Assessing Players' Cognitive Load in Games

Alberto Faria Coutinho da Silveira Ramos

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Carlos António Roque Martinho
Prof. Marta Barley de La Cueva Couto

Examination Committee

Chairperson: Prof. Nuno João Neves Mamede
Supervisor: Prof. Carlos António Roque Martinho
Member of the Committee: Prof. Ana Paula Boler Cláudio

January 2021

Acknowledgments

First and foremost, I would like to thank my two advisors – Prof. Carlos Martinho and Prof. Marta Couto – for the exemplary guidance provided, since the first day. This work would not have been possible without them.

Secondly, I would like to thank my parents for their tireless support throughout my entire life. I hope you know that I am extremely grateful for the education you have given me, and that my work ethic is the result of that.

I would also like to dedicate a special "thank you" to my grandparents. My paternal grandparents, not only for their consistent presence and availability throughout my life, but also for the question they asked me every weekend; something along the lines of "So what about the thesis, when do you finish it?"; which served as a weekly reminder to put my procrastination aside. My maternal grandfather, for the example of an outstanding person that he is – and always has been. Since I was little, I have had this persistent ambition of "being like him when I grow up", which is something that I am continuously working towards achieving.

Finally, I would like to thank the rest of my family members who are not highlighted above; I promise you are equally relevant in my life. I simply wanted to keep this section concise. Nonetheless, I would like to thank you for keeping me optimistic and, most importantly – happy – over these past few months. It is important to emphasize that this last paragraph also includes my closest friends; not only for their camaraderie, but also for boosting my good mood every single day. For me, the line that separates these friendships from "family" is very diffuse.

To each and every one of you – Thank you.

Abstract

Due to the exponential growth of computer technologies, video games are becoming more complex each passing year; with tasks and challenges that, very often, defy the player's cognitive abilities. Handling limitations of the Working Memory and proper Cognitive Load management is crucial when dealing with problem-solving tasks; however, these concepts appear to be highly undervalued, or even unknown, in the gaming industry.

To address this problem and help game designers to better understand the intrinsic complexity of their games, this work applies the attention-shifting principles of the Time-Based Resource Sharing (TBRS) Memory Model in the game Way Out (a game we have developed from scratch). We formulated the idea of Attention-Grabbing Events and tried to incorporate them into the game, aiming to create a tool-set that estimates the player's Cognitive Load while playing a video game. To validate our hypothesis, we compared the data collected from the game with the questionnaire NASA TLX – a subjective method that assesses the mental workload experienced during a task.

Although we were unable to directly estimate the player's Cognitive Load, we believe that this work was a step forward towards achieving that goal. The amount of Attention-Grabbing Events and gameplay time, when compared with the NASA TLX, seem to be a good indicator of Cognitive Load levels. However, the TBRS Cognitive Load formula, in its current form, does not appear to be reliable when directly applied in a general gameplay scenario – at least following the approach we did.

Keywords

Cognitive Load; Working Memory; Time-Based Resource Sharing; Video Game; Game Development; NASA TLX.

Resumo

Devido ao crescimento exponencial tecnológico, os video jogos estão a tornar-se mais complexos a cada ano que passa; com tarefas e desafios que, muitas vezes, desafiam as habilidades cognitivas dos jogadores. Lidar com as limitações da Memória de Trabalho e gerir a sua Carga Cognitiva é crucial para a resolução de problemas; no entanto, estes conceitos parecem ser altamente subvalorizados, ou mesmo desconhecidos, na indústria de jogos.

Para abordar este problema e ajudar os designers de jogos a entender melhor a complexidade intrínseca de seus jogos, este trabalho aplica os princípios de atenção do modelo de memória "*Time-Based Resource Sharing*" (TBRS) no jogo Way Out (um jogo que desenvolvemos de raiz). Formulámos a ideia de "*Attention-Grabbing Events*" e tentámos incorporá-la no nosso jogo, com o objetivo de criar um ferramenta que estime a Carga Cognitiva dos jogadores enquanto jogam video jogos. Para validar a nossa hipótese, comparámos os dados recolhidos do jogo com o questionário NASA TLX – um método subjectivo que avalia a Carga Cognitiva sentida durante uma tarefa.

Mesmo não tendo conseguindo calcular diretamente a Carga Cognitiva dos jogadores, acreditamos que o nosso trabalho serviu para nos aproximar-mos desse objetivo. A quantidade de *Attention-Grabbing Events* e tempo de jogo, quando comparados com o NASA TLX, parecem ser bons indicadores dos níveis da Carga Cognitiva dos jogadores. No entanto, na sua forma atual, a fórmula do TBRS que calcula a Carga Cognitiva não parece ser adequada quando diretamente aplicada em video jogos – pelo menos seguindo a nossa abordagem.

Palavras Chave

Carga Cognitiva; Memória de Trabalho; Time-Based Resource Sharing; Video Jogo; Desenvolvimento de Jogos; NASA TLX.

Contents

List of Figures

List of Tables

Listings

Acronyms

LTM	Long-Term Memory
STM	Short-Term Memory
WM	Working Memory
CL	Cognitive Load
CLT	Cognitive Load Theory
TBRS	Time-Based Resource Sharing
AGE	Attention-Grabbing Event
UI	User Interface

1

Introduction

Contents

1.1 Motivation

Due to the exponential growth of computer technologies in the last decades, video games are becoming more complex and diversified than ever. From deep and intriguing storytelling to complex game mechanics, it is unquestionable that the video game industry is doing a proper job in keeping up with this growth and creating games that are becoming more realistic and immersive each passing year.

Back in the 70s and 80s, when video gaming was emerging and becoming mainstream, games were much simpler and had straightforward mechanics that a joystick and a few set of buttons could handle. Space Invaders, for example, a fixed shooter created by Tomohiro Nishikado in 1978 that is considered one of the most influential video games of all time, consists of controlling a space cannon horizontally while firing descending alien forces. The enemy spaceships approach the player more rapidly as time passes, making the game harder the longer it's played. The mechanics, however, are quite simple and easy to memorize.

Nowadays the story has diverged immensely – each year, thousands of new video games are released with complex mechanics that take much longer to master and require entire keyboards to be played with. An example of this can be observed in the game Dark Souls, an action role-playing game that was developed by *FromSoftware* and released in 2011. In this game, the player assumes the role of an undead character that explores the virtual kingdom of Lordran to seek the fate of his kind. A game well known for its hard boss fights that, in order to be beaten, forces the player to learn from past mistakes by memorizing the enemies movements and weaknesses. Demanding mechanics like these require a great amount of attention and cognitive resources and, if not dealt with accordingly, can easily lead to negative emotions such as frustration or anger.

Handling limitations of Working Memory and proper Cognitive Load management is crucial when dealing with problem solving tasks and is proven to positively influence effective performance and learning [10]. Since the Working Memory has a limited capacity and is believed to only retain information for a small period of time of approximately twenty seconds, it is easily overloaded if more than a few chunks of information need to be simultaneously processed.

These limitations and concepts, which are highly important in neurological and physiological matters, appear to be quite undervalued and ignored in the gaming industry. If Cognitive Load and the overall correct management of Working Memory's resources are taken into consideration by game designers in early phases of game development, *highly beneficial results could be obtained.* By estimating the amount of Cognitive Load that a players' Working Memory is using while playing a video game, game designers would have, in addition to play testing feedback, an extra source of reliable information that would be an indicator of their game levels complexity. Hence, excessively demanding tasks could be detected and adjusted accordingly earlier, facilitating and cutting costs in the play testing phase and allowing the developers to focus on other aspects of the game.

1.2 Problem

Play testing can be often misleading. Finding the right people who give honest reviews and clearly point the flaws of a game can be challenging, specially when testing happens at a studio. When feedback is given face-to-face to the developers, testers tend to be reluctant to provide negative feedback, even if they feel that's the truth. This enters in the spectrum of *social desirability* [11], where testers or survey participants tend to answer questions in a manner that will be seen as more favorable to others. This can lead to biased reviews that are unhelpful to game designers since they become unaware of what to change and improve. Play testing often relies on honesty, and honesty is not always present in game reviews.

If appropriate tools to measure physiological data are incorporated into games, developers can have a much better understanding of the quality of their games before release. More specifically, if there was a way to measure the players' Cognitive Load – which is, in a simplified way, the amount of resources that the Working Memory uses, amongst multiple other things, to solve problems and learn novel information – game designers would have an additional source of feedback that is, inevitably, honest.

1.3 Hypothesis

When novel information enters our Working Memory, it is either repeated or trained and can integrate existing chunks in our Long-Term Memory or formulate new knowledge structures that can be more or less complex – called schemas.

Our brain creates these schemas because it makes it easier to save space in our Working Memory, reducing the Cognitive Load. Therefore, when people are faced with a task that is a novelty for them, they tend to have higher levels of Cognitive Load because schemas related with that task were not yet created or are still too simplified. In contrast, people that perform that same task on a regular basis already have created schemas in their Long-Term Memory, allowing them to "skip steps" of the task that they are so used to execute – which translates to a lower execution time and, consequently, in *lower levels of Cognitive Load*.

Assuming it is possible to estimate the duration of time in which the player's attention was fully grabbed during a game, it is theoretically possible to apply the principles of a Memory Model to assess the players' Cognitive Load. Therefore, we hypothesise that if the attention-shifting principles of the Time-Based Resource Sharing (TBRS) Memory Model are incorporated in games, and if the model's formula to assess Cognitive Load is correctly used, it would be possible to estimate the amount of cognitive resources used by a player's Working Memory, while playing a video game.

This work aims to confirm whether or not our hypothesis is valid, by integrating this model within a game that we have developed from scratch. We will compare the collected game data with a subjective method that also estimates a users' Cognitive Load during an activity – the NASA TLX questionnaire.

1.4 Contribution

This work mainly aims to provide game designers more in-depth insights about the experience they are developing – by estimating the amount of cognitive resources used by players while playing their games. Additionally, in a general sense, our hypothesis seems refreshing when it comes to assess Cognitive Load in multimedia applications, since it tries to incorporate the principles of a Working Memory Model within an application (in this case, a video game).

For that purpose, we started by reviewing the literature related with the Working Memory, its evolution and different ways of assessing Cognitive Load; followed by an overview of some related work that also values the users' Cognitive Load in multimedia environments.

To validate our hypothesis, we established a computational model that adapted a memory model based on attention-shifting – Time-Based Resource Sharing – to a game that we have developed from scratch.

Along side the obtained results of our work, all the documentation of this research will be provided with the expectation to contribute to a possible solution to the problem mentioned above.

1.5 Document Outline

This work was written conform the following structure:

- **Chapter 1: Introduction**, presents the main motivation behind the development of this work; highlights the importance of assessing Cognitive Load in games and briefly explains the approach we took in trying to achieve our goal.

- **Chapter 2: Background**, starts by clarifying the two main theoretical concepts of our work – the Working Memory and Cognitive Load – and then overviews the evolution of the Working Memory Models throughout the years. It finishes by analysing the main approaches to assess one's Cognitive Load.

- **Chapter 3: Related Work**, highlights some of the different approaches made throughout the years to assess and/or control Cognitive Load levels in diverse multimedia applications and learning environments. Even though most are not directly correlated with games, the listed projects were, nonetheless, a source of inspiration for our work – and should be taken in consideration when developing interactive multimedia applications that may require the user's cognitive resources.

- **Chapter 4: Implementation**, starts by describing the approach we took when trying to apply the attention-shifting principles of the Time-Based Resource Sharing Memory Model into games, in order to assess the player's Cognitive Load. To successfully test our hypothesis, we developed a game from scratch; this chapter also extensively describes the entire process of its development.
- **Chapter 5: Procedure and Results**, starts by describing the data acquisition procedure and justifies some of the decisions made to validate our game data. Additionally, it extensively analyses and compares the obtained data with a trustworthy questionnaire that also assesses the mental workload during a task.
- **Chapter 6: Conclusion**, concludes this dissertation by summarising the work done and by presenting possible future approaches to continue our research.

2

Background

Contents

2.1 The "Short-Term Store"

The distinction between the nowadays called "Short-Term Memory" and "Long-Term Memory" was firstly, somewhat, controversial. It was argued that such division was useless and would unnecessarily complicate the concept of memory. However, evidence that such division would, in fact, make sense, started to emerge around the 60s. A strong argument in favor of a dichotomy in the memory system was noticed by Milner, while studying patients with hippocampal lesions [12], who appeared to became incapable of either store or retrieve information from the Long-Term Memory (LTM) but could still process and register immediate input for short periods of time. This inspired R. C. Atkinson and R. M. Shiffrin to deepen the studies of the memory and the dichotomy of the LTM and this new "Short-Term Store", leading them to conceive the first Memory Model [1] (further explained in Section 2.3.1).

The Working Memory (WM), initially named Short-Term Store and, nowadays, often called Short-Term Memory (STM), is now commonly known as a *cognitive system crucial for reasoning and decision-making that can hold information for a short period of time*. Additionally, contrary to the LTM, the WM has a *limited capacity and a certain amount of resources available to properly work*.

In the context of our work, *Cognitive Load (CL) refers to the amount of resources used by our WM to properly function* (*i.e.* to solve problems, learn novel information, react to stimulus, etc.). These resources are limited and need to be properly managed to avoid *cognitive overloading* [13].

The following sections will explore the origin of these concepts and their evolution.

2.2 Cognitive Load Theory

The limited capacity of the WM was first suggested in the 1950s by George A. Miller [14]. After numerous experiments, Miller proposed that the humans' STM capacity is limited and only able to hold seven plus or minus two units of information at a given time. The term "unit" was later updated to *chunk* or **schema** – which refers to a pattern of thought that categorises and gathers related information according to the manner in which they will be used, creating a mental structure or framework in the LTM. Schemas, for instance, are what allows us to tell that certain objects are trees to which we can react in a common way even though no two trees have identical elements [10].

