 18.8% for B-versions with the same amount of ties. When combining the individual enjoyment comparisons the result was even clearer, with A-versions gathering 68.5% of the preference, leaving B-versions with 18.3% in addition to 13.2% ties. Additionally, the sample as a whole preferred the A-versions in all cases (seen in tables 19, 43, and 62), and as the more realistic mechanics were also perceived to be quite realistic by the sample (as shown in tables 20 and 42), the result means that according to this study, it is more enjoyable to solve puzzles in virtual reality with more realistic core mechanics.

In relation to the testing sample utilised in the study, the testers seemed to be at least somewhat reliable sources of information, as their responses were mostly consistent on duplicated questions (as noted in the analysis of Table 42), although the response biases identified in Section 6.1.1 most definitely had an impact on the reliability. Additionally, there was healthy variance in most of the real-life abilities compared with in the experiment, however, there was one quality which the sample lacked: the phobias of interest (as seen in Table 6), and due to this the first hypothesis about the effects of those phobias on the enjoyment remains unexplored. As a sidenote, the claustrophobia did come up for a few testers, but it happened during testing the Swimming puzzle, unlike the design which planned it for the Crawling and climbing.

When considering the individual enjoyment categories, the more realistic mechanics were clearly favoured by the testers who had more previous gaming experience (shown in tables 25, 51, and 64), the so called "hardcore half" (HC), while the B-versions also had higher or equal preference among the HC testers, when compared to the more "casual half" in most cases. This partly confirms the second hypothesis from Section 5.2, which was that testers with more gaming or VR experience would prefer all the puzzles more than the less experienced ones, even though the

VR experience did not appear to provide similar results, apart from the Swimming puzzle where increase in the previous experience strictly corresponded to increased enjoyment and the preference of the A-version.

In addition to the previous results, a major interest of the study was to find out whether testers' real-life abilities had anything to do with the enjoyment levels, and if possessing skills among the tested mechanics would affect the version preference. The major tools for this were measuring the correspondence between testers perceived swimming, diving and climbing skills, in addition to athleticness, and the results indicate that the real-life swimming and diving skills did have an impact on the enjoyment levels of the Swimming puzzle (can be seen in tables 28 and 29).

The impact seems to have been that higher perceived swimming or diving abilities seemed to amount to having higher general enjoyment levels for version A on average. For version B, the general enjoyment was lower for more skilled testers, indicating that the perceived lesser realism had a negative impact on those who knew what to expect. This was further confirmed with the immersion statistics being lower in version B for the more experienced testers. Additionally, there were, quite naturally, only very few master level swimmers and divers, and all of them rated their Swimming puzzle enjoyment to be among the highest in each category, which could be a telltale sign of the response bias of demand characteristics (see Section 6.1.1), as it is possible that the master level swimmers and divers thought that as the puzzles are designed to simulate real-life actions which they master, they are expected to enjoy them because they enjoy the actions in the real life so much.

The climbing skill and athleticness however, did not seem to have any clear correlations with the enjoyment levels for this sample, as was seen in tables 30, 49, and 50. This causes the third hypothesis about more athletic testers preferring the more realistic versions to be disproven.

Additionally, the importance of the testing order was evaluated in this experiment, the result showing that being familiar with the puzzle environment beforehand, meaning testing another version of the same puzzle first, seems to significantly increase the performance in the second version (as in decreasing the completion time), despite any other characteristic that the testers may have, like age or previous gaming experience (results showcased in tables 31, 54, and 67. This result was completely unanimous across the three puzzle types, which leads to believe it was not caused by coincidence, especially as completion time is not even something that is decided based on opinion like many other results in this thesis, it is a calculated fact.

Also, other results considering the testing order showed that in the Swimming and the Crawling and climbing puzzles, the version which was tested as first, albeit taking longer to complete, was perceived to be more enjoyable than the second (displayed in tables 32 and 53). This result indicates that while knowing the puzzle environment beforehand boosts performance, it seems to also decrease enjoyment, probably because the novelty value decreases simultaneously with experience gained. The result may sound obvious, but for some reason it is not unanimous, as for the Seeking and finding puzzle it was the version which was tested as second which appears to have been slightly more enjoyable, reasons for that unknown.

