The computer games journal manuscript No.
(will be inserted by the editor)

Exploring virtual reality mechanics in puzzle design

Taneli Nyyssönen · Jouni Smed

Received: date / Accepted: date

Abstract We explore 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 objects in virtual spaces. Each mechanic has two different variations: realistic and game-like, and the main goal is to test which variation is more enjoyable to use. Convenience sampling is used in the study and the sample size is 22 volunteers. Both qualitative and quantitative data are collected. The data collection methods used are questionnaire and observation. The enjoyability of the mechanics is 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 is to observe whether real-life skills corresponding to the studied mechanics affect the enjoyment and performance levels in the respective mechanics. The more realistic mechanics turn out to be more enjoyable by a significant margin, suggesting that they should be utilised in the future. Additionally, the real-life diving and swimming skills, puzzle variation testing order, previous gaming experience, and testers' familiarity with the testing supervisor are identified to have a clear impact on the enjoyment levels.

Keywords Virtual reality · Puzzle · Game design · Experimental · Mechanics · Interaction design

University of Turku, Department of Computing
FI-20014 Turku, Finland
Tel.: +358 29 450 5000
E-mail: ttjnyy@protonmail.com
E-mail: jouni.smed@utu.fi

1 Introduction

This article focuses on analysing custom virtual reality (VR) puzzles based on test subject performance and feedback from the viewpoint of game design (Nyyssönen (2020)). The main goal is to determine whether solving virtual reality puzzles with more realistic core mechanics is more enjoyable than utilising a game-like approach. Additionally, we look into other aspects which could affect the enjoyment, including 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 cannot be generalised because of the small sample size, these observations could act as a basis for further research in the field.

A large motivating factor for this research is its novelty: there are few publications related to evaluating the effects of realism of VR game mechanics, and most of that research is based on comparing real-life activities with virtual ones, not alternative VR mechanics. Existing publications related to the topic are, e.g., *AquaCAVE* (Yamashita et al. (2016)), and *The collaborative cube puzzle: a comparison of virtual and real environments* (Wideström et al. (2000)). The former is an augmented swimming environment combined with immersive surround-screen in VR, which is essentially a real-life swimming simulator enhanced by a virtual environment, while the latter investigates the differences between solving a puzzle-cube similar to a Rubik's Cube in VR and in the real world.

The mechanics studied in our work include VR-simulated swimming, crawling, and climbing. Additionally, we compare different hiding locations for objects inside virtual environments. The mechanics are tested as a form of AB-testing, where testers get to test similarly themed puzzles with two different types of movement or other mechanics.

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

The basic concepts of virtual reality and games are explained in detail by Nyyssönen (2020), while this article begins by explaining some of the advantages and challenges of designing games for VR, after which the puzzles developed for this thesis are presented alongside with some of the testing methodology and results. Finally, the study is concluded with discussion about possible extension for the research in question.

2 Designing games in virtual reality

Game design refers to the thought processes behind most aspects of 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 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. That is why game design is not *just* engineering or art, it is both, it is a *craft*. (Adams (2014))

When it comes to designing for VR, the game designer has an additional set of challenges to overcome when compared to more conventional games. The following sections discuss some of the main advantages and disadvantages.

2.1 Player-controlled camera

In VR games, it is common to use the first-person camera model with an avatar-based interaction model, although the "avatar" is rarely visible, except for its hands, as the avatar is the player themself. The puzzles in our work also utilise the first-person model combined with an avatar-based interaction model, where the only parts of the avatar the player can see are its hand models. The reason behind this is that it would be difficult to realistically simulate the mechanics otherwise.

This brings us to one major aspect about VR in comparison to conventional games: the camera movement. In conventional games, the developers need to plan the restrictions for the camera movement which determine the parts of the game world the player will be able to see. In VR games, this problem does not exist in a similar way as the player has autonomy over the camera movement by default. This in turn creates opportunities for VR game developers, for example, they can hide information or objects in locations which in conventional games would be too clearly indicated to the player by the camera movement restrictions.

For example, in a conventional game where a player is instructed to search for a key, the player is restricted by the camera movement options which reduce the possibilities for the hiding location of the key. Moreover, the player could test whether they are able to enter openings or go under surfaces in order to rule those areas out of being possible key locations. As a contrast, in VR it is possible to enter anything that the player's virtual head fits through, creating near infinite possible hiding locations for objects without revealing any hints of their position (the Seeking and finding puzzle version A, see Section 3.4, 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 a heightened sensitivity to input lag and potential nausea from rapid camera movements in some VR game activities.

2.2 Designing movement

Creating realistic movement inside a virtually created environment poses challenges for game developers, and one big reason for this is that unlike the room where the playing is taking place, a virtual space is technically endless. In order to attempt solving this problem, several different types of VR-movement have been created. Some of these types are room-scale, locomotion, and teleportation. For a broader review, see (Nyyssönen (2020), pp. 16-17).

2.3 Design challenges

In VR, the player is "mapped to themself", meaning that their real-life actions mirror the actions of their character inside VR, thus they *are* the character. In comparison, in conventional games, the player *plays as* another character or entity. This means that in VR, the player's natural aspects, especially height and posture, are fully present. This causes problems, for example, when the player is required to reach out or into something, as different players have different reaching capabilities. The problem is unique to VR, as in conventional games the character's height and all other aspects are set by the developers. Whereas for VR games, the player height is immutable, and rather than trying to change it (which would cause a major disconnection between the player and the game world), the games have to change to accommodate this "physical" challenge.

One solution to the innate player height problem in VR game design, 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 (see sections 3.3 and 3.4), although it can cause the tallest of players to have a (slightly) less challenging experience, which could have an effect on their enjoyment level. Another solution is 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.