In the late 1980s, while studying the complexity of problem-solving, and based on the most influential model of the WM which is the Multi-Store Memory Model [1], John Sweller proposed the **Cognitive Load Theory (CLT)**. In Swellers' own words: *"Cognitive Load Theory has been designed to provide guidelines intended to assist in the presentation of information in a manner that encourages learner activities that optimize intellectual performance"* [10]. This theory differentiates CL in three types: **Intrinsic**, **Extraneous** and **Germane** CL.

Intrinsic CL is related to the inherent difficulty and complexity associated with the information that a person is paying attention to, which cannot be manipulated. For instance, if a student is paying attention to a lecture and the lecturer is explaining a very complex diagram with a lot of elements, that is going to translate in a high intrinsic load for the student.

Extraneous CL is the amount of resources that the WM is using with unrelated matters to the learning process – it essentially refers to any distraction affecting learning. Taking the previous example, if there are background conversations between students while the lecturer is explaining the diagram, the extraneous load will be high for those who are trying to focus.

The third type, *Germane CL*, is associated with the mental effort that is being used towards the development of schemas. It can be promoted, for example, through the use of mnemonics, including rhyme schemes and acrostics[1], all of which are used to make the learning process and the creation of schemas easier.

The sum of Intrinsic, Extraneous and Germane CL gives the total amount of CL being used, which should not surpass the total capacity of one's WM. The central problem that CLT identified was that, when the limited capacity of humans WM is exceeded, learning becomes much more challenging. Thus, in order to enhance the learning process and to better consolidate information in the LTM, since Intrinsic CL cannot be manipulated, *designers and instructors should instead focus on lowering Extraneous CL and improve Germane CL*.

In psychology and the context of this work, CL simply refers to the total amount of resources used by the WM to process and maintain novel information.

2.3 Memory Models

In Luis Buñuel own words, *"You have to begin to lose your memory, if only in bits and pieces, to realize that memory is what makes our lives. Life without memory is no life at all... Our memory is our coherence, our reason, our feeling, even our action. Without it we are nothing."*

Our memory is what defines our existence. It allows us to recall the past, make decisions and encode and store new information that is retrieved when required. It is, in neurological and physiological terms, a complex set of encoded neural connections in the brain. Memory, however, it is a hypothetical construct and thus, everything that is known about it is composition is purely theoretical.

[1] "An acrostic is a piece of writing in which a particular set of letters – typically the first letter of each line, word, or paragraph – spells out a word or phrase with special significance to the text. Commonly written as a form of poetry, but can also be found in prose or used as word puzzles"

2.3.1 Multi-Store Memory Model

Different theories about the composition of the memory have been proposed in the last decades. The first and undoubtedly one of the most influential was the **Multi-Store Memory Model**, published by Atkinson and Shiffrin in 1968 [1] (Figure 2.1). This model proposes that the memory is a one way system that is composed by three stores: the Sensory Memory, the STM and the LTM.

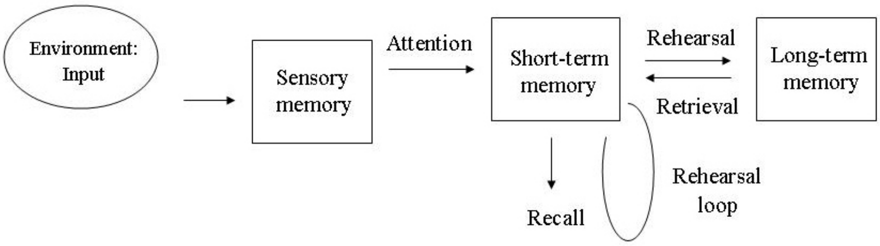


Figure 2.1: Multi-store Memory Model (1968) [1].

The Sensory Memory (or sensory register) is what allows us to retain impressions of the information provided by our five senses right after the original stimuli have ended. It absorbs and retains as much information as possible for about one second and prioritizes which data is passed to the STM, based on its relevance. If, for instance, a student is in a class trying to learn a new subject, his sense of taste is irrelevant when compared to the sense of sound, which will be prioritized and passed to the STM since, in this scenario, the goal is to learn the matter that is being taught.

The STM[2] takes the seven most relevant chunks of information [14], filtered by the Sensory Memory, and can hold them for a short period of around 15 to 30 seconds. After that point, the information held will either be forgotten or rehearsed.

In this model, the rehearsal loop is what keeps the information stored in the STM before reaching a *rehearsal threshold* where it is passed to the LTM – which has an unlimited capacity and can store information from a few minutes to a lifetime.

Even though very influential, this model was criticized due to its simplistic and linear approach. Both the STM and the LTM, for example, are believed to be made up of multiple sub-components that are not taken into consideration in this model. This model has been refined over the years and led to new, and more complete, memory models.

2.3.2 Working Memory Model

In order to fill the gaps that the Multi-store Memory Model had, specifically on the STM, Baddeley and Hitch proposed a new **Working Memory Model** in 1974 [2] (Figure 2.2).

[2] "Short-Term Memory" (STM) is also commonly called "Working Memory" (WM); we mainly use that terminology in future chapters.

This new model proposed that the WM is divided in 3 subcategories: Central Executive, Phonological Loop and Visual-spatial Sketchpad. The **Central Executive**, even thought the information about it is scarce, is what drives the WM. It is believed that this element of the model is both responsible for allocating the information received into the slave systems, as well as dealing with multiple cognitive tasks, such as problem solving and mental arithmetic.

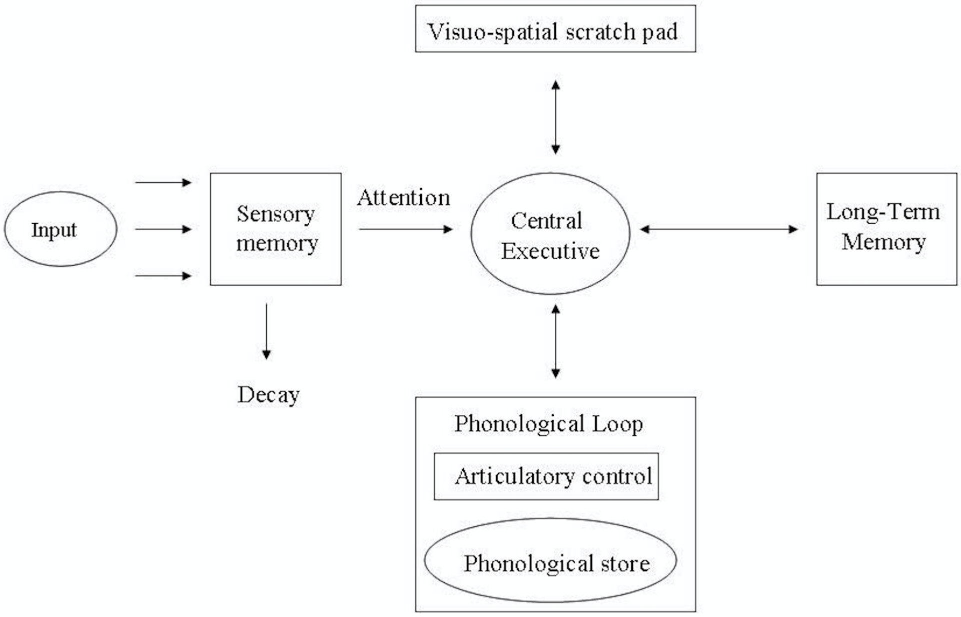


Figure 2.2: Working Memory Model (1974) [2].

Onto the slave systems, the **Phonological Loop** is a temporary store that holds verbal and auditory information. It is divided in two sub-systems: The Phonological Store, that functions as a inner ear, since it stores what is heard; and the Articulatory Process, that functions as an inner voice, since it rehearses words or sounds to keep them in WM while needed. The other slave system that was initially proposed is the Visual-spatial scratch pad – also known as **Visual-spatial Sketchpad** – and it works as a passive screen; since it stores, processes and manipulates every visual or spatial information.

In 2000 the model was updated and a fourth component was introduced – the **Episodic Buffer** [3] (Figure 2.3). As the Central Executive, not much is known about this element of the model, but is believed that it stores information in a multidimensional way, *i.e.*, is capable of binding information from the subsidiary systems, and from LTM, into a unitary episodic representation.

Using a Dual-Task paradigm, it is possible to prove that the elements of this model work properly [15]. People are able to complete two tasks simultaneously if different processing systems are being used. However, it becomes substantially more difficult to complete two tasks at the same time, if the same processing system is required for both (*e.g.* trying to rub the stomach while patting the head, is a known difficult exercise to execute because both tasks require the usage of the Visual-spatial Sketchpad).

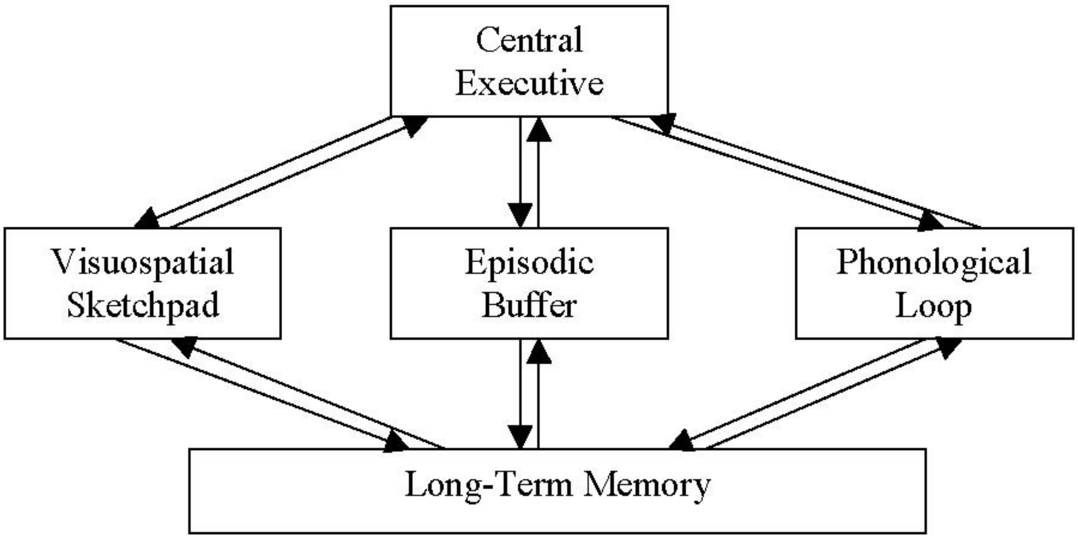


Figure 2.3: Working Memory Model (2000) [3].

The gaps and answers that the Central Executive leaves unknown, led to different theories and studies to come up in the recent years, and some believe that categorizing the WM in different compartments is outdated and unrealistic. In 2016, Robert H. Logie suggested that the Central Executive "acted as a placeholder umbrella term for aspects of cognition that are complex" that "were poorly understood at the time"; instead, he suggests that it does not work as a single executive homunculus, but rather as a *collection of multiple strategies and abilities that work cooperatively as a self-organizing system*. [16]

2.3.3 Time-Based Resource Sharing

Time-Based Resource Sharing (TBRS) is a fairly recent model that was introduced by Barrouillet and Camos in 2004 [4], and is based on four main assumptions.

The *first*, is that both the processing and maintenance of information require and share the same resource, which is attention. The *second* assumption is that as soon as attention is switched away, the activation of the memory traces suffers from a time-related decay. Additionally, the refreshment of these decaying memories traces, requires their retrieval from memory by attentional focusing. The *third* assumption is that any processing that captures attention, disrupts maintenance by preventing the refreshment of memory traces; therefore, WM functioning is limited by a *central bottleneck*. Which leads to the *fourth* and final assumption: since attention can only be devoted to one process at the time, maintenance and processing cannot occur concurrently, meaning that, to maintain information in WM (to avoid forgetting) it is required that the individual regularly switches attention from processing. This means that the central bottleneck allows only one central process at time, **making the sharing of attention time-based**.

The authors believe that, as Kahneman suggested in 1973 [17], tasks that impose a heavy load on the WM, necessarily impose severe time-pressure. A simple task that requires continuous attention can be as cognitively demanding as solving complex mathematical equations.

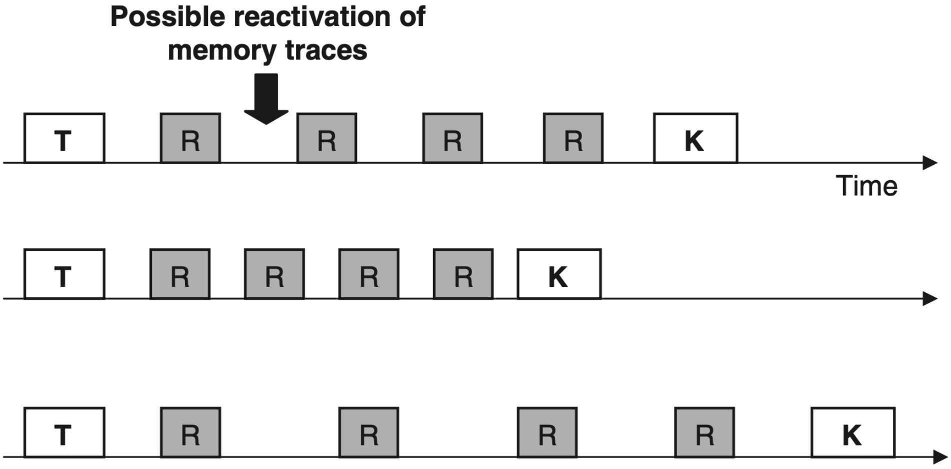


Figure 2.4: TBRS - “Reading digit span task” exercise [4].