In addition to the previous, the fourth hypothesis, which was related to the testing order, namely that testers would find the puzzles tested earlier more enjoyable than later ones, was not confirmed in the slightest, as while for Crawling and climbing and Seeking and finding the favourite order of testing was either as first or second, for the Swimming puzzle testing as third was the ultimate favourite. This result means that other factors, supposedly at least gained VR experience during the testing and the lack of experienced VR sickness, had a larger impact on the enjoyment than was expected.

There was also a fifth hypothesis about the testers not rating any of the puzzles highly on average, because of the lack of aesthetics and story diminishing their experience. This hypothesis can be disproven, as most of the puzzle versions got above average (2.5) enjoyment results across the testing sample as a whole (as discussed in Section 6.7). This probably relates to the testers understanding of what to expect in the testing, although there was one lenghty feedback about how the poor visuals ruined the experience of (at least) one tester.

The final major correspondence was discovered in the category of familiarity with the testing supervisor, and suggests that the enjoyment levels were indeed higher for testers who were more familiar with the testing supervisor (6th hypothesis), compared to those who were complete strangers (see tables 24, 48), and 63. This result confirms the sixth hypothesis and is a tremendously interesting one, especially when the result was almost completely unanimous among the three puzzles. In detail, the result means that testers who were either vaguely or closely familiar with the testing supervisor, preferred whatever versions they tested, more, than the testers who had not met with the supervisor beforehand.

When it comes to the difficulty aspect (analysed in sections 6.4.5, 6.5.5, and 6.6.4) the results were partly quite expected: testers seemed to enjoy challenge (which is subjectively defined) but the enjoyment levels went down when the puzzles were perceived to be too difficult or too easy. In addition to challenge, the effects of setbacks had varying results: in some cases they did not seem to have an effect on the enjoyment levels at all, while in others the extreme cases (very few and very many setbacks) affected either negatively or positively on the enjoyment level. Additionally, the sense of control and intuitivity of the mechanics had an overwhelming effect on the enjoyment, meaning that the more control over the mechanics and the easier they were to use, the higher the enjoyment levels were, which was quite an obvious result.

All in all, as already discussed in Chapter 5, the results are not generalisable, as the amount of testers in the study was low (22) and the sample did not reflect the general population very well, as the testers were acquired using convenience sampling, which causes the results to only apply to male Finnish tertiary students (previous or current) living in the Turku region in Finland.

7.1 Discussion

The results comparing the testing order with the completion time showed that being familiar with the puzzle environment beforehand, meaning testing another version of the puzzle first, seems to significantly increase the performance in the second version, despite any other characteristic that the testers may have. It would be interesting for another researcher to attempt to recreate this in some way and see whether they would get similar results and try to find out why the effect is so seemingly massive, if it truly is so. After all, the puzzle versions were just *similar*, not identical, so even though the previous knowledge about the *type* of the puzzle is certain to reduce at least misunderstanding the rules, it still does not give many straight answers to the puzzle itself.

Additionally, it would be of major interest to see whether the phobias of the sample would have a statistically significant effect or not, although in that case the study probably would have to require testers with the phobias in question to participate, which would again skew the results as the sample would not reflect the whole population that well anymore. Most likely the only way to truly test this would be to have a large enough sample to try to ensure a considerable amount of testers with phobias to be part of the experiment, although the sample probably would have to be so large that the study might be quite overwhelming for a single researcher to conduct.

Overall, this experiment can be considered a success, as it reached an overwhelming conclusion about the realism of the VR mechanics, although a completely different problem is to evaluate how well those mechanics were created in the first place. A hint related to this aspect can be seen when comparing the final statistic, the future vision of VR (the results before testing displayed in Table 13), which was the only question asked in both the pre- and post-questionnaires. Unfortunately, the experiment did not change any tester's view from hopeless to hopeful or the other way around, but the two testers who were yet to make up their minds did rate that the experience had a positive impact on their view, despite VR not completely gaining their trust. The process of evaluating the design of the mechanics is another aspect that maybe some interested researchers would like to tackle in the future.

References

[1] S. Yamashita, X. Zhang, and J. Rekimoto, "Aquacave: Augmented swimming environment with immersive surround-screen virtual reality," in *Proceedings of the 29th annual symposium on user interface software and technology*, pp. 183–184, 2016.