On the subject of "physical" design challenges, VR is infamous to cause motion sickness (or simulator sickness), which the design of VR-games (and other applications) have to take into account. The main cause for the motion sickness is that the user's brain is not 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 referred to as balance), that we are not moving. Simultaneously we receive sensations from our proprioception system (sense of our skin and muscles to determine our limb positions), which can add to the sensory conflict. (Ebenholtz (1992), Kim et al. (2018))

Some design solutions for this challenge are to avoid moving the camera without initiation by the player, avoiding acceleration, minimizing the input lag, and fading the player's field of view when necessary, for example, during teleportation transitions (this is utilised in the teleportation mechanic used in the Swimming puzzle version B, see Section 3.2). Additionally, it may be possible to ease the confusion caused to the player's brain by visualising the movement trajectory in some way, although this has produced varying results. (Bonato et al. (2008), Fernandes & Feiner (2016))

On the topic of "mental" challenges, the first that often causes serious issues is handling virtual wall collisions; the player needs to be restricted by the virtual environment's boundaries somehow. One way to answer this problem is by making it impossible for the player to reach any walls, although this limits the possible design space. Another common solution, is to teleport the player a small distance away from the entered wall inside the virtual environment, artificially enforcing the boundary. While this solution can be effective for handling virtual wall collisions, it does not solve the issue of motion sickness and it can break the immersion.

The second "mental" challenge is related to grabbing virtual objects and how to make that both as functional and immersive as possible. There are essentially 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 is usually more functional. Using the first option can 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 centre of mass, and the parts of the hand(s) currently touching it. The second option then again can cause problems with interacted objects ending up inside walls or other objects, as the grabbed positions of objects are fixed to specific locations in relation to the controllers (hands) which pick them up, ignoring the surrounding boundaries.

The third challenge is related to the second, specifically about interaction with objects which need to move slower than the hand that is interacting with them, such as heavy doors. In conventional games, these interactions are often handled via toggling the interaction on and off while the interaction itself happens with an animation. For VR, the interaction is completely controlled by the player from start to finish, and while it is possible to have rotators (e.g., doors) follow the player's hand movements precisely, it requires giving up friction completely, causing the rotators to function unrealistically. The more realistic option is 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 create the feeling of friction too well, as in real life heavy objects need more force, not less.

3 Practical application

We have created and tested three different VR puzzles, all of which have two variations. 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 (The Stonefox (Github username), Extend Reality Ltd (2019)). In this section we present the basic functionalities, namely the player and the tracking system, followed by the three puzzles, their mechanics and variations.

3.1 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.

In these puzzles the VR-device used was the original HTC Vive. In total it consists of two base stations utilising the *Lighthouse*-technology, which detect the player's headset and controllers using infrared light, two controllers, and the actual headset. The tracking system combined with the headset and controllers is called *SteamVR*. For more details on the setup and controls, see (Nyyssönen (2020), p. 24).

The player itself is formed by three different main colliders: head, body, and feet. The head collider's position is calculated based on the real-life location of the headset, measured via physical "lighthouses" (base stations) which track the VR-headset and controllers. 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. The colliders and their dependencies are all created by the VRTK and displayed in Figure 1.

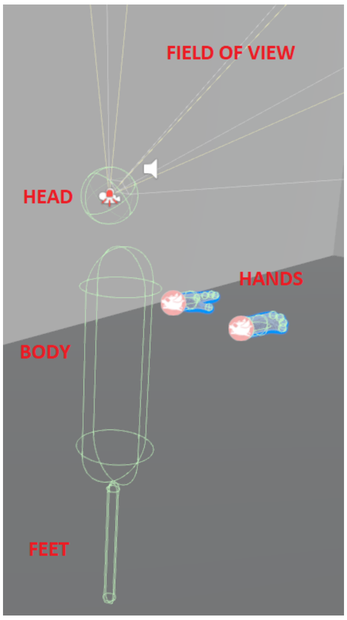

Fig. 1 Player's collider composition, consisting of the hand, head, body, and feet colliders.

3.2 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 locate a key object at the end of the path. Completing the puzzle also involves climbing a 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 the 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 virtual forehead in addition to lanterns located in various parts of the puzzle. The lanterns light up when approached by the player, making it easier to locate themselves in the otherwise dark water, while simultaneously creating a sense of progression. Additionally, the water areas contain bubbles which move upwards to help players retain a sense of movement. The puzzle overview can be seen in Figure 2.