To validate their hypothesis, they came up with a simple task where participants were asked to maintain letters in memory while simultaneously performing a secondary task that involved reading a series of digits that were presented, one at the time, on a screen (Figure 2.4). The idea is that if time pressure is applied to a task as simple as this, it can easily become much more demanding. Thus if the digits from the secondary task are presented at a faster pace, maintaining letters in WM becomes much harder, since there is less time to reactivate memory traces – leading to a higher CL. However, if the digits are presented at a slow or comfortable pace, there is time to reactivate memory traces – leading to a low or moderate CL.

In the case of this model, CL refers to the total amount of time during which attention was fully captured and can be formulated as:

$$CL = \frac{\sum_{i=1}^{N} a_i}{T} \tag{2.1}$$

a_i reflects the latency in which the i_{th} event fully captured attention.

T refers to the total duration of the task or activity.

If the total number of processes N is known, the formula can be simplified by using average processing times:

$$CL = \frac{\overline{a}N}{T} \tag{2.2}$$

To illustrate the concept of CL, *i.e.* the balance between the competing actions that are processing and maintenance, and grabbing the example from Figure 2.4; suppose that a participant has to say 10 letters out loud, each takes 200 milliseconds to be said and the time available is 4 seconds. The resulting CL of this example would be 10 x 200 / 4000 or 0.5. However, if the time available doubled, the resulting CL would be cut in half (10 x 200 / 8000 or 0.25).

A computational implementation of this model was proposed in 2011 – the TBRS* – aiming to deepen the knowledge of the WM and to validate the assumptions underlying the TBRS [18].

2.4 Methods to assess Cognitive Load

When it comes to measuring CL, the main challenge is knowing if the methods used are valid, reliable and practical. Conventionally, there are two main approaches to assess the WM's capacity: **Objective** and **Subjective** [19].

2.4.1 Objective Approach

The **Objective Approach** mainly relies on behavioral data collected from the users while performing a task. Whilst commonly more reliable, this approach may affect a users' focus from the task itself, since it often requires the usage external and intrusive equipment. Two well known methods that, by collecting behavioral data, analyse the learning process, are the Dual-Task Technique and Task-invoked pupillary response.

The *Dual-Task Technique* requires the testers to perform two tasks simultaneously. In a multimedia scenario, the secondary task usually translates in a simple visual monitoring activity that requires the testers to react as soon as possible when they spot a change. In a game play scenario it could be, for example, a light at the right top corner of the screen that, from time to time, changed color from green to red. Every time the players spotted this change, they would have to press a key – the reaction time in this secondary monitory task would then be used to assess the players' CL.

Task-invoked Pupillary Response is one of the most well known objective ways for measuring CL and it consists of analyzing the pupil activity. Pupil dilatation occurs when there is high CL, while pupil construction is associated with low CL [5]. One of the main advantages of using this technique is the direct correlation between pupil activity and the WM, allowing it to accurately measure CL in scenarios unrelated with learning, which is the case of our work. However, using this method in a game play scenario could be very intrusive and affect the players' experience. Figure 2.5 is from a study that explored the effects that pupillary responses had on CL [20] – it illustrates cognitive tasks (both easy and hard) and how they affect the respective pupil dilation.

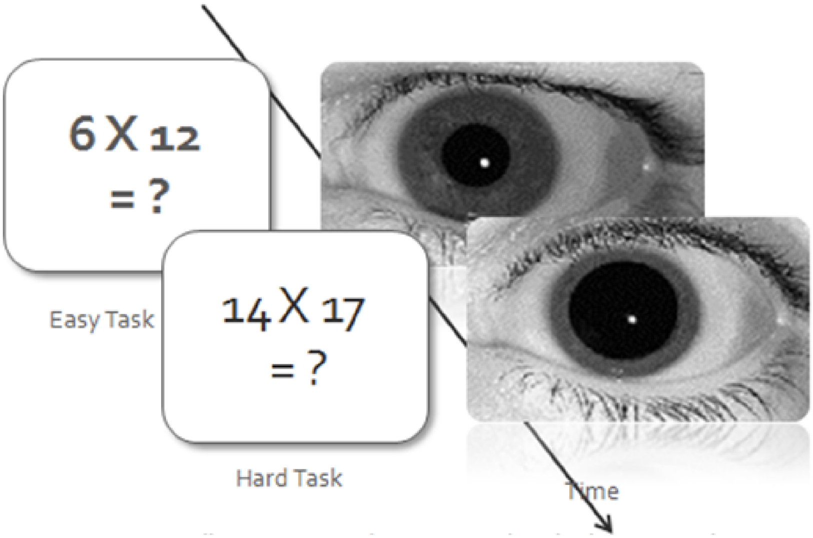


Figure 2.5: Mental multiplication task and pupil size change [5].

2.4.2 Subjective Approach

The **Subjective approach** is probably the most common and, as the name implies, requires the subject to do some sort of self-report after completing a task. Usually these subjective self-reports require the respondents to rate the perceived mental effort or task difficulty in a numerical scale.

Initially, this approach presented some limitations; perhaps the most important of them being related with the content validity, *i.e.* knowing which of the three types of CL originated the mental effort (Intrinsic, Extraneous or Germane). Was it caused by a poor design? By the subject's personal problem-solving skills? Perhaps by the material presented itself? This problem, however, has been mitigated throughout the years with new questionnaires that are much more sensitive to small differences in CL, while keeping their reliability and unintrusive nature [21]. Since it requires the users to introspect about their cognitive processes after performing an activity, other limitation of this approach is the fact that it is not possible to monitor the CL variations in real time – it simply returns a briefing of the overall CL demanded by the task. Nonetheless, when properly applied, this approach returns high indexes of data validity [21].

Possibly the biggest strength of the subjective approach is its simplicity; as it does not require any external equipment or manipulation of the information that is being presented to the subject. After completing a task, the subject simply has to answer a series of questions and, based on the responses provided, the CL can be estimated. Additionally, this approach does not require the usage of external equipment to collect the user's behavioral data during an activity, making this a non-intrusive approach.

Given the relevance that this approach has to our work, it will be further explored in the following subsections.

2.4.2.A Questionnaires

Questionnaires or self-reports are a subjective approach to assess CL.

Before 1992, the absence of proper methods to directly measure CL was noticeable, which inspired Paas (1992) to come up with a new way that helped the enhancement of the first subjective method to measure CL during activities. He concluded that subjects are capable of self-evaluating their mental effort and workload during an activity, and that the “intensity of that effort” could be considered as an index of CL [22]. The first self-reports used a 9-point Likert scale to measure CL that ranged from 1 (*very, very low mental effort*) to 9 (*very, very high mental effort*) [22].

Later, in 2003, the definition of mental effort was refined to “the aspect of cognitive load that refers to the cognitive capacity that is actually allocated to accommodate the demands imposed by the task. Thus, it can be considered to reflect the actual cognitive load” [21].

There are two most commonly used techniques for subjectively assessing mental workload [23] – the **NASA TLX** and the **SWAT**. They both divide the workload in multiple subscales and are proven to provide quite similar results [23]. However, for the context of our work, since it assesses a wider variety of mental workload components involved in the experience, we ended up opting to use the NASA TLX technique.

2.4.2.B NASA TLX

The NASA TLX (NASA Task Load Index), developed in 1981 by Sandra G. Hart of the NASA Ames Research Center [24] is one of the most known subjective techniques to assess CL. It has been used in various domains such as healthcare, aviation, and others of similar technical complexity. It is a subjective workload assessment technique that relies on a multidimensional construct to derive an overall workload score based on a weighted average of ratings on six subscales: Frustration, Effort, Temporal Demand, Physical Demand, Mental Demand and Performance [25]. These subscales of the workload are based on the assumption that some combination of these dimensions are likely to represent the overall *workload* experienced by most people performing most tasks [26]. Three of the subscales focus on the demands imposed on the subject (mental, temporal and physical demand), whereas the other three explore the interaction of the subject with the task (effort, performance and frustration levels) (Table 2.1).

Table 2.1: NASA TLX Rating scale descriptions.

Mental Demand	How much mental and perceptual activity was required (*e.g.*, thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	How much physical activity was required (*e.g.*, pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	How insecure, discouraged, irritated, stressed, and annoyed or secure, gratified, content, relaxed, and complacent did you feel during the task?

This technique consists of two parts: **ratings** and **weights**. The first part usually occurs right after the completion of the task, and consists of rating each of the six subscales individually in a numerical scale ranging from 0 to 100 (least to most demanding). The second part requires the subject to rate each sub-scale pairwise based on their perceived importance and relevance to the activity in question. The weights are calculated from the score of these choices from 15 possible pair combinations from the six subscales and range from 0 to 5 (least to most relevant) [25].

Workload ratings are usually recorded immediately after a task was performed however, a 15 minutes delay is tolerable and does not significantly interferes with the recall of ratings, as Moroney demonstrated in 1992 [27].

This is the questionnaire we chose to validate our hypothesis; therefore, its technicalities will be further explained in a future chapter (Chapter 5, Section 5.1.1).

2.5 Summary

The limited capacity of the WM was first suggested in the 1950s by George A. Miller [14]. After experimenting, he came to the conclusion that the human's WM was only able to hold seven plus or minus two chunks of information – this discovery had a major impact in future studies related with the WM.

In the 1980s John Sweller introduced the CLT. His theory differentiated CL in three types: *Intrinsic, Extraneous and Germane* [10]. Very briefly, *Intrinsic* CL refers to the inherent difficulty of an activity, *Extraneous* CL is related with external factors to the activity that might capture the participants attention, and *Germane* CL is related with the mental effort that is being used towards the development of schemas – related with the way the information is passed and processed. The sum of the three types gives the total amount of CL that is being used in a task or activity.

In 1968, Atkinson and Shiffrin published the first memory Model, the **Multi-Store Memory Model** [1]. The model proposes that our memory is a one way system that is composed by three stores: The *Sensory Memory*, that retains as much information as possible and filtrates and prioritizes the data that passes to the STM. In a nutshell, it is what "decides" what is or is not "worthy of your attention". The *Short-Term Memory* (STM), that processes and holds the information passed by both the Sensory Memory and the LTM for about 15 to 30 seconds. It can rehearse the information, keeping it fresh in the STM. And finally, the *Long-Term Memory* (LTM) that has an unlimited capacity and organizes the information in schemas.

Due to its simplicity, Baddeley and Hitch proposed a new **Working Memory Model** [2] in 1974, that divides the STM in 3 subcategories: The Central Executive, the Phonological Loop and the Visual-Spatial Sketchpad. Each of these subdivisions had it is own purpose to process different types of information. In 2000 they updated their previous model and added a fourth component, the Episodic Buffer, which is responsible for binding information from the other stores.

Time-Based Resource Sharing (TBRS) [4] is a fairly recent Memory Model that defends that *attention* is the core component of our WM; and that both the processing and maintenance of information that enters in our WM requires this component. However, processing and maintenance cannot occur simultaneously, so there needs to be a constant switch of attention throughout time, hence the name of the model. This model is quite simple, yet one of the most important for our project, and we will discuss it in depth in Chapter 4.

Conventionally, there are two main approaches when it comes to measuring Cognitive Load: the **Objective** and the **Subjective Approach** [19].

The *Objective Approach* is usually far more precise when it comes to assessing CL, since it extracts behavioral data in real time from the users during an activity. However, it commonly requires external equipment to extract that data; making it, in a general way, an intrusive approach. Some of the most known objective methods to assess CL are, amongst others, the *Dual-Task Technique*, *Task-Invoked Pupillary Response* and *Heart-rate blood pressure*.

As the name implies, the *Subjective Approach* requires the participant to self-evaluate their workload during an activity after performing it – usually in the format of a questionnaire. One of the most known questionnaires and, therefore, subjective approaches, is the **NASA TLX** questionnaire. Given the importance of this questionnaire to our work, it will be reviewed with more detail in Chapter 5.

3

Related Work

Contents

WM and CL are not commonly conceived as important in the game development industry; therefore, there is a noticeable lack of studies that directly correlate the concepts introduced in the previous chapter, with games. Nonetheless, we grasped inspiration in other studies that aimed to assess and reduce CL in various Multimedia environments. This chapter will introduce some of those studies that, even though not directly correlated with the approach we opted to follow, were, nonetheless, a source of inspiration; and we believe that they should be taken in consideration when developing interactive multimedia applications that may require the user's cognitive resources.

We will start by analysing some important negative repercussions that extraneous CL may have in multimedia learning; followed by a brief explanation of how the widely known eye-tracking technique can be applied in multimedia environments to improve the users experience; and finalizing with the analysis of some studies related with CL in educational games.

3.1 Cognitive Load in Multimedia environments

As modern technology evolves, new multimedia learning environments emerged; and the correlation between these new environments with the correct management of cognitive resources has been a subject of many recent studies – which is understandable since the WM is crucial for processing and learning novel information.

In these new multimedia learning environments, techniques that simultaneously present visual and auditory information, in both static or dynamic ways, are often used as a way to try to enhance learning. However, from a cognitive point of view, they frequently ignore a possible "cognitive overload", leading to inefficient learning [28].

The following subsections will explore some of the work done to mitigate or assess CL in multimedia environments.

3.1.1 Negative repercussions of Multimedia Learning

A frequently discussed problem in multimedia learning is the effects that the simultaneous presentation of images/animations with explanatory text might have on the learners. Whereas it can bring positive results in learning, it can also easily lead to two main negative repercussions.