[2] H. Lee, M. Moon, T. Park, I. Hwang, U. Lee, and J. Song, "Dungeons & swimmers: Designing an interactive exergame for swimming," in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, UbiComp '13 Adjunct, (New York, NY, USA), p. 287–290, Association for Computing Machinery, 2013.

[3] J. Wideström, A.-S. Axelsson, R. Schroeder, A. Nilsson, I. Heldal, and r. Abelin, "The collaborative cube puzzle: A comparison of virtual and real environments," in *Proceedings of the Third International Conference on Collaborative Virtual Environments*, CVE '00, (New York, NY, USA), p. 165–171, Association for Computing Machinery, 2000.

[4] S. Benford, C. Greenhalgh, G. Reynard, C. Brown, and B. Koleva, "Understanding and constructing shared spaces with mixed-reality boundaries," *ACM Transactions on computer-human interaction (TOCHI)*, vol. 5, no. 3, pp. 185–223, 1998.

[5] Virtual Reality Society, "History of virtual reality." `https://www.vrs.org.uk/virtual-reality/history.html`, 2017. (Accessed on 14.9.2019).

[6] P. Américo, "Battle of Avay." `https://images8.alphacoders.com/494/494481.jpg`.

[7] D. Brewster, *The Stereoscope; Its History, Theory and Construction, with Its Application to the Fine and Useful Arts and to Education.* John Murray, 1856.

[8] E. A. Link, 1937. U.S Patent 2,099,857.

[9] M. L. Heilig, "Sensorama simulator." `http://www.mortonheilig.com/SensoramaPatent.pdf`, Aug. 1962. U.S Patent 3,050,870 (Accessed on 25.9.2019).

[10] I. E. Sutherland, "A head-mounted three dimensional display," in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pp. 757–764, ACM, 1968.

[11] M. W. Krueger, T. Gionfriddo, and K. Hinrichsen, "VIDEOPLACE-an artificial reality," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 35–40, 1985.

[12] P. Luckey, "Oculus rift." https://upload.wikimedia.org/wikipedia/commons/thumb/0/08/Oculus-Rift-CV1-Headset-Front_with_transparent_background.png/1920px-Oculus-Rift-CV1-Headset-Front_with_transparent_background.png, 2016.

[13] Valve and HTC Corporation, "HTC Vive." https://cdn.arstechnica.net/wp-content/uploads/2016/05/jt3a0498-1.jpg, 2016.

[14] "TPCast Wireless Adapter for Vive." https://www.tpcast.cn/h_en/index.html.

[15] "Varjo Technologies Oy." https://varjo.com/.

[16] © 2018 OppenheimerFunds Distributor, Inc., "Investing in the Soaring Popularity of Gaming." https://www.reuters.com/sponsored/article/popularity-of-gaming?utm_source=reddit.com. Accessed on 17.4.2020.

[17] E. Adams, *Fundamentals of Game Design, Third Edition*. Pearson Education Inc., 2014.

[18] "Virtuix Omni." https://www.virtuix.com/. Accessed on 30.3.2020.

[19] "Beat Saber." https://beatsaber.com/. Accessed on 10.4.2020.

[20] Elon Musk, "Neuralink." https://neuralink.com/.

[21] The Stonefox (Github username), Extend Reality Ltd, "Virtual Reality Toolkit version 3.3.0." https:

//github.com/ExtendRealityLtd/VRTK/releases/tag/3.3.0, Feb. 2019.

[22] R. Likert, “A technique for the measurement of attitudes.,” *Archives of psychology*, 1932.

[23] D. M. Tice, J. L. Butler, M. B. Muraven, and A. M. Stillwell, “When modesty prevails: Differential favorability of self-presentation to friends and strangers.,” *Journal of personality and social psychology*, vol. 69, no. 6, p. 1120, 1995.

[24] C. Bergland, “Neuroscientists Discover the Roots of "Fear-Evoked Freezing".” https://www.psychologytoday.com/us/blog/the-athletes-way/201405/neuroscientists-discover-the-roots-fear-evoked-freezing, 2014. Accessed on 20.10.2019).

[25] A. L. Edwards, “The social desirability variable in personality assessment and research.,” 1957.