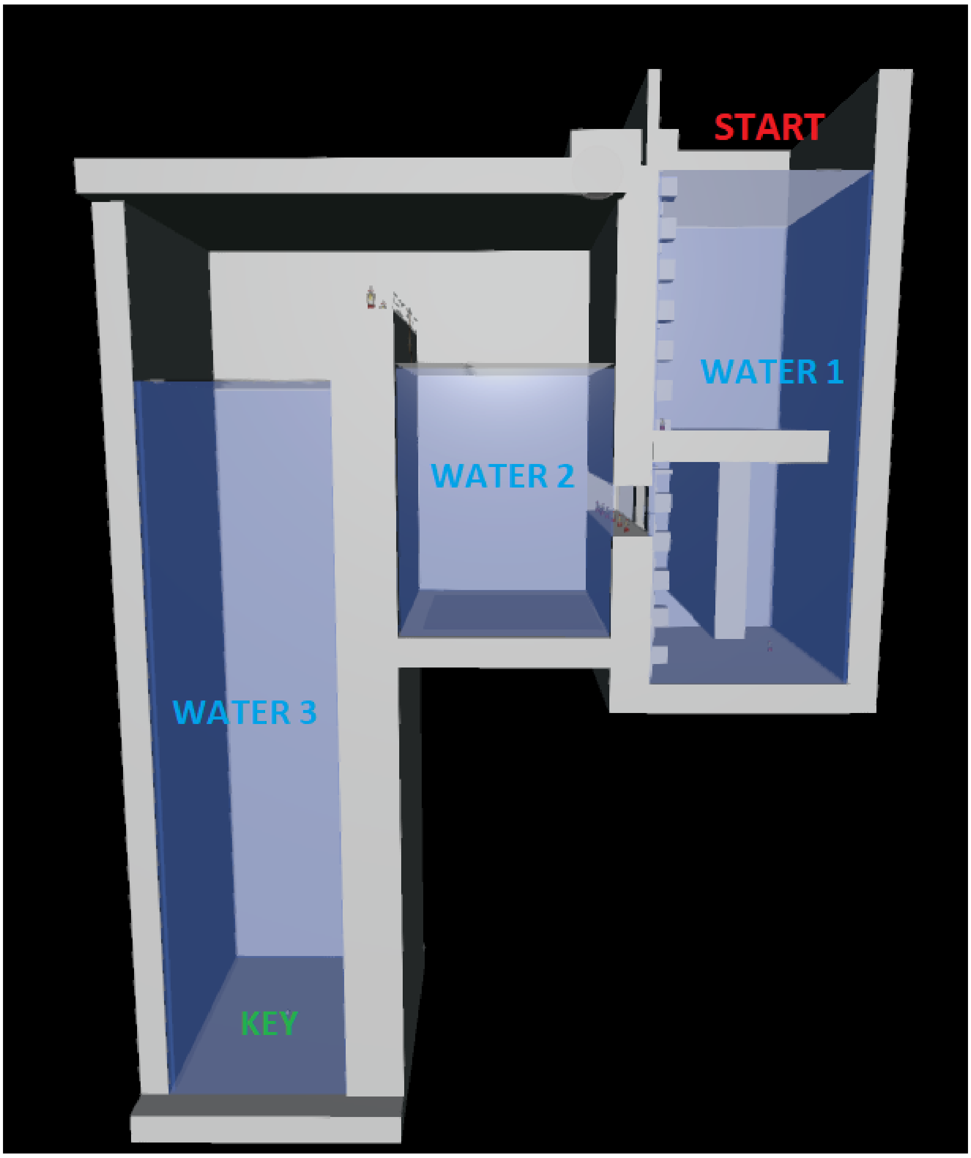

Fig. 2 The overview of the Swimming puzzle, consisting of the three water areas, start location (in red) and the key (in green).

The puzzle has two variations: the more realistic version A, and the less realistic version B, which utilise different swimming mechanics, which are thoroughly explained by (Nyyssönen (2020)). The puzzles themselves are otherwise almost identical, although the hidden path that the player needs to find is altered slightly so that 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.

3.3 Puzzle 2: Crawling and climbing

This puzzle aims to experiment with two different ways of how crawling and climbing could be performed in VR. The puzzle consists of alternating tunnels and walls, with three tunnels to choose from each time as seen in Figure 3. 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 that the player should choose.

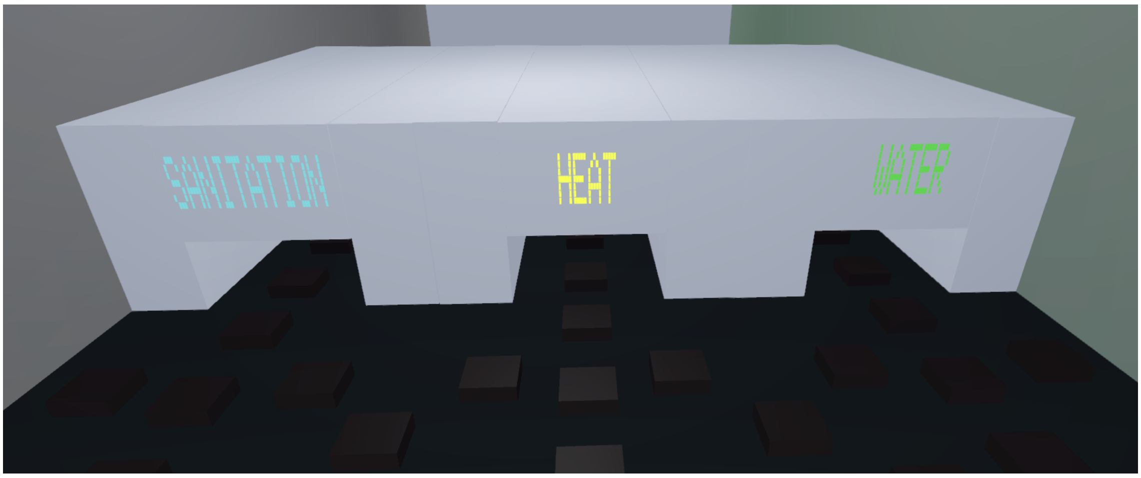

Fig. 3 One of the path choices the player has to make in the Crawling and climbing puzzle. The floor contains climbable squares for the A-version and each of the three paths is labelled with a keyword (sanitation, heat, water).

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. Although, the first choice is designed 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 low also because testing one path takes quite a lot of time and effort, especially with 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 emerging victorious, or they will be sent back to the beginning to redo the puzzle. The average number of setbacks will partly define the perceived difficulty of the puzzle and will be used to determine whether it was too difficult or too easy.

The puzzle also contains an attempt to simulate climbing in a tight space, 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 mechanics, which are explained by (Nyyssönen (2020)), are that the keywords and sounds and thus the paths they form are completely different.