Firstly, it can lead the learners to the **split attention effect**, which occurs when it is required for them to *to split their attention between two or more mutually dependent sources of information (e.g. explanatory text and a diagram) which have been physically or temporally detached*. These sources have to be later mentally restructured, which consequently increases extraneous CL and can have a negative impact in learning [29].

Secondly, it can also lead to the **redundancy effect**. The simultaneous presentation of images and explanatory text with very similar information can also lead to an increase of CL due to the competition of resources in the visual WM [30].

To curtail problems like these, a study from 2003, after 12 years of research, proposes nine ways to reduce CL in multimedia learning environments; presenting different scenarios often observed in these environments and the respective solutions [6]. This study aims to help designers understand the small changes that can me made to substantially reduce the learners CL.

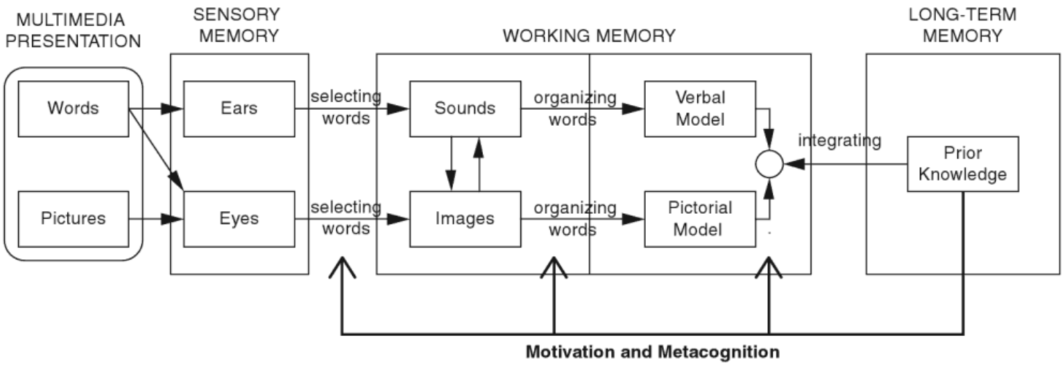


Figure 3.1: Diagram from Cognitive Theory of Multimedia Learning [6].

For instance, to reduce the *split attention effect* during multimedia presentations which can happen, for example, when the eyes are processing both text and pictures (as seen in Figure 3.1), they suggest *off-loading* the information. By instead presenting the words in the form of narration, it is possible to process the words in the verbal channel (Figure 3.1: *words* →*ears*) whereas the animation or picture can be processed in the visual channel (Figure 3.1: *picture* →*eyes*). This would reduce the processing demands from the visual channel and allow the learner to pick the most relevant aspects of the picture for further processing (Figure 3.1: *eyes* →*images*) [6].

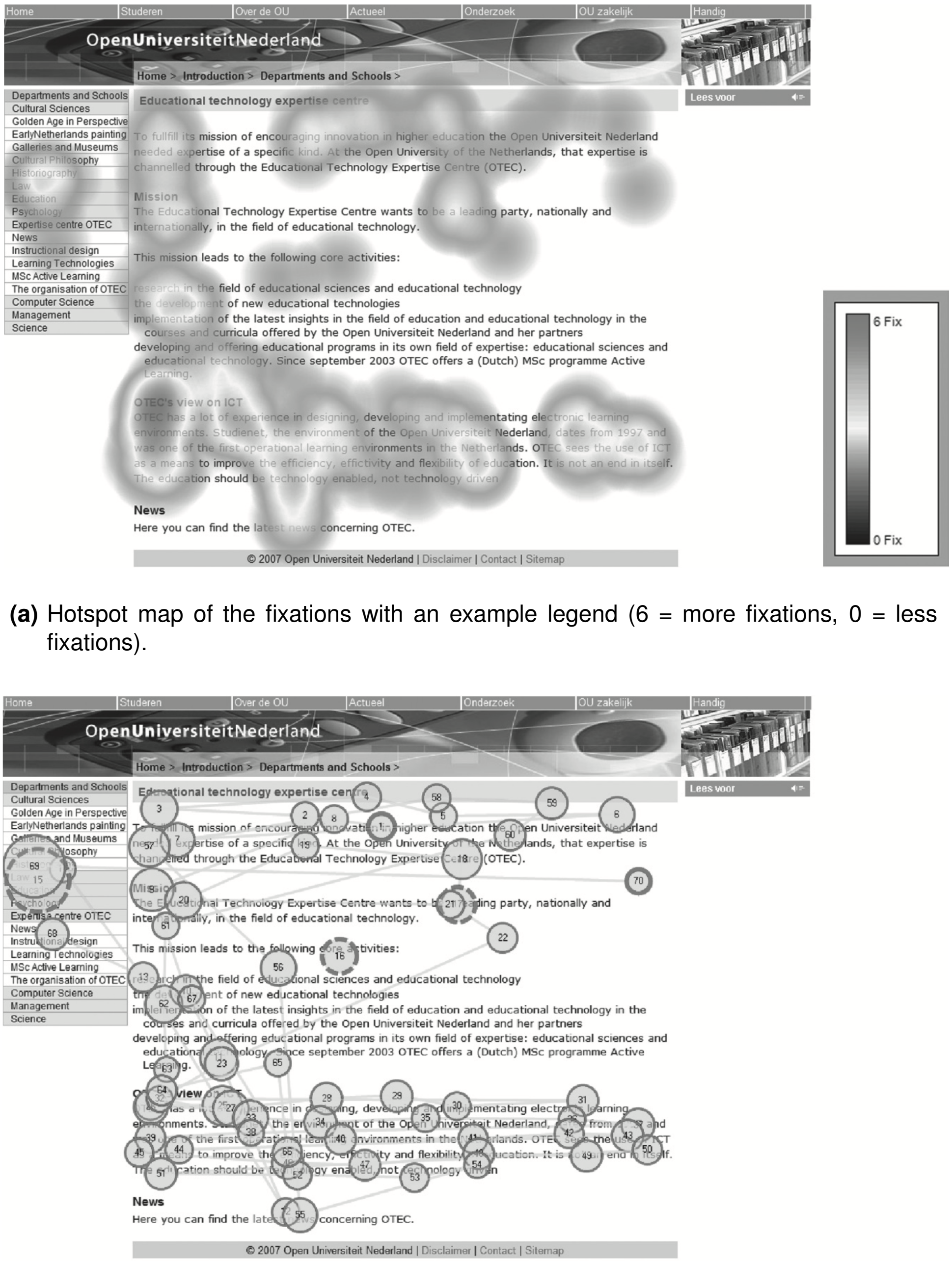
Following the principles of Cognitive Load Theory of Multimedia Learning could be useful, for example, when designing and developing an educational game.

3.1.2 Eye-tracking technique

A technique with great potential of providing useful insights related with the cognitive resources used in computer-based learning environments is the **eye-tracking technique** [28], and research on this topic has been increasing in the last few years [7]. It consists of recording eye movement data while users are working on a task. Through this technology, as seen in Figure 4.2, two main measurements can be observed: **Fixations** and **Saccades** [7].

Fixations, as the name implies, refer to the periods of time in which the eye was stable looking at a particular area or point in the screen. These are usually represented as a color-coded "hotspot" image in the multimedia environment (Figure 3.2(a)).

Saccades describe the quick eye movements between dwell times, which represent the changes in focus of the learner during the time it was measured (Figure 3.2(a)).



Educational technology expertise centre

To fullfill its mission of encouraging innovation in higher education the Open Universiteit Nederland needed expertise of a specific kind. At the Open University of the Netherlands, that expertise is channelled through the Educational Technology Expertise Centre (OTEC).

Mission
The Educational Technology Expertise Centre wants to be a leading party, nationally and internationally, in the field of educational technology.

This mission leads to the following core activities:

research in the field of educational sciences and educational technology
the development of new educational technologies
implementation of the latest insights in the field of education and educational technology in the courses and curricula offered by the Open Universiteit Nederland and her partners
developing and offering educational programs in its own field of expertise: educational sciences and educational technology. Since september 2003 OTEC offers a (Dutch) MSc programme Active Learning.

OTEC's view on ICT
OTEC has a lot of experience in designing, developing and implementating electronic learning environments. Studienet, the environment of the Open Universiteit Nederland, dates from 1997 and was one of the first operational learning environments in the Netherlands. OTEC sees the use of ICT as a means to improve the efficiency, effictivity and flexibility of education. It is not an end in itself. The education should be technology enabled, not technology driven

News
Here you can find the latest news concerning OTEC.

© 2007 Open Universiteit Nederland | Disclaimer | Contact | Sitemap

(a) Hotspot map of the fixations with an example legend (6 = more fixations, 0 = less fixations).

(b) Gaze Plot.

Figure 3.2: Images exemplifying the two measurements from the Eye-Tracking Technique [7].

It is believed that there is a relationship between eye movements and cognitive processes [31]. According to cognitive theory of multimedia learning, there are three cognitive processes that build coherent mental representations: **Selecting**, **Organizing** and **Integrating** [32].

- *Selecting* can be identified by the first fixations and represents the learners' inner selection of the relevant elements to carry into WM.

- *Organizing* occurs when the learner connects words or images in their WM. Longer fixation indicate deeper processing [7], and can also be used to determine CL – since there seem to be a functional link between the item that is being fixated and the cognitive processing allocated to that item [33].

- *Integrating* is the process that connects the organized information (the previously organized pictorial or the verbal models) with relevant prior knowledge. This process can be observed by looking at the transitions between pictures and text.

When using the eye-tracking technique and after a comprehensive analysis of these three processes, it is possible to estimate the amount of resources used by the WM during a multimedia presentation or an individual slide.

This approach could also be useful to assess CL in a game where, for instance, the player needed to interact with multiple interfaces. If successful, it could substantially facilitate the work of UX/UI engineers when constructing elements of a game such as navigation, usability and ergonomics.

3.2 Cognitive Load in Games

Direct studies of CL in game-based environments are extremely limited, making this a very interest field to study and experience on. Slava Kalyuga and Jan L. Paas explore in detail [34] the importance of CLT and why it should be taken into consideration in educational games. They discuss the different methods that could be applied, in order to improve learning and the consolidation of schemas in our LTM.

Other study explores how can CLT and the **4C/ID framework** (Figure 3.3) be applied in game design [9]. The 4C/ID model is a four-component instructional design model based on CLT developed in the early 1990s [8], that aims to train complex cognitive skills by minimizing extraneous CL, while increasing germane CL. The basic concept of the model is that complex learning environments can always be described in terms of four interrelated blueprint components that are central to complex learning [35]: **Learning Tasks**, **Supportive Information**, **(Just-In-Time) JIT Information** and **Part-Task Practice**.

- *Learning Tasks*: Concrete, authentic, whole-task experiences provided to the learner that primarily aim induction, i.e. promote schema construction.

- *Supportive Information*: Information that supports the non-recurrent aspects of the learning tasks. Functions as a bridge that connects the learners' prior knowledge with the learning tasks, by establishing relationships between novel information and what is already known by the learner.

- *(Just-In-Time) JIT Information*: The information that is prerequisite to the learning tasks (*e.g.* rules, principles). Apart from being used in the learning tasks, JIT Information is also relevant to the Part-Task Practice.

- *Part-Task Practice*: Practice items provided to the learners that aim to promote rule automation for selected recurrent aspects of the whole complex task.

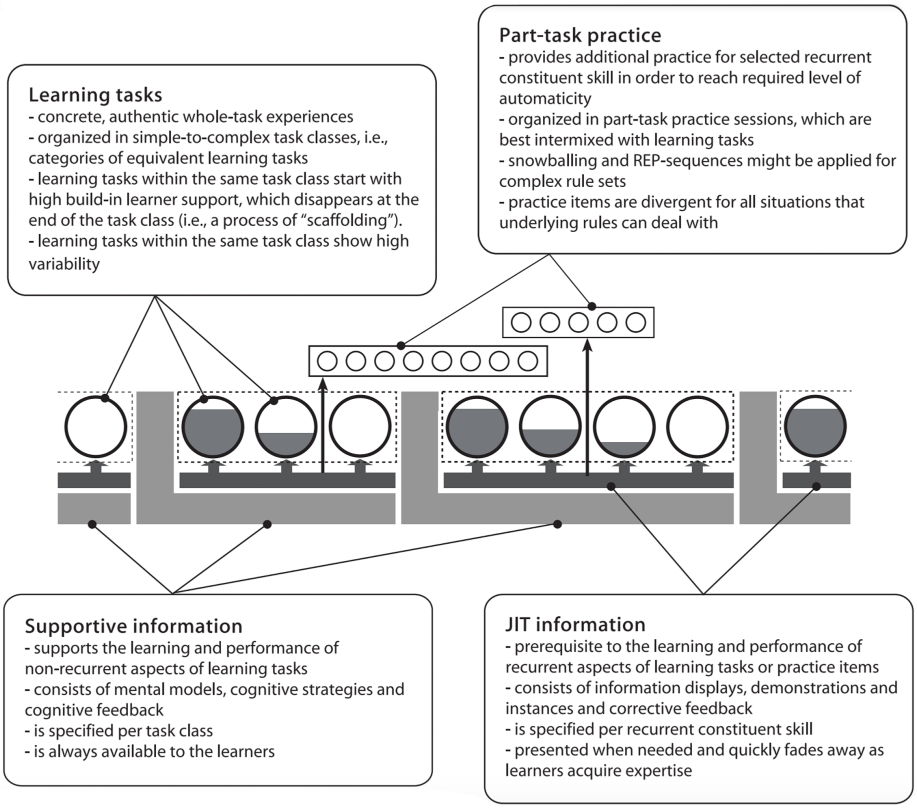

Learning tasks
- concrete, authentic whole-task experiences
- organized in simple-to-complex task classes, i.e., categories of equivalent learning tasks
- learning tasks within the same task class start with high build-in learner support, which disappears at the end of the task class (i.e., a process of "scaffolding").
- learning tasks within the same task class show high variability
Part-task practice
- provides additional practice for selected recurrent constituent skill in order to reach required level of automaticity
- organized in part-task practice sessions, which are best intermixed with learning tasks
- snowballing and REP-sequences might be applied for complex rule sets
- practice items are divergent for all situations that underlying rules can deal with
Supportive information
- supports the learning and performance of non-recurrent aspects of learning tasks
- consists of mental models, cognitive strategies and cognitive feedback
- is specified per task class
- is always available to the learners
JIT information
- prerequisite to the learning and performance of recurrent aspects of learning tasks or practice items
- consists of information displays, demonstrations and instances and corrective feedback
- is specified per recurrent constituent skill
- presented when needed and quickly fades away as learners acquire expertise

Figure 3.3: 4C/ID Framework [8].