[26] S. Messick and D. N. Jackson, “Acquiescence and the factorial interpretation of the MMPI.,” *Psychological Bulletin*, vol. 58, no. 4, p. 299, 1961.

[27] Cabooter, Elke and Millet, Kobe and Weijters, Bert and Pandelaere, Mario, “The ‘I’in extreme responding,” *Journal of Consumer Psychology*, vol. 26, no. 4, pp. 510–523, 2016.

[28] A. Blankenship, “Psychological difficulties in measuring consumer preference,” *Journal of Marketing*, vol. 6, no. 4_part_2, pp. 66–75, 1942.

[29] M. T. Orne, “On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications.,” *American psychologist*, vol. 17, no. 11, p. 776, 1962.

[30] J. Kruger and D. Dunning, “Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments.,” *Journal of personality and social psychology*, vol. 77, no. 6, p. 1121, 1999.

A Puzzle instructions

Swimming puzzle version A

Task:

- In this puzzle the goal is to find a key object (easily visible when found) and GRAB it after finding it.
- Navigate in the water using a light attached to your (imaginary) forehead and find the hidden path(s).
- After reaching an ending please note down (on provided paper) the completion time and how many times you drowned (these will be shown to you after reaching the ending).

Controls:

- GRAB buttons can be found on the both sides of the controllers and they need to be pressed usually either with the palm of your hand or your thumb (test that you found them).
- GRAB a hold of the water while inside it and then DRAG yourself towards a direction in order to move. Be quick as grabbing only last for a short time until forced release.
- Alternate hands in swimming motion (while grabbing and being released) in order to swim faster and rise up against the force of gravity.
- Holding down the TOUCHPAD towards any direction will also move you slowly towards that direction (necessary when not in water). Hold down both in order to increase speed.
- Entering any walls will change your vision to blue and in some cases teleport you away from the wall. This should only occur when moving in real life room space. (Move back in order to revert)

- In order to CLIMB objects (like a rope for example) press and hold the GRAB button when your hand is touching the climbable object until your other hand grabs another climbable object. Releasing the GRAB button will stop grabbing the object.

Fail conditions:

- You have a limited amount of oxygen, which is used up while underwater and refreshed upon reaching a surface.
- When oxygen is getting low, a sound indicating possible drowning starts to play, it is recommended to head to the surface as soon as possible when hearing it.
- Another indicator of low oxygen is that the view gets gradually darker.
- If you drown, you will be taken back to the beginning.

Swimming puzzle version B

Task:

- In this puzzle the goal is to find a key object (easily visible when found) and GRAB it after finding it.
- Navigate in the water using a light attached to your (imaginary) forehead and find the hidden path(s).
- After reaching an ending please note down (on provided paper) the completion time and how many times you drowned (these will be shown to you after reaching the ending).

Controls:

- Touching TOUCHPAD and moving your finger on it will move you in the direction of your finger movement (pressing down is not needed).

- In order to TELEPORT, press and hold the TRIGGER, which can be found under the TOUCHPAD (on the other side of the controller) and aim to a location. A valid location is indicated by the colour of the teleport beam (RED indicates invalid). After selecting a location and while holding the TRIGGER continuously, press the TOUCHPAD down (once) and a teleport should occur.
- Inside water TOUCHPAD movement works and if close to a floor (any horizontal surface) , TELEPORTING is possible.
- GRAB buttons can be found on the both sides of the controllers and they need to be pressed usually either with the palm of your hand or your thumb (test that you found them).
- In order to get back to the surface, press one of the GRAB buttons and you will start slowly rising upwards. Press the TOUCHPAD down (once) in order to let gravity take over again.
- If you reach the surface you will stay on the surface until the TOUCHPAD is pressed and during this time it is possible to move on top of the surface by touching the TOUCHPAD (not pressing).
- In order to CLIMB objects (like a rope for example) press the GRAB button when your hand is touching the climbable object (holding down is not required. Press the button again in order to ungrab or grab some other object with your other hand.

Fail conditions:

- You have a limited amount of oxygen, which is used up while underwater and refreshed upon reaching a surface.
- When oxygen is getting low, a sound indicating possible drowning starts to play, it is recommended to head to the surface as soon as possible when hearing it.