The main idea with this puzzle is to test especially crawling in VR, as it is a 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.

An overview of the puzzle layout can be seen in Figure 4.

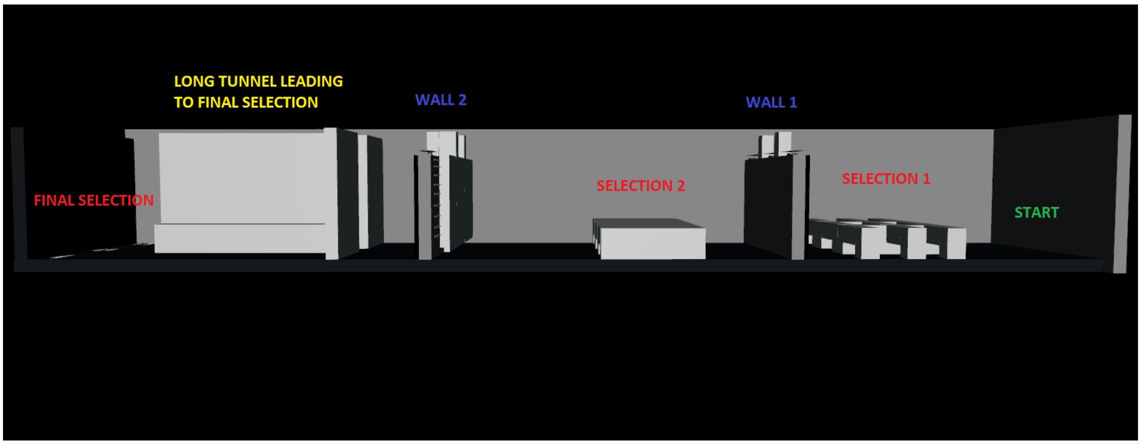

Fig. 4 A side view of the Crawling and climbing puzzle, highlighting the start area (on the right in green), path selection locations (in red) and climbable walls (in blue).

3.4 Puzzle 3: Seeking and finding

The Seeking and finding puzzle tests how difficult locations should objects be hidden in puzzle games in order to attain optimal perceived enjoyability, and how VR affects designing these types of puzzles. The puzzle area contains block-structures of various shapes and among them two hidden keys which the player must find. Both variations of the puzzle offer hints scattered around the puzzle area to the player about the locations of the keys.

In VR 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 2.1). Since there is no explicit need to indicate whether entering objects is possible to the player, they need to test everything, and this inspired the idea for the puzzle. This feature of VR creates possibilities for hiding objects within objects without giving any hints about their position in the form of game mechanics. In contrast, a player in a first-person non-VR game would instantly know whether it is possible that there are objects hidden under ingame objects, such as a table, or not based on if they are able to crouch in the game.

In summary, VR puzzle games in comparison to non-VR puzzle games provide more scope to move in a scene. The additional mobility enables hidden object puzzles in VR to be designed in a way that resembles finding objects in real life. The variations of the Seeking and finding puzzle test the enjoyability of searching and finding objects as you would closer to in real life compared to how you would look for and find objects in a non-VR puzzle game.

There are two variations of the puzzle, version A and version B. The hiding places in version A of the puzzle are designed to work only in VR conditions. The hiding places in version B of the puzzle could also be used as hideouts in non-VR games. The main focus of the research for this puzzle is to see which version is more enjoyable. Additionally, the locomotion movement type is enabled for this puzzle and all of the objects in the scenes are climbable for maximum exploration possibilities.

The differences between the two version are explained by (Nyyssönen (2020)), while an example of the puzzle overview can be seen in Figure 5.

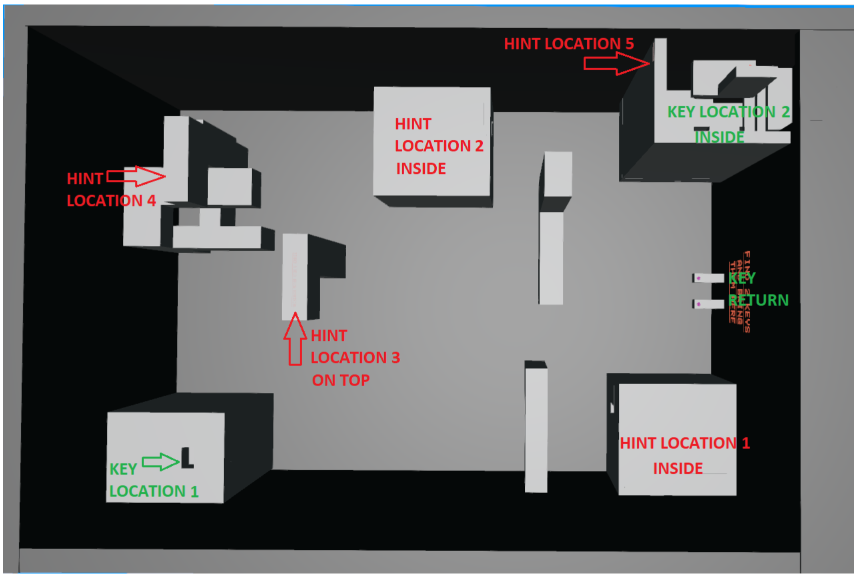

Fig. 5 The overview of the Seeking and finding puzzle version A, highlighting the locations of both of the keys (in green), hints (in red), and the area where the keys need to be taken to when found (key return).