The article [9] explores how can the 4C/ID-model design be applied in multiple game characteristics, with the goal of optimizing the design outcome. They started by listing 12 game characteristics that could potentially support the schema construction process (*e.g.* Challenge, Competition, Rules, Goals,...). For each of them, a hierarchical list of priority components from the model that ranged from 1 (lowest priority) to 3 (highest priority) was assigned, resulting in a radar graph for each game characteristic (as exemplified in Figure 3.4). The radar graph in each game characteristic visually depicts the design priority suggested in the book.

With a priority list based on the 4C/ID model assigned to each game characteristic, educational game developers can obtain a comprehensive view of how to properly create their games without disturbing the limitations of the learners' WM.

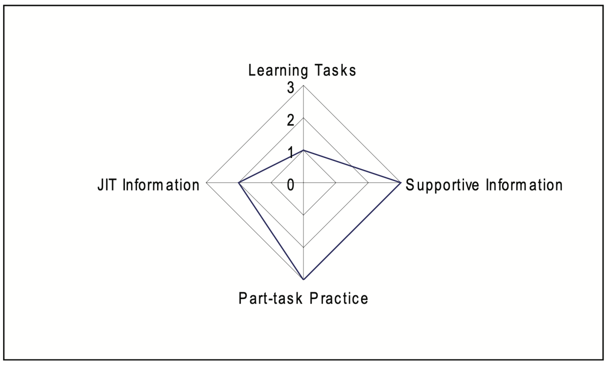


Figure 3.4: Example of a graph radar for the *Goals* game characteristic [9].

3.3 Summary

This chapter reviews some of the work done throughout the years that also aimed to control or assess CL in multimedia environments.

It starts by highlighting how can some of the typical negative repercussions of multimedia learning environments affect the learners CL, and overviews some studies that tried to mitigate this problem [6]. It also explores how the objective eye-tracking technique can be used in multimedia environments (*e.g.* a presentation, a slide or a website) to assess the users' CL [7]. The chapter ends with a review of a study that tried to minimize extraneous CL while increasing germane CL in educational video games, by exploring how can the 4C/ID-model design be applied in multiple game characteristics [9].

Even though one of the main functionalities of the WM is to process and consolidate novel information, its functions go way beyond learning. It is a cognitive system responsible for a wide panoply of functions such as guidance of behavior, translating instructions into action plans, considering alternatives, understanding the relation between items or ideas, amongst numerous others [36].

Our goal is to find a way that assesses CL and can easily be integrated into games, whether educational or not. A simple formula that collects gameplay data and estimates the CL that the player's WM is using during the gameplay experience, without the need of modifying the core components of the game. An effective, automatic and non-intrusive way that informs game designers about the mental workload that their games are demanding from the players – and that is precisely what we will explore in detail in the next chapter.

4

Implementation

Contents

Our goal is to create a tool-set that assesses CL in a non-intrusive way and can easily be integrated into video games. This chapter will explore the approach we took to achieve this goal. We will start by explaining how can the TBRS hypothetically be integrated in games and then describe the implementation of the game we have developed from scratch – to fully and freely test our hypothesis.

4.1 Time-Based Resource Sharing in Games

In order to obtain reliable CL results, we need to be coherent with existing studies so far related with the WM and its models. After analyzing and discussing each Memory Model, we came to the conclusion that a slightly altered version of the TBRS Memory Model could be suitable for the purpose of this work – mainly due to its simple formulation of the CL, based on attention-shifting.

This Memory Model [4] from 2004 was introduced in Chapter 2 (Section 2.3.3) however, its relevance for our work is of extreme importance. Therefore, we will do a more in-depth analysis of this model, focusing on its possible usages in the context of our work.

A central concept in cognitive psychology is the concept of – **bottleneck**. We all have independent systems that simultaneously process information; however, their capacity is limited and, as such, not all information is processed. For example, we can easily follow a conversation with a friend in a crowded restaurant; but listening and perceiving every conversation that's happening simultaneously, becomes impossible. Our attention, in this case, needs to be properly managed (if I pay attention to my friend's conversation, I'm following what he's saying; but if my attention is shifted towards processing the next table's conversation, I immediately lose track of what my friend is saying). *TBRS, in its third assumption [4] (see Chapter 2, Section 2.3.3), grabs this concept to argument that the WM has a central bottleneck that only allows one process at the time – one's attention is either shifted towards processing or maintaining information in WM*.

To simplify the theory behind the model, picture a relatable scenario: You are a student attending to a lecture and the professor is explaining some important theory about the subject; the usual procedure is listening and taking notes so you are able to revise them later and consolidate that information on your LTM. However, while listening, your attention is shifted towards processing that novel information. If you switch to maintenance (*i.e.* try to memorize the information) on the spot, you lose track of the class because your attention is not allocated to process the information anymore. This is, in a nutshell, how the TBRS Memory Model describes the functioning of the WM.

Apart from its simple way in explaining how the WM functions, the fact that it has a clean and elegant formula solely based on time that calculates CL during an activity, was another major reason why we leaned towards this model. *We were curious to find out whether or not its simplicity would be capable of capturing the subtleties of tasks typically observed in video games*. In short, the models' formula boils down to the fact that CL results from the total amount of time in which a participant was fully paying attention to a task, divided by the total duration of that task (Equation (4.1)).

$$CL = \frac{\sum_{i=1}^{N} a_i}{T} \tag{4.1}$$

As mentioned earlier, CL is mainly referred in learning scenarios; since the WM includes reasoning, decision-making and problem solving processes (amongst numerous others) for the information being held – which is, in fact, crucial for learning. However, there is a wider panoply of activities that also require these types of processes such as, more and more frequently, video games.

But how can TBRS specifically be applied to video games? How can one be sure of what aspect of the game the player is paying attention to at any time of the gameplay? Well, the correct answer is that we can never be entirely certain if the player is fully paying attention to a game or not. There are multiple factors in a video game that can grab the player's attention. But imagining it is a game testing scenario and, since it is their job, the testers have to be focused on the experience. There are certain actions that hint that a player is paying attention at a specific time of the gameplay. If, for example, the game has really challenging bosses and the player has to constantly dodge the enemies' attacks; or if it is a puzzle type game and the player has to interact with multiple objects; or if the player opens a User Interface (UI) and is actively interacting with it. These are all indicators that, in that specific moment of the game, the player's attention was shifted towards processing some kind of information from the game.

If the periods of time in which the players' attention was most certainly being captured, and having the total duration of the gameplay, a formula similar with the TBRS to measure CL could be applied and could be a lead towards assessing CL in games in an intrinsic, non-intrusive way.

We therefore propose to adapt the CL TBRS formula (Equation (4.1)) to games, where the a_i will reflect the time of the i_{th} *Attention-Grabbing Event (AGE)* (further explained in Section 4.2.3), whilst the T refers to the *total gameplay time*.

With this being said, we have created a game where we put the concepts of this theory in practice, and the next section (Section 4.2) will explain and go through the whole process of developing it.

4.2 The Game: Way Out

To test our hypothesis, we decided to create a game from scratch; since it didn't impose restrictions in our creativity and gave us the necessary flexibility to create a satisfying game environment in which it made sense to fully test our hypothesis.

4.2.1 Concept

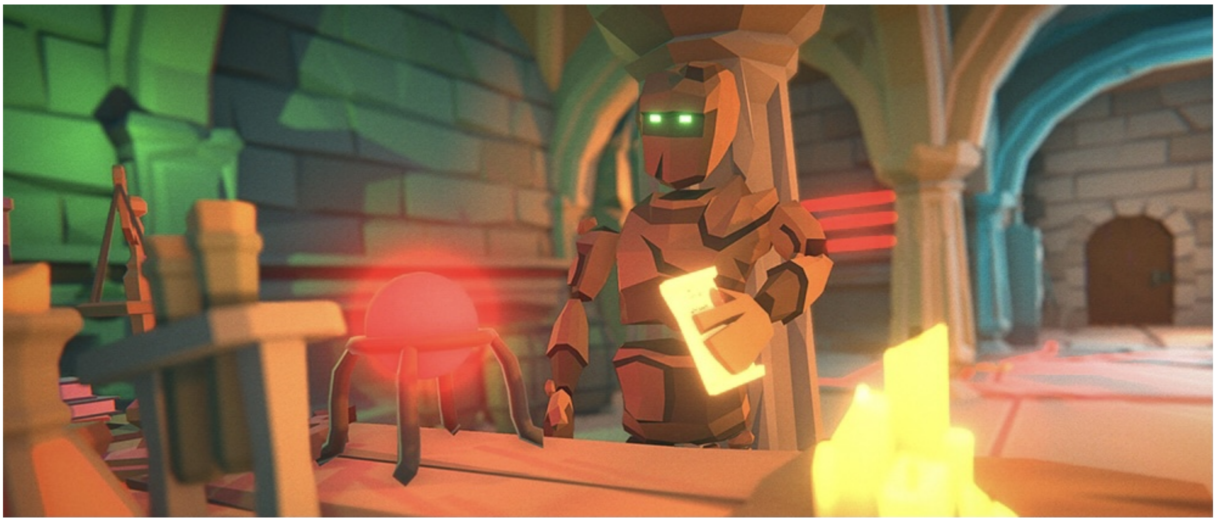


Figure 4.1: Way Out screenshot.

Way Out (Figure 4.1) is a 3D low poly puzzle game where you play as a golem who just woke up in a mysterious laboratory and is trying to figure out the purpose of his existence. To do so, he has to solve puzzles and challenges in a dungeon-like environment to both progress through the map and find clues about himself.

The hidden plot is that a human scientist has become the first to achieve full conscience transmutation. The puzzles the golem has to solve were created by the golem himself in his human form, and are a simple way to determine if the scientist's cognitive and reasoning skills have remained intact in his new body.

However, due to the nature of this work, the small demo that we have created and tested mainly explores the puzzles and challenges of the game and not the plot itself. In the following subsections we will have an in-depth analysis of each particularity of our game and the reason for its implementation.

With the goal of analysing possible CL variations, a total of four versions of the game were developed. Each version's puzzles had particular tweaks and changes to analyse this eventual discrepancy. These key particularities of the game will also be explored in the following subsections.

4.2.2 Development Environment

The engine upon which the game was built is Unity 3D[1] – a cross-platform game engine designer to support and develop 2D and 3D video games, computer simulations, virtual reality environments (Unity, 2020), among numerous others.

[1] https://unity.com/

It was also the engine of choice due to our previous experience with it and its satisfying results. Furthermore, due to the uncertain final result of our model, we were unsure about what type of game and challenges we wanted to implement; and the engine provided the needed flexibility for customizing and easily adapting our game according to early feedback, and the different stages of development.

At first, we also considered modding an existing game to implement our model, but given the problems and restrictions it could bring in the future, we opted for the safer option of implementing a game from scratch.

4.2.3 Attention-Grabbing Events

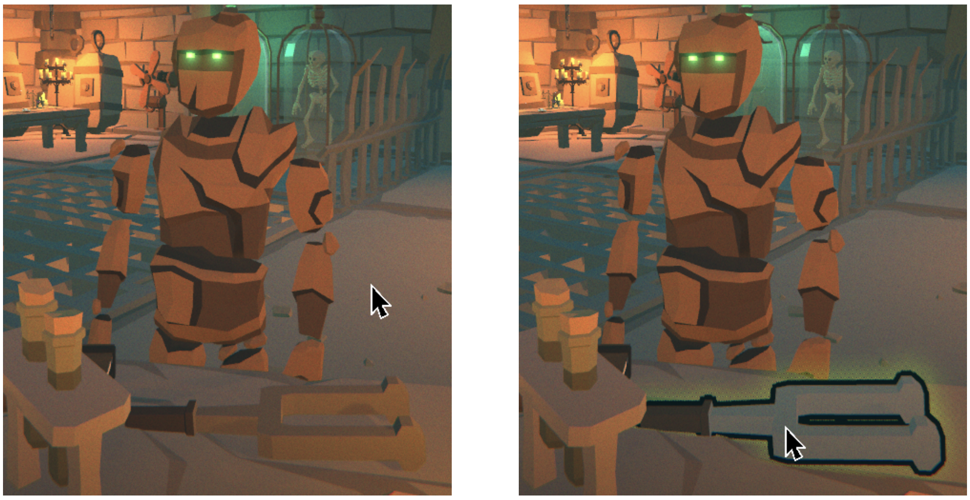
According to TBRS memory model, CL is the result of the total attention time of a task divided by the total time of that task (Equation (4.1)).

With the goal of adapting the TBRS attention-shifting principles and its CL formula to game development, we decided the following: even though they are most likely very distinct from game to game, through any game, the player has to execute certain actions or events to progress, which usually take a certain chunk of time to be performed. Whenever one of these events occurs, its duration (*i.e.* the time since the event begins until it ends) could be translated into a period of time in which the player's attention was supposedly shifted towards processing information. We call these – **Attention-Grabbing Events (AGE)**.