- Another indicator of low oxygen is that the view gets gradually darker.
- If you drown, you will be taken back to the beginning.

Crawling and climbing puzzle version A

Task:

- In this puzzle the goal is to determine a correct path by choosing out of 3 options each time. The correct path is formed by connecting words and sounds.

- Listening to the sounds in the puzzle is also necessary in order to determine the correct path.

- The first choice cannot be incorrect.

- The movement happens via crawling and climbing in VR.
  After reaching an ending please note down (on provided paper) the completion time, how many times you were reset to the beginning and the ending KEY (these will be shown to you after reaching the ending).

Controls:

- GRAB buttons can be found on the both sides of the controllers and they need to be pressed usually either with the palm of your hand or your thumb (test that you found them).
- In order to CRAWL, press the GRAB button once when touching a black square on the floor and drag yourself towards another black square. As the goal of this puzzle is to simulate crawling, the easiest way to accomplish this is to be in a crouching/sitting position on the (real-life) floor. The square you are holding on to will automatically

release your hand if you grab another square or if your head is too far away from your grabbing hand. Manual release is not possible in order to prevent throwing yourself forward.

- In order to CLIMB the walls, press and hold down the GRAB button when your hand is touching the climbable object (black square or ladder piece). Releasing the GRAB button (or grabbing another one with the other hand) will stop grabbing the object.

Fail conditions:

- Choosing an incorrect path combination (only checked after all the choices have been made) will reset you to the beginning once reaching the end.

Crawling and climbing puzzle version B

Task:

- In this puzzle the goal is to determine a correct path by choosing out of 3 options each time. The correct path is formed by connecting words and sounds.

- Listening to the sounds in the puzzle is also necessary in order to determine the correct path.

- The first choice cannot be incorrect.

- The movement happens via crawling and climbing in VR.
  After reaching an ending please note down (on provided paper) the completion time, how many times you were reset to the beginning and the ending KEY (these will be shown to you after reaching the ending).

.

Controls:

- In order to move, press and hold down the TOUCHPAD towards the direction of desired movement. Press and hold down both for increased speed.
- In order to move inside the tunnels (CRAWL), an amount of crouching is needed to fit in.
- GRAB buttons can be found on the both sides of the controllers and they need to be pressed usually either with the palm of your hand or your thumb (test that you found them).
- In order to CLIMB the walls, press the GRAB button (no need to hold down) when your hand is touching a climbable object (black square or ladder piece).
- Pressing the GRAB button again or grabbing another object will stop grabbing the object.

Fail conditions:

- Choosing an incorrect path combination (only checked after all the choices have been made) will reset you to the beginning once reaching the end.

Seeking and finding puzzle version A

Task:

- In this puzzle the goal is to find 2 keys hidden somewhere in the room and deliver them to their indicated positions.
- It is possible to go inside some of the objects and it is required.

- After reaching an ending please note down (on provided paper) the completion time (this will be shown to you after reaching the ending).

Controls:

- In order to move, press and hold down the TOUCHPAD towards the direction of desired movement. Press and hold down both for increased speed.
- In order to CLIMB the scene objects, press and hold down the GRAB button when your hand is touching any object other than the outer walls or ceiling and drag yourself to a direction.
- Releasing the GRAB button or grabbing another object will stop grabbing the object.

Fail conditions:

- There is no way to fail this task, other than giving up.

Seeking and finding puzzle version B

Task:

- In this puzzle the goal is to find 2 keys hidden somewhere in the room and deliver them to their indicated positions.
- It is NOT possible to go inside any of the objects and it is NOT required.
- After reaching an ending please note down (on provided paper) the completion time (this will be shown to you after reaching the ending).

Controls:

- In order to move, press and hold down the TOUCHPAD towards the direction of desired movement. Press and hold down both for increased speed.
- In order to CLIMB the scene objects, press the GRAB button (no need to hold down) when your hand is touching any object other than the outer walls or ceiling and drag yourself to a direction.
- Pressing the GRAB button again or grabbing another object will stop grabbing the object.

Fail conditions:

- There is no way to fail this task, other than giving up.

Additionally, the testers were provided with a picture of the HTC Vive -controller which had all the necessary inputs mapped in. This mapping is displayed earlier in the thesis in Figure 9.