4 Testing

Convenience sampling (i.e., using testers who are the easiest to access) was used to find testers. The voluntary testers were acquired through the networks of the University of Turku and the Turku Game Hub in addition to people familiar to the researcher. The HIVE - Turku Game Hub provided some volunteer testers who had extensive knowledge about games and game or VR development, so that the sample would also include some experts for comparability. In relation to the ethical guidelines, the testing process did not collect any sensitive personal information (e.g., names, addresses or other contact details) about the testers, and the participants cannot be identified from the data.

The test subjects were grouped by which version and puzzle order the individuals would test the puzzles, in an attempt to minimise the effects of both VR sickness and familiarity affecting the final test result. Four testers were allocated for each puzzle order, while two of those testers had the same version order as well, meaning that they would test either the more or less realistic puzzle version first for all of the three puzzles. Moreover, each tester tested individually from their assigned group. Due to a lack of volunteers, one of the groups contained only two testers, leaving the version order of that puzzle order without comparisons.

The volunteers tested the puzzles separately during the testing period that took place in November-December of 2019. The testing equipment used was a HTC Vive VR-headset with a 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 room with a $3\text{m}\times 3\text{m}$ sized play area. The test involved first reading the instructions, which explained the goals, fail conditions, and inputs of the different versions to the testers, followed by testing the puzzles and filling out the questionnaire. The puzzle instructions can be viewed in the appendices of (Nyyssönen (2020)), while the questionnaire will be briefly explained here next.

The questionnaire consisted of two parts, a pre- and a post-testing questionnaire. The pre-questionnaire was filled before testing and contained questions related to the testers' demographics, personal preferences, and abilities, while the post-questionnaire was answered after testing and contained questions only related to the testing itself. Most of the questions utilised the Likert scale (Likert (1932)) with four response options, which are considered to have equal intervals, while the remaining questions were open-ended. The response options were the following:

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

The picking criteria for the testers and the traits of the testing sample are explained in detail by (Nyyssönen (2020), pp. 47-54).

4.1 Hypotheses

The first hypothesis was that testers who identify themselves with certain phobias will experience VR related to those phobias stronger than those who do not. For example, people who are afraid of tight spaces may identify themself as having claustrophobia and this may influence their performance when tight spaces are experienced during the test. Based on the common reactions people tend to have when confronting their fears, these players may be faster or slower than average in clearing the Crawling and climbing puzzles if their phobia affects their speed. These common reactions are “flight or fight” and the latest addition *freeze*, as highlighted by Bergland (2014). The testing sample did not admit to having any phobias, so this hypothesis was difficult to verify.

The second hypothesis was that testers who had more gaming or VR experience would find all of the puzzles more enjoyable than testers with less experience, mainly because they would not be as confused by the controls and they would have more realistic expectations about what is possible in VR.

The third hypothesis was that the testers who rate themselves as more athletic than average would prefer the more realistic versions of the puzzles because they enjoy the real-life correspondents of those actions already, while the testers who rate themselves as less athletic than average would prefer the less realistic versions of the puzzles as they will require less physical effort to solve. The result of this hypothesis has been 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, the testers would enjoy the puzzles that they test first more than the puzzles that they test last. This is because they would become used to being in VR and the test structure, causing their initial amazement of trying out something novel to fade by later tests. There is also an increasing chance that the testers will develop VR sickness the longer the testing continues. To increase validity, the test subjects were assigned a unique testing order based on their grouping, consisting of puzzle and version (realistic (A) or game-like (B)) orders (as described in the beginning of Chapter 4.

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 will find the puzzles on average less enjoyable than testers who are. This is based on the assumption 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. The hypothesis was one of the ways of estimating the existence of *social desirability bias* (Edwards (1957)) in this study.

Additional major interest in this study was to explore the effects of pre-existing real-life skills or abilities related to the tested mechanics (e.g., swimming, diving, and climbing abilities), but it was difficult to hypothesise what the results would show, as testers were bound to rate their abilities subjectively and experience the mechanics differently based on their personalities.

4.2 Results and analysis

The main idea of the experiment was to compare the puzzle mechanics in a scale of game-likeness versus realism and to see if the realistic or the game-like aspects (puzzle variations) would 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 utilising the mechanics and performance.

The observations were recorded both during the testing and filling out of the questionnaires. They consist 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 results and observations are divided into general, puzzle-specific, and combined. The general results discuss the whole testing process, puzzle-specific ones highlight aspects related to a certain puzzle type, while the combined results section displays the combined enjoyment statistics of all of the puzzles. Only the major results are highlighted in this article, while the full numerical data with complete analysis, as well as the recognised error factorials are explained by Nyyssönen (2020).

General results and observations

Concerning the clarity of the instructions and the questionnaire, many testers needed to ask clarifying questions related to the vocabulary, for example, related to words such as “immersion” and “intuitive”. Furthermore, a lot of the testers were not prepared to answer questions related to some of their personal traits and abilities, struggling to decide their perceived skill levels.

The reliability of the testers was measured by asking about their perceived elevation level when climbing in both versions of the Crawling and climbing puzzle, while the climbed height stayed the same. The results showed that the responses stayed overall consistent, suggesting that the testing sample can be considered reliable, at least to some degree. Although, 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* (Kruger & Dunning (1999)).

In terms of the personal qualities of the testing sample, the previous gaming experience had the highest measured impact on puzzle version preference and general enjoyment. The results showed that the half of the testers who spent more time on gaming on a weekly basis (referred to as the “hardcore” half) had a clear preference over A-versions (the more realistic ones), while the more “casual” half preferred the B-versions. Additionally, the hardcore half enjoyed even the B-versions more than the casual half, indicating that previous gaming experience increased the overall enjoyment levels in all cases. The previous VR-experience, however, did not produce the same result, but was instead rather mixed varying based on the puzzle type, so these results only partly confirm the second hypothesis.