Having the total gameplay and AGEs duration, it is theoretically possible to recreate the TBRS CL calculation. However, since all games are different and we are trying to generalize our model to cover any game type, *we highly emphasize that the game designers are the ones who should ponder and choose the AGEs, taking into account the type of game being developed.*

This being said, we will now explain which events were considered attention-grabbers in our game and the reason behind that consideration. In Way Out, we considered the following AGEs:

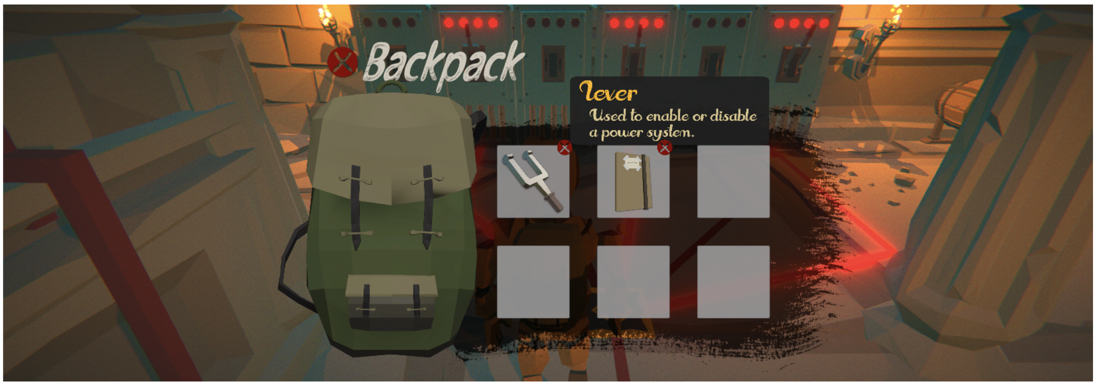
- **Object Interactions**: Since the game is oriented towards puzzle solving, interacting with objects is mandatory. Thus, even though object interactions usually only take fractions of a second, we consider the duration of each interaction as an AGE – hence, throughout the game, we count the total amount of time spent hovering interactive objects (Figure 4.2(b)) and add it to the formula.
- **Interfaces**: To solve different puzzles, querying certain UIs can be extremely useful. Thus, whenever an interface is opened (*e.g.* Inventory, Notebook, Sphere placeholder menu) we assume that the player is paying attention to the game during that period of time (Figure 4.3). Additionally, in order mitigate any question related with the game controls, a "Help" button was also implemented – this allows the player to access the game's instructions at any time during the gameplay (just like the others, the active time of this interface is also counted as an AGE.)



(a) Mouse outside an interactive object.

(b) Mouse inside an interactive object.

Figure 4.2: Way Out: Object Interactions.

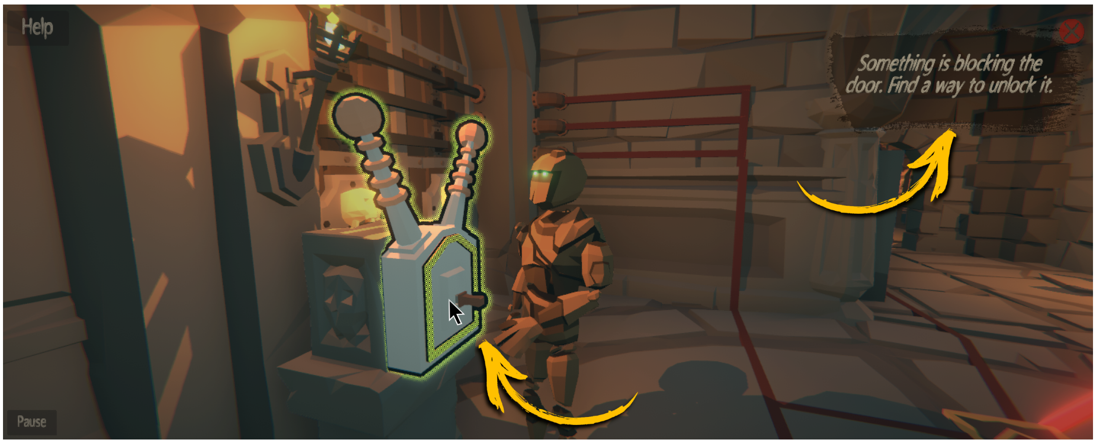


Used to enable or disable a power system.

Figure 4.3: Way Out: Interface Example (Inventory).

- **Notifications**: While playing, the players will encounter certain challenges and puzzles (Section 4.2.5) which, without any guidance, could unnecessarily increase the gameplay time. Thus, we added a notification system that, when needed, displays useful tips and warnings on the top right corner of the screen (Figure 4.4). Notifications are displayed by default for 5 seconds but, if desired, the player can opt to close them before. We assume that for the short time that a notification is displayed, the players' attention is shifted towards processing the information that it provides. Thus, the period of time that a notification is displayed on the screen is also considered an AGE.

However, the overlap of events would interfere with the formula, since it would mean that the player's attention would be shifted towards processing multiple events at once. An example of this happening can be observed in Figure 4.4, where the player is interacting with an object whilst a notification is simul-



Something is blocking the door. Find a way to unlock it.

Figure 4.4: Way Out: Notification Example.

taneously being displayed. Pressing the button triggered the notification, but the player kept hovering the interactive object.

Therefore, in order to mitigate these temporal overlays, we found useful to create an hierarchy for our game's AGEs (Figure 4.5).

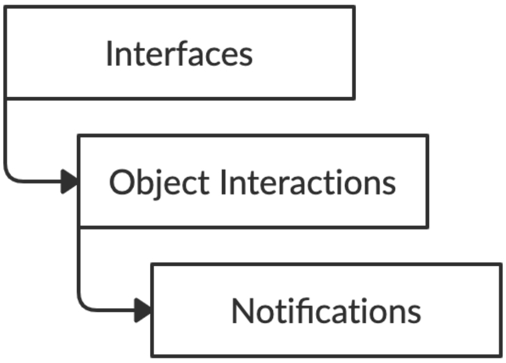


Figure 4.5: Attention-Grabbing Events (AGEs) hierarchy.

We intentionally designed the interfaces to occupy a large chunk of the screen, so we assume that whenever the player is actively interacting with an UI element (*e.g.* the is Inventory open), object interactions and notifications become disabled, mitigating a possible overlap of attention.

The same principle applies to object interactions – if the player starts interacting with an object while a notification is being displayed on the screen, it is assumed that the player's attention is being shifted towards the interaction, not taking into consideration notification's display time in the equation.

Listing 4.1 is a simplified version of the code, that exemplifies how this hierarchy was implemented; where *InterfacesActive*, *Interacting* and *NotificationsActive*, are booleans that dynamically vary between *true* or *false*, depending on the player's interactions throughout the gameplay. Apart from the *TotalAtten-*

tionTime, we also register the time spent in each type of AGE individually.

```
void Update(){
    // ...
     if(InterfacesActive){
        InterfacesActiveTime += Time.deltaTime;
        TotalAttentionTime += Time.deltaTime;
    }if(!InterfacesActive && Interacting){
        InteractionsTime += Time.deltaTime;
        TotalAttentionTime += Time.deltaTime;
    }if(!InterfacesActive && !Interacting && NotificationsActive){
        NotificationsDisplayTime += Time.deltaTime;
        TotalAttentionTime += Time.deltaTime;
    } // ...
}
```

Listing 4.1: Simplified version of the code that represents the AGEs hierarchy.

4.2.4 Game Versions

According to the TBRS memory model, a task is more cognitive demanding when it requires a larger amount of attention-shifting. In the case of our game, as mentioned in the previous section (Section 4.2.3), we consider object interactions, notification and UI display times as our main AGEs. Therefore, theoretically, for the same gameplay duration, the greater the number of AGEs, the greater the value of the CL will be.

That being said, and since the intrinsic difficulty of a task is proven to be correlated with higher levels of CL [23], we started by creating two versions of our game – **"Easy"** and **"Hard"** – with the *goal of analysing if the data collected from the Easy versions would indicate lower levels of CL, when compared with the Hard ones*.

Both versions contain the exact same type of puzzles, challenges and possible interactions. However, the puzzles from the Easy version were intentionally twisted to require a lesser number of interactions for its resolutions, which would overall result in a less amount of AGEs.

Furthermore, we will also wanted to validate our model focusing on **time**. Once again, TBRS defends that the CL is the result of the attention time dedicated to a task divided by the time of that task [4]. However if the player is, for example, trying to solve a problem and has to move through the map without interacting with any objects, the gameplay time is counting but the attention time is not which, according to the formula (Equation (4.1)), would result in a lower CL. And this is precisely the point that we want to verify. *If, for the same puzzles, in order to solve them the player is forced to move around the map, would this increase, maintain or decrease the player's CL?*

Table 4.1: Game Versions.

	Normal Movement	**Additional Movement**
Easy	A1	B1
Hard	A2	B2

Thus, as seen in Table 4.1, we ended up creating four versions of the game.

Two that solely explore the contrast between the intrinsic difficulty of the game – **A1** and **A2** – where all the items required for the resolution of the puzzles are all relatively close to each other, not forcing the player to move through the map in order to solve them. Note that *"Normal Movement"* means there is no extra movement, *i.e.* all the items required for the resolution of the puzzle are relatively close to each other.

The other two – **B1** and **B2** – beyond exploring the intrinsic difficulty of the puzzles, also explore the effects that additional movement and, therefore, gameplay time has on the player's CL. In these versions, the items required for the resolution of the puzzles are scattered around the map, forcing the player to move more through the map in order to solve them – hence the *"Additional Movement"* in Table 4.1. This will theoretically increase the overall gameplay time and, consequently, using the TBRS CL formula, decrease the CL.

The puzzles implemented were designed to verify the effects that the variations in AGEs and gameplay time had on the players' CL. For that purpose, we have developed two main puzzles that slightly vary between the four versions of the game.

To test the discrepancies between the players' attention through the versions (A1 vs A2 and B1 vs B2), both puzzles require more or less AGEs for their resolution. To test the differences in gameplay time through the versions (A1 vs B1 and A2 vs B2), we changed the items disposition between versions of the puzzles – forcing the player to move more or less, depending on the version played.

The following subsection will explore in detail the implementation of the game puzzles for each version of the game.

4.2.5 Game Puzzles

To test our hypothesis, we created two main puzzles – the Lever Puzzle and the Orb Puzzle. This section will describe the puzzles and explore their differences, according to the game version being played.

4.2.5.A The Lever Puzzle

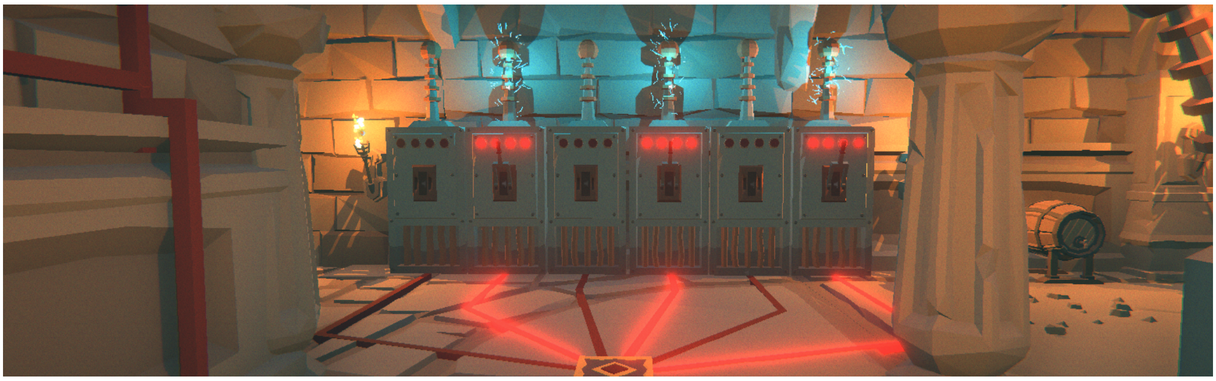


Figure 4.6: Way Out: Lever Puzzle.

The first main puzzle of our game – The Lever Puzzle[2] – requires the player to move 6 levers to unlock a door. Initially, of the 6 levers, only 3 are correctly positioned and ready to move (Figure 4.6). For all four versions of the game, the player has to first find the 3 missing levers and place them on the machines that still require one.

The difference between the Easy and Hard versions of the game are the effects each lever has on the machines. In the Easy versions (A1 and B1), each lever only affects its machine. For instance, Lever #3 only affects the state of Machine #3, turning it on (lever up) or off (lever down). Therefore, the easy versions' solution is fairly simple – the player solely has to find and place the missing levers correctly and turn on the machines (by moving each lever up).

On the other hand, in the Hard versions of the game, each lever movement can affect the state of multiple machines. For example, Lever #5 affects the state of Machines #4, #5 and #6, by either turning them on or off depending on their current state at the time. This leads to a theoretical higher number of interactions – and AGEs – since the solution is not as straight forward as the opposite versions.

Additionally, in the hard versions, the notebook has a different initial note that helps the players by indicating the effects that each lever has on the machines (Figure 4.7). Theoretically, this will also influence the overall attention time, since it is expected for the players to pay more attention to the notebook in these versions.

[2] A full walk-through of the four versions of the Lever Puzzle is available at: https://www.youtube.com/watch?v=6LSi81yiB28&t=3s&ab_channel=AlbertoRamos

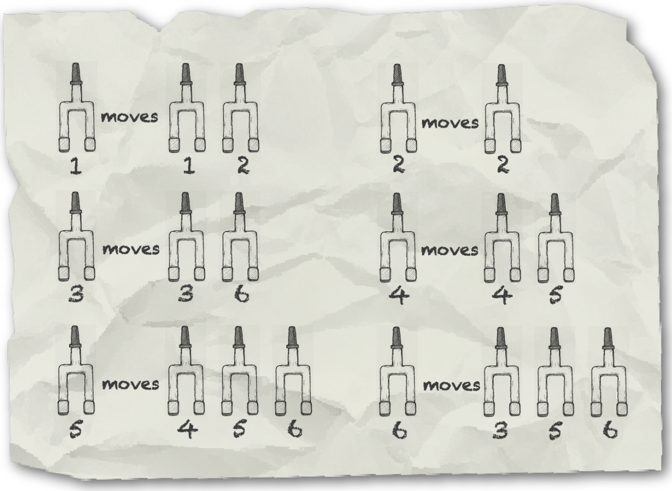
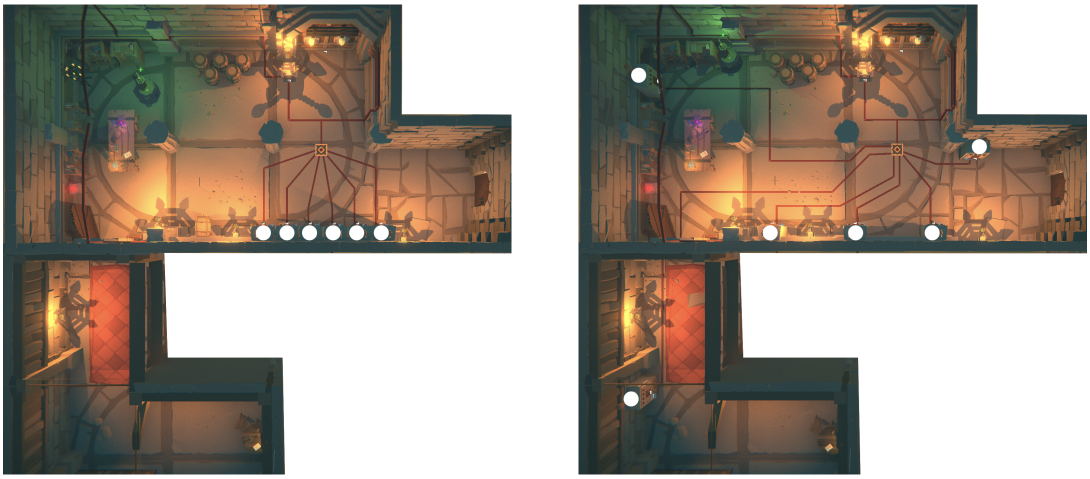


Figure 4.7: Lever Puzzle - Hard Versions' initial note.

When it comes to the players' movement, the A and B versions of the game differ from the position of the machines – represented as white dots in Figure 4.8.

In both the A versions (Figure 4.8(a)), all the machines are close to each other, allowing the player to clearly see the effect that each lever interaction has on the puzzle. While in the B versions (Figure 4.8(b)), the machines are scattered around two rooms. Hence, to analyse the effect that each lever interaction has on the puzzle, the player has to move around.



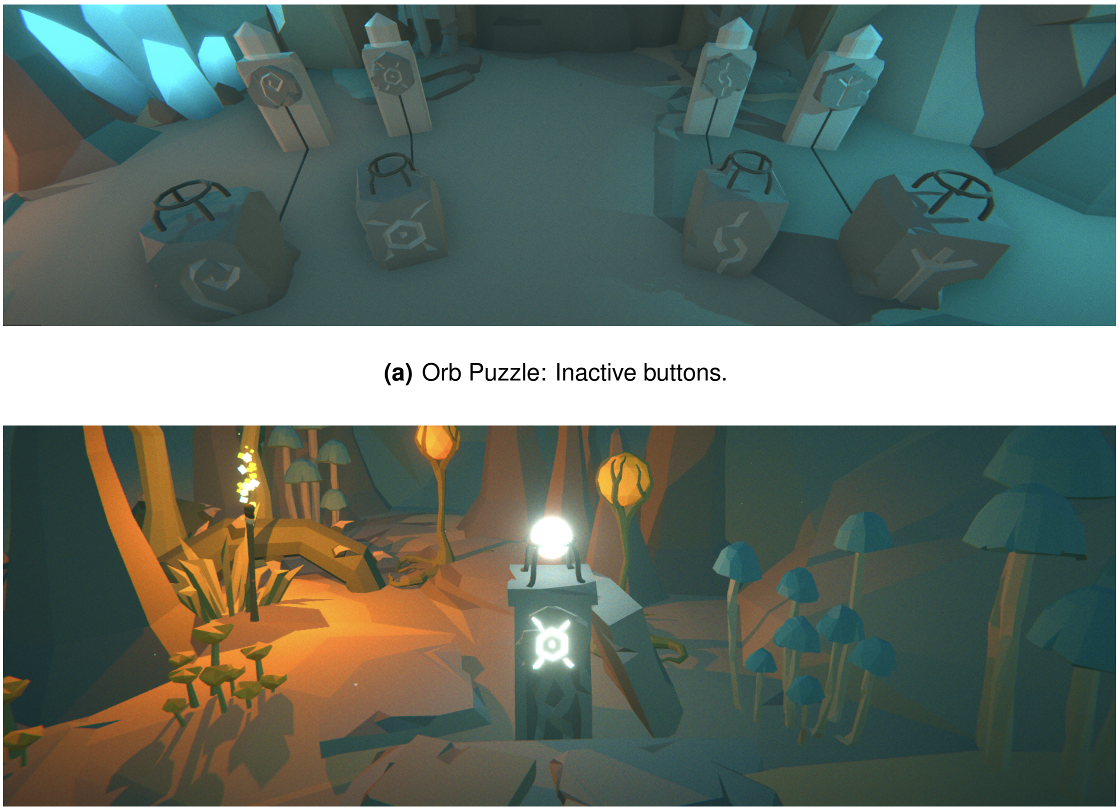
(a) Lever Puzzle: Versions A.

(b) Lever Puzzle: Versions B.

Figure 4.8: Way Out: Lever Puzzle (Versions A and B).

4.2.5.B The Orb Puzzle

The second puzzle of our game – The Orb Puzzle[3] – requires the player to press four buttons (identified by distinct runes) in a specific order to unlock the final door. However, each button needs a power source – a glowing orb. The orbs are initially scattered around the map on "orb stands". Below each stand, a rune is emitting light of the same color as the orb (see Figure 4.9(b)). The runes match each orb to the different button it activates. For example, observing both Figure 4.9(a) and Figure 4.9(b), the white orb activates the second button (from the left), because they share the same rune.



(a) Orb Puzzle: Inactive buttons.

(b) Orb Puzzle: White Orb on its initial stand.

Figure 4.9: Way Out: Orb Puzzle.

When all buttons are powered (*i.e.* all the orbs are correctly placed), the player is able to press them and unlock the final door. However, they need to be pushed in a specific order that the player can discover by trial and error, or by finding a note that contains that information and add it to the Notebook.

Onto the different versions of the game, when it comes to the intrinsic *difficulty* of the puzzle ,*i.e.* the comparison between the Easy and Hard versions, the main differences are the following:

- In the hard versions, whenever the player fails the correct button sequence or powers a button with

[3] A full walk-through of the four versions of the Orb Puzzle is available at: https://www.youtube.com/watch?v=8ge565wPE9I&feature=youtu.be&ab_channel=AlbertoRamos

the wrong orb, every rune color change. This forces the player to change the position of every orb, since they now power a different button (Figure 4.10), leading to a much higher number of AGEs. In contrast, in the easy versions, if the player fails the sequence he simply has to repeat it from the start.

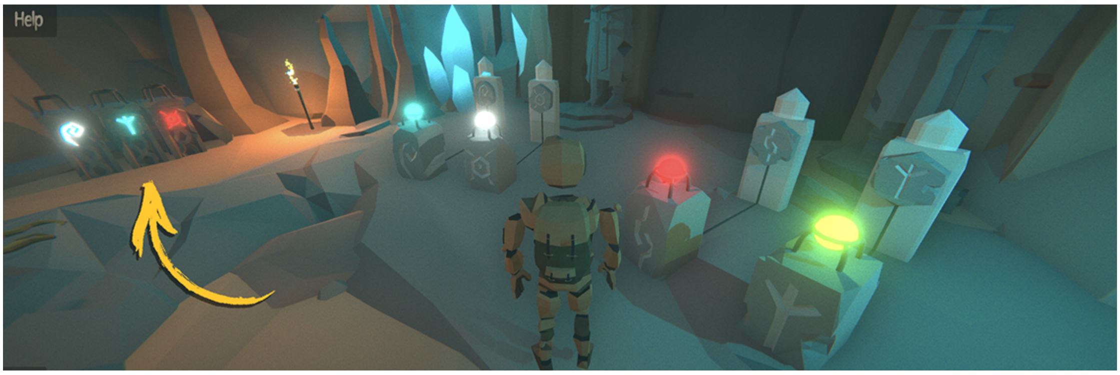
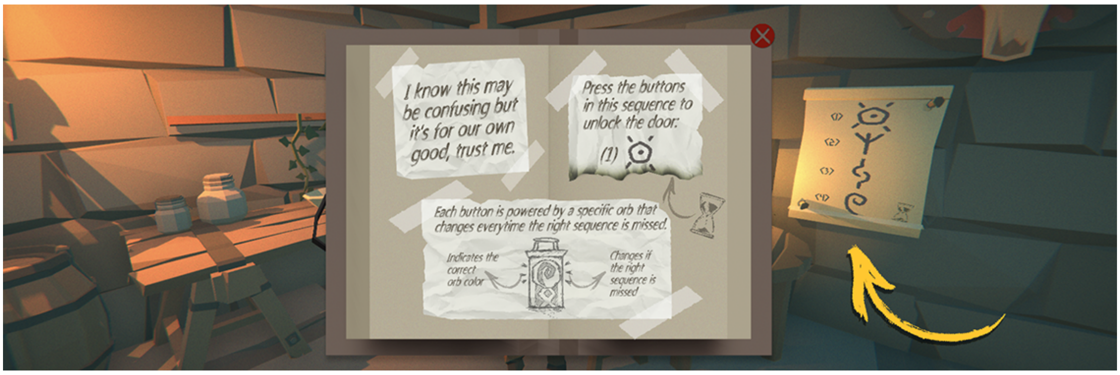


Figure 4.10: Orb Puzzle (Version A2) Example: Player failed the sequence so the rune colors changed.

- In the easy versions, if the player places an orb correctly, there is a light that ignites immediately, functioning as a visual indicator that the player has made the right choice. On the other hand, in the hard versions, this light only ignites once the player starts the sequence and presses the buttons in the correct order.
- Contrary to the easy versions, the note that contains the information regarding the correct button sequence is burned in the hard versions. However, in the same room where the note is located, there is an intact poster that also contains the correct sequence (Figure 4.11). The player cannot take the poster with him nor add it to his Notebook, therefore the information about the sequence needs to be stored in WM.



I know this may be confusing but it's for our own good, trust me.

Press the buttons in this sequence to unlock the door:

(1)

Each button is powered by a specific orb that changes everytime the right sequence is missed.

Indicates the correct orb color
Changes if the right sequence is missed

Figure 4.11: Orb Puzzle (Hard Versions): Burned Note.

- Lastly, we added time pressure to the hard versions of this puzzle. Once the first button was pressed, the player only has 15 seconds to press the other three, and get the sequence right; if he fails the sequence, or does not manage to complete it in time, the rune colors change, forcing him to change the position of every orb (Figure 4.12). The easy versions do not have a time limit.

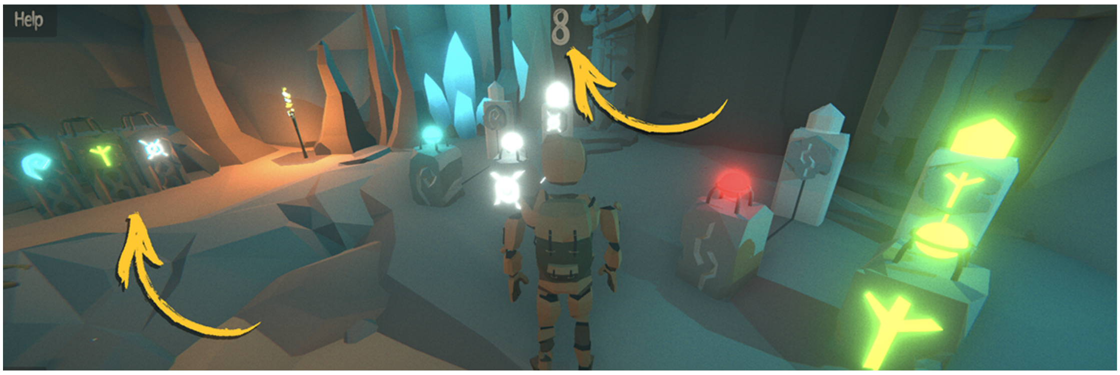
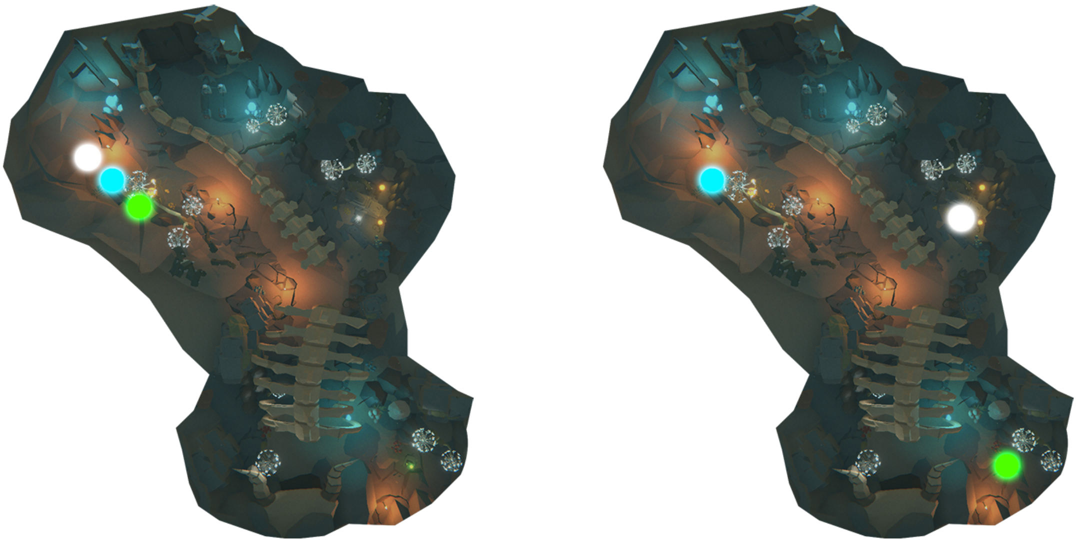


Figure 4.12: Orb Puzzle (Version A2) Example: time to complete the sequence is almost over.