Conversely, a personal quality which did not have a clear measured impact on the enjoyability, unlike expected based on the third hypothesis, was perceived athleticness, which produced rather varying results, causing the third hypothesis to be disproven.

Another personal “feature” which had a high impact on the enjoyability result, was the familiarity with the testing supervisor. The result very clearly showed that testers unfamiliar with the supervisor enjoyed all of the puzzles less than those who were more acquainted, confirming the sixth hypothesis. This result was thought to most likely stem from the experience being better overall because of a familiar face present, in addition to the fact that the personal bond could have had an increasing effect on the overall ratings because of the *social desirability bias* Edwards (1957).

When it comes to the aspect of VR sickness, the testing sample contained a low amount of testers who showed any signs of nausea or dizziness during the testing. Furthermore, only two testers had to abort the testing because of these reasons, indicating that either the puzzles were designed well enough to not cause that much sickness, or the testing sample just happened to be more resistant to it. This is quite surprising because such a large percentage (16 out of 22) of the testers had none or little previous VR experience.

Additionally, the importance of the testing order was evaluated in this experiment, and the order in which the puzzle types were tested did not appear to have a clear correlation with the puzzle enjoyment, disproving the fourth hypothesis. The puzzle version order results, on the other hand, showed that being familiar with the puzzle environment beforehand, meaning testing another version of the same puzzle first, significantly increased the performance (i.e., decreased the completion time) in the second version. This result was completely unanimous across the three puzzle types, suggesting it was not caused by a coincidence. For enjoyment, however, the results indicated the opposite: testing a puzzle version as second seemed to decrease enjoyment, probably because the novelty value decreases simultaneously with experience gained. However, the result was 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.

The final measured aspect, perceived difficulty, did not reveal any major surprises: testers enjoyed challenge, a concept that is subjectively defined. This means that testers who felt a puzzle was too easy or too difficulty enjoyed it less than testers who felt that the difficulty was just right for them. Additionally, the setbacks in the puzzles like resets to the beginning had varying impact on the enjoyability ratings, indicating that they probably did not play a highly crucial role in defining them.

Swimming results and observations

This section highlights some of the major results related to the Swimming puzzle followed by the main observations. For this puzzle, the perceived enjoyment was divided into two categories: immersion (mostly due to the realistic appearing water utilised) and general puzzle enjoyment.

A major interest in this study was to see whether higher perceived swimming or diving abilities would amount to higher enjoyment levels among the testers on average. According to the results, this seemed to have been the case for the more realistic Swimming puzzle version A. On the other hand, for version B the enjoyment levels were lower for the more skilled testers, indicating that the perceived lesser realism had a negative impact on those who knew what to expect. The result indicates that pre-existing real-life swimming or diving abilities did impact the enjoyability levels of some of the testers, which is a highly interesting result.

The following are the major observations related to the Swimming puzzle:

1. For some testers of the Swimming puzzle, the drowning did not serve its purpose as a setback to be avoided. Instead, those testers realised that as the drowning did not cause a serious penalty, it could be utilised to speed up the exploration process. This means that those testers intentionally drowned in an attempt to be able to explore areas further away faster. This behaviour could have provided an advantage over the testers who always returned to the surface to resupply oxygen. 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.

2. Some of the testers became confused about the first two water areas (see Figure 2), as they thought that they had returned to the beginning when in fact 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, causing some testers to lose time and in some cases become frustrated. An additional source of confusion was the rope the testers needed to climb between water areas two and three, as many 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.

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

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 occurred when the testers were attempting to climb the rope, and most likely had minor impacts on the enjoyment levels of testers affected.

5. The movement mechanic in the version B was reported to be less immersive mainly because it utilised teleportation. 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.

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 versions of the Swimming puzzle because of it.

Crawling and climbing results and observations

This section highlights some of the major results related to the Crawling and climbing puzzle followed by the main observations. For this puzzle, the perceived enjoyment had only general puzzle enjoyment category, as the puzzle was not designed to look very immersive unlike the Swimming puzzle which included realistic appearing water.

For the Crawling and climbing puzzle, the perceived pre-existing climbing ability was one of the main areas of focus when evaluating the enjoyability levels, but unlike in the Swimming puzzle where the swimming and diving skills had an impact, the perceived ability to climb did not have a clear correspondence with the enjoyment levels. The main mechanic of interest in the puzzle, crawling, did not receive perceived pre-existing skill ratings by the testers, as it was not included in the pre-questionnaire. This was the case because crawling is not seen as a skill in the general sense, but more like an early stage of human development, so it would have been difficult to gather reliable data related to it.

In relation to the puzzle's design, which includes three possible correct paths for the testers to find out in each version, the main takeaway is that one of the paths was extremely favoured in both of the puzzle variations. This implies that the favoured paths might have been too easy by design for most of the testers or, alternatively, it is possible that the other remaining paths were too difficult.

The following are the major observations related to the Crawling and climbing puzzle:

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 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 in some testers having trouble unattaching themselves from the walls. This is why no matter the result, the toggling feature should most likely not be used for climbing in VR.

2. The puzzle version A was reset a few times for some of the testers if their virtual hands became stuck on climbable surfaces. It is unclear why some testers virtual hands could get stuck on climbable surfaces as it occurred infrequently and was difficult to recreate. Apart from this bug, the testing for this puzzle was otherwise relatively smooth.

3. Some testers incorrectly assumed 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.

4. Some of the vocabulary and wordplays used in the puzzle were difficult to understand for some testers, although the testing supervisor helped when prompted to do so. This could have been avoided by having the puzzle diction in Finnish instead of English, but then the possible volunteer base would have decreased as well. In case of utilising two separate questionnaires in different languages, the meaning of some of the vocabulary in the questions would have had to be slightly different by nature, which could have caused more unreliability in results.

Seeking and finding results and observations

This section highlights some of the major results related to the Seeking and finding puzzle followed by the main observations. For this puzzle, the evaluations 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 conventional games, thus they are less “game-like”, which creates the difference in the “realism” of the two puzzle versions.

Due to the hiding mechanics, 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 methods of hiding objects can occur in the real world. Additionally, the level was not designed to look realistic, so it would have been illogical to ask questions related to the realism.

Due to the unmeasurable realism aspect, the Seeking and finding puzzle focuses on measuring enjoyability. The enjoyment categories include key searching enjoyment and the surprise factor of finding the keys in addition to the general puzzle enjoyment, which is measured in all of the three puzzle types. The surprise factor is a measurement combining difficulty and enjoyment, and it exists to measure the impact of the perceived challenge of searching the keys on the perceived enjoyability. However, the surprise factor is more than a plain difficulty to enjoyability measurement, as a hidden object can be difficult (or at least time-consuming) to find even if the player would know where to search. Thus the surprise factor measures the design of the hideouts; if the surprise factor is high then most likely the hideout was designed well, meaning it fulfilled its purpose.

The results turn out to be almost identical for the two variations in many aspects, although the surprise factor was rated considerably higher for the “more realistic” A-variation in general. This indicates that the hideout design worked as intended, as the hideouts in version A were designed to be more unconventional and thus surprising.

The following are the major observations related to the Seeking and finding puzzle:

1. In version A, one of the keys was dominantly more difficult to discover than the other one. 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.

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. Although some testers understood the hint logic almost immediately, the overall design of those type of hints could have used revising.

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 ending up 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.

4. The problem related to the players' innate height (described in Section 2.3) 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.

Combined results

This section sums up the enjoyability comparisons together in order to see the final result about which mechanic type was perceived to be the most enjoyable. The comparisons are calculated based on all the different result categories aside from the general results, which consider the sample as a whole and are analysed separately.

The enjoyability calculations are divided into two types, individual and combined, in order to see the contrasts between them. The individual comparisons (see Table 1) calculate the sum of all of the enjoyment level comparisons between A and B-versions for singular enjoyment categories, such as immersion (Swimming puzzle), across the result tables. For each comparison, the version which has a higher enjoyment level receives a point for that comparison, while ties give points to neither and are calculated on their own. For the combined comparisons (see Table 2) the calculations function otherwise in the same way, but this time only combined enjoyability categories, such as "immersion + general enjoyment" (Swimming puzzle), are taken into account. The only exception is the Crawling and climbing puzzle, which will be included in both results in the same way, as it only features one enjoyment category.

Based on this scoring system, the most enjoyable mechanic is the more realistic one (A-versions), as its average combined and individual enjoyability results were higher in all three of the puzzles in each comparison category. It is also worth noticing that the differences between the individual and combined ratings are in some cases quite vast. This shows that combining the results is less reliable, as solely focusing on the combined results can lead to the enjoyment differences appearing more or less radical than they are.

Overall, the total amount of preferred categories for the individual comparisons for A-versions is 232/372, which is around 62.4% of the total, while B-versions 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 A-versions gaining 135/197 or around 68.5% of the total preference, leaving B-versions with just 36 or 18.3% in addition to 26 or 13.2% ties.

The highest enjoyment differences between the versions can be 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 80 and 80.8 percentages for version A.

As a contrast, the Seeking and finding puzzle has the closest enjoyment ratings between versions, 47.9% (A) to 26.1% (B) in the individual comparisons, although the combined ratings are less close (72.9% (A) to 18.8% (B)), highlighting the unreliability of combining result categories. Another interesting outcome related to the Seeking and finding results is that the tester based category in the individual results contains a large number of ties. The category scored 27 ties, which is approximately a third of the total amount of individual comparisons (78) in the category. This result indicates that perhaps the correlation between the testers' measured traits and preferred version is insignificant or at least not very clear in the Seeking and finding puzzle.

Additionally, the general enjoyment statistics across all testers for every puzzle show that the Seeking and finding puzzle version A was the most enjoyable (3.18), followed by the Crawling and climbing A (3), and Swimming A (2.96). For B-versions, the preference order is the same, numerically 3, 2.85, and 2.46. When the versions are combined the order naturally stays the same, resulting into 3.09, 2.93, and 2.71, respectively. This result means that the sample as a whole preferred A-versions over B-versions. Additionally, 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 also disproves the fifth hypothesis about testers not rating any of the puzzles highly on average, as most puzzle versions had a higher enjoyment level than the half-way point (2.5).

Table 1 The individual result categories for all puzzles. This table 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, e.g., 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 2 The combined result categories for all puzzles. This table 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 1. 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	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

5 Conclusion

The results from our study point out that the more realistic mechanics were favoured by the testers overall, which imply that those type of mechanics would be more enjoyable to use in at least these type of VR games. However, the sample size was quite low (22) and the testers were gathered from a narrow scope of the population, so the result cannot be generalised to a larger population. It can be argued though that the result serves as an important stepping stone to further research in VR mechanics and puzzle design in VR games.

According to our results, the testers' pre-existing real-life abilities related to the studied mechanics (swimming, diving, and climbing) seemed to have a partial impact on the perceived enjoyability. It seems that better perceived swimming or diving skills correlate to increased enjoyment of the Swimming puzzle in the more realistic A-version, while correlating to decreased enjoyment in the less realistic B-version. The perceived climbing ability, on the other hand, was found to have no correlations in the Crawling and climbing puzzle.

In relation to the testers' personal traits, the previous gaming experience was found to have a significant impact on the perceived enjoyability. The testers who had a lot of gaming experience preferred the more realistic A-versions in addition to having higher overall puzzle enjoyment levels compared to the testers who had less or no previous experience, who preferred B-versions. The testers' familiarity with the testing supervisor was also found to be a significant factor, as the results showed that the testers who were not familiar with the supervisor enjoyed all of the puzzles less than the testers who were. This result is likely to be at least partly caused by the social desirability bias.

The results comparing the testing order with the completion time showed that being familiar with the puzzle environment beforehand (i.e., testing another version of the puzzle first) seems to significantly increase the performance in the second version, while decreasing enjoyment. It may be interesting to see our tests replicated to learn whether the results would be similar as well as find a reason for the testing order's effect on performance and enjoyability.

In addition, a replicated study with testers who have phobias may discover whether the phobias of the sample have a statistically significant effect or not. On the other hand, testing with a focus on people with phobias is less representative of the whole population. Most likely the only way to truly test this would be to have a large enough sample size to try to ensure a considerable amount of testers with phobias to be part of the experiment.

To summarise, this experiment provides a clear conclusion about the impact of the realism of the tested VR mechanics, although the process of evaluating the design of the mechanics is an aspect worth tackling in the future.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References and further reading

Adams, E. (2014), *Fundamentals of Game Design, Third Edition*, Pearson Education Inc.

Benford, S., Greenhalgh, C., Reynard, G., Brown, C. & Koleva, B. (1998), 'Understanding and constructing shared spaces with mixed-reality boundaries', *ACM Transactions on computer-human interaction (TOCHI)* **5**(3), 185–223.
URL: *https://www.dourish.com/classes/ics203bs04/11-BenfordMixedReality.pdf*

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

Blankenship, A. (1942), 'Psychological difficulties in measuring consumer preference', *Journal of Marketing* **6**(4_part_2), 66–75.

Bonato, F., Bubka, A., Palmisano, S., Phillip, D. & Moreno, G. (2008), 'Vection Change Exacerbates Simulator Sickness in Virtual Environments', *Presence: Teleoperators and Virtual Environments* **17**(3), 283–292.
URL: *https://doi.org/10.1162/pres.17.3.283*

Bowman, D. A. & Hodges, L. F. (1997), An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments, *in* 'Proceedings of the 1997 symposium on Interactive 3D graphics', pp. 35–ff.

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

Cabooter, Elke and Millet, Kobe and Weijters, Bert and Pandelaere, Mario (2016), 'The 'I'in extreme responding', *Journal of Consumer Psychology* **26**(4), 510–523.

Ebenholtz, S. M. (1992), 'Motion sickness and oculomotor systems in virtual environments', *Presence: Teleoperators & Virtual Environments* **1**(3), 302–305.

Edwards, A. L. (1957), *The social desirability variable in personality assessment and research.*, Dryden Press.

Fernandes, A. S. & Feiner, S. K. (2016), Combating VR sickness through subtle dynamic field-of-view modification, *in* '2016 IEEE Symposium on 3D User Interfaces (3DUI)', pp. 201–210.

Kim, J., Kim, W., Ahn, S., Kim, J. & Lee, S. (2018), Virtual Reality Sickness Predictor: Analysis of visual-vestibular conflict and VR contents, *in* '2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)', pp. 1–6.

Kruger, J. & Dunning, D. (1999), 'Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.', *Journal of personality and social psychology* **77**(6), 1121.

Lee, H., Moon, M., Park, T., Hwang, I., Lee, U. & Song, J. (2013), 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, Association for Computing Machinery, New York, NY, USA, p. 287–290.
URL: *https://doi.org/10.1145/2494091.2494180*

Likert, R. (1932), 'A technique for the measurement of attitudes.', *Archives of psychology* .

Messick, S. & Jackson, D. N. (1961), 'Acquiescence and the factorial interpretation of the MMPI.', *Psychological Bulletin* **58**(4), 299.

Nyyssönen, T. (2020), Exploring virtual reality mechanics in puzzle design, Master's thesis, University of Turku.
URL: *http://urn.fi/URN:NBN:fi-fe2020052539004*

Orne, M. T. (1962), 'On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications.', *American psychologist* **17**(11), 776.

Robinett, W. & Holloway, R. (1992), Implementation of flying, scaling and grabbing in virtual worlds, *in* ‘Proceedings of the 1992 Symposium on Interactive 3D Graphics’, I3D ’92, Association for Computing Machinery, New York, NY, USA, p. 189–192.
URL: *https://doi.org/10.1145/147156.147201*

The Stonefox (Github username), Extend Reality Ltd (2019), ‘Virtual Reality Toolkit version 3.3.0’. (Accessed on 10.5.2019).
URL: *https://github.com/ExtendRealityLtd/VRTK/releases/tag/3.3.0*

Tice, D. M., Butler, J. L., Muraven, M. B. & Stillwell, A. M. (1995), ‘When modesty prevails: Differential favorability of self-presentation to friends and strangers.’, *Journal of personality and social psychology* **69**(6), 1120.

Wideström, J., Axelsson, A.-S., Schroeder, R., Nilsson, A., Heldal, I. & Abelin, r. (2000), The Collaborative Cube Puzzle: A Comparison of Virtual and Real Environments, *in* ‘Proceedings of the Third International Conference on Collaborative Virtual Environments’, CVE ’00, Association for Computing Machinery, New York, NY, USA, p. 165–171.

Yamashita, S., Zhang, X. & Rekimoto, J. (2016), 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.