When it comes to the players' movement, the only difference between the A and B versions of this puzzle is the initial position of the orbs. In the A versions (Figure 4.13(a)), the orbs are initially all close to each other and to the main puzzle area. Whilst in the B versions (Figure 4.13(b)) the orbs are initially scattered around the map; therefore, if the rune colors ever change, the player has to move throughout the map to verify the new color associated with each rune (*i.e.* each button), leading to a theoretical longer gameplay time.



(a) Versions A.

(b) Versions B.

Figure 4.13: Way Out: Orbs initial position (Versions A and B).

4.3 Data Gathering and Cognitive Load calculation

To follow the principles of the TBRS memory model and in order to use its formula [4], we need to collect relevant gameplay data that estimates the players' attention time during the game. Therefore and, as mentioned in a previous subsection (Section 4.2.3), beyond the *Total Gameplay Time*, we will mainly collect data related with the duration of the multiple AGEs that occur throughout the game, namely:

- *Interface Interactions:* The time spent with the Inventory, Instructions, Notebook and Sphere placeholder menu opened.

- *Object Interactions:* The time spent interacting with interactive objects.

- *Notifications:* Time in which notifications were shown on screen.

The total sum of AGEs mentioned above will return the **Total Attention Time** (Equation (4.2)) during the game. The **Total Attention Time** will after be used as the dividend in the adapted TBRS CL formula; whereas the **Total Gameplay Time** will be the divisor (Equation (4.3)).

$$TotalAttentionTime = \sum_{i=1}^{N} AGE \tag{4.2}$$

$$CL = \frac{TotalAttentionTime}{TotalGameplayTime} \tag{4.3}$$

Furthermore, in order to support any possible unexpected values, we also collected the number of times each type of AGE happened (*e.g.* number of times the Inventory was opened).

When the player completes the game, all the data listed above is stored on an online Google Sheets document for further analysis. Additionally, when the game ends, a randomly generated code name is given to the player and also stored on the same online Google Sheets document. This code name functions as a bridge that allows us to compare the gameplay data with the questionnaire answers (further explained in the next chapter).

Apart from the data listed above, we also need to store the version played by the player (*e.g.* A1, A2, B1 or B2). When the player starts the game, a query is made to the online Google Sheets document to verify which version was played by the last participant – so a different one can be loaded – as Figure 4.14 shows. If there is no stored data, a random version is loaded. This system aims to evenly distribute the total amount of players across all four versions of the game.

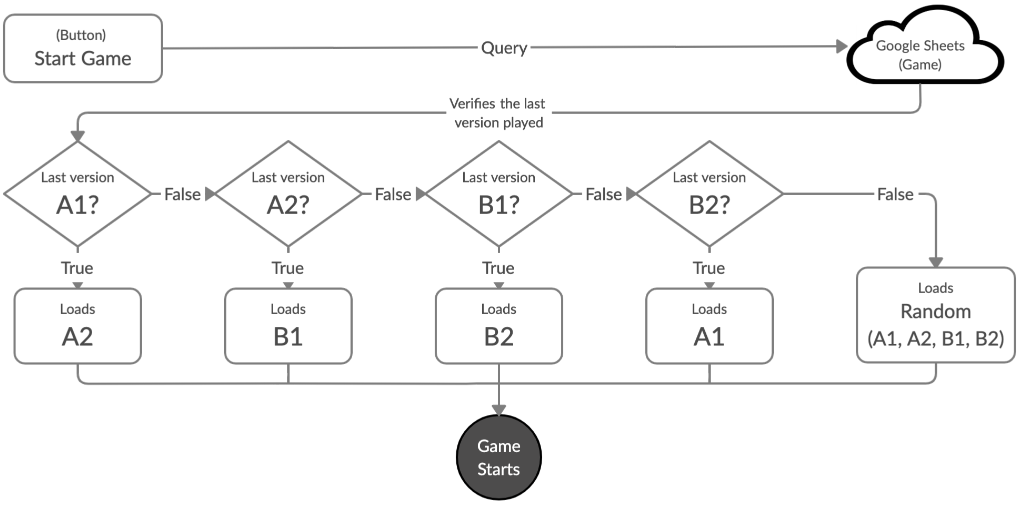


Figure 4.14: Diagram showing how the game version is selected when the game initiates.

4.4 Summary

In order to estimate the players' CL, we developed a game from scratch called Way Out (Section 4.2) where we applied the attention-shifting principles of the TBRS memory model (Section 4.1) to collect relevant gameplay data.

We adapted the CL TBRS formula – Equation (4.1) to Equation (4.3) – to our game. The a_i from the original formula now reflects the time of the i_{th} AGE (Section 4.2.3), whilst T refers to the Total Gameplay Time.

In the case of our game, we consider the periods of time of the following events as attention-grabbers (Section 4.2.3): Object Interactions, Active Interfaces, Notifications. However, since we are trying to generalize our model to cover any game type and, since all games are different, we highly emphasize that the game designers are the ones who should ponder and choose the AGEs, taking into account the type of game being developed.

Four versions of the game Way Out were created (Section 4.2.4) in order to analyse the possible CL variations when the difficulty (A1 vs A2 and B1 vs B2) and gameplay time (A1 vs B1 and A2 vs B2) increases or decreases. In order to obtain a similar number of participants in all versions, we followed the structure shown in Figure 4.14 to select which version is loaded when a player starts the game.

The entire project is available in a public GitHub repository[4] and the game can be downloaded from a website[5] (available for Windows and MacOS).

[4] https://github.com/albertoramos1997/WayOut

[5] https://web.tecnico.ulisboa.pt/ ist194117/Tese/

5

Procedure and Results

Contents

This chapter will start with a systematic description of our procedure and its structure; followed by the in-depth analysis of the obtained results.

5.1 Procedure

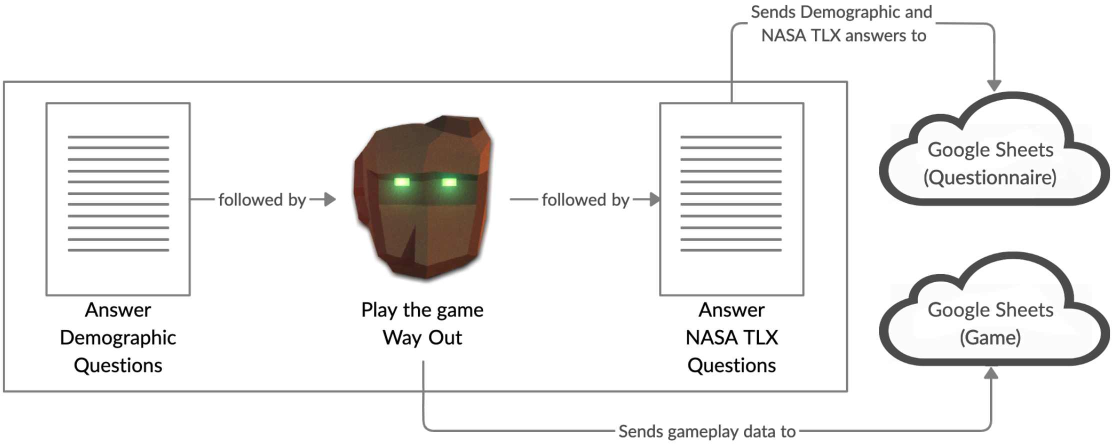

Sends Demographic and NASA TLX answers to
Answer Demographic Questions
followed by
Play the game Way Out
followed by
Answer NASA TLX Questions
Google Sheets (Questionnaire)
Google Sheets (Game)
Sends gameplay data to

Figure 5.1: Procedure to acquire data.

In short, the structure of the followed procedure is summarized in Figure 5.1, and the entire questionnaire can be seen in Appendix A.

With the goal of seeking basic information about the respondents and understand where they fit in the general population, the *first* part of the procedure consists of asking the participants the following demographic questions: "Age", "Gender", "Mother Tongue", "How often do you play video games?", "Do you enjoy point and click puzzle games?".

Once collected, this data allows us, if needed, to divide the population of respondents in various groups, which will be useful in the overall analysis.

In the midst of the questionnaire, after answering the demographic questions, the participants are asked to play the game Way Out, which is the *second* part of the procedure – extensively explained in the previous chapter. After playing and finishing the game, a random code name is generated and provided to the player, so it can be pasted onto the questionnaire, linking the game data with the questionnaire answers.

The *third* and final part of the procedure consists of asking the participants questions related with their workload during the game, in order to validate our hypothesis. To do so, we need to compare the game data that may affect the CL with an existing valid and trustworthy method that accurately measures the workload of a task. In a general sense we are examining the "workload" experienced by the player during the gameplay. Cognitive Load and Mental Workload are often used as synonyms and the relationship between workload factors and CL types was analyzed in depth by Galy, Cariou and Mélan (2011) [23].

For that purpose, we will use the **NASA TLX** [25] which, as mentioned on a previous chapter (Chapter 2), is a subjective workload assessment tool technique that divides the workload in six different subscales: Frustration, Temporal Demand, Physical Demand, Mental Demand, Performance and Effort. Therefore, after playing the game, the participants were asked to answer a few questions related with their overall workload during the game. Since this questionnaire was only briefly introduced in Chapter 2, the following subsection (Section 5.1.1) will clarify its technicalities, with the goal of better understanding how it actually assesses CL.

5.1.1 How to assess CL with NASA TLX

The NASA TLX consists of two parts: **weights** and **ratings**.

Generally, the first requirement is for the participant to evaluate the contribution of each subscale – its weight – of the workload during the gameplay (the weights themselves also provide diagnostic information as the nature of the workload imposed by the game).

To do so, there are 15 possible pairwise comparisons of the six subscales of workload. Each pair (for instance, Temporal Demand versus Mental Demand) is presented at the time, and the subject has to choose the member of each pair that contributed more to the workload of the task performed (in our case, the game). At the end of every pairwise comparison, we count the number of times that each subscale was selected instead of the others. It can range from 0 (never selected in a pairwise comparison) to 5 (selected in every pairwise comparison) – this is the resulting weight assigned for that specific subscale.

The second requirement is to obtain individual numerical ratings for each subscale – which reflect the magnitude of that factor in the game. Thus, the respondents are asked to rate each subscale individually from 0 to 10 or 0 to 100 (least to most taxing).

Notice that it is common to switch the order of the NASA TLX parts, *i.e.* start with the ratings and end up with the pairwise comparisons (weights), or even discard the pairwise part entirely (RAW TLX [26]).

The adjusted ratings for each of the six subscales of the workload is computed by multiplying their respective weight with their raw rating (Equation 5.1). For example, if the weight and rating of Temporal Demand was 4 and 50 respectively, its Adjusted Rating would be 4 x 50 = 200.

$$AdjustedRating = Weight * RawRating \tag{5.1}$$

Using the NASA TLX, the overall workload of a task, *i.e.* its resulting CL, is the result of the sum of the Adjusted Ratings divided by 15 (which is the total amount of pairwise comparisons) (Equation (5.2)).

$$Workload_{NASATLX} = \frac{\sum AdjustedRatings}{15} \tag{5.2}$$

Using the shorter version of the questionnaire instead – the RAW TLX – the resulting Workload of the task would simply be the sum of the individual Raw Ratings divided by 6 (total amount of subscales) (Equation (5.3)); since there are no pairwise comparisons in this simplified version of the questionnaire.

$$Workload_{RAWTLX} = \frac{\sum RawRatings}{6} \tag{5.3}$$

5.2 Pilot

Before broadening the experience to a larger sample of participants, we opted to first test it with a small sample – aiming to correct eventual game bugs and to better understand whether the questionnaire was adequate. During this phase, we specifically asked the participants to be extra critical and transparent, since our goal was precisely to adjust any eventual flaws with the experience.

Apart from a few game bugs pointed out, a consistent feedback received during this phase was that the pairwise comparisons, at the end of the questionnaire, were somewhat confusing. Some even went as far as saying that the comparisons looked all very similar and that “in the end, they selected almost randomly”.

However, since the RAW TLX is a “trimmed” version of the NASA TLX without the pairwise comparisons, we ended up providing the full version of the questionnaire in the actual experiment; with the premise that the first thing to analyse was the possible discrepancies between the two versions (NASA TLX versus RAW TLX) – and whether it was justified to use the shorter version of the questionnaire instead, when analysing and comparing the collected data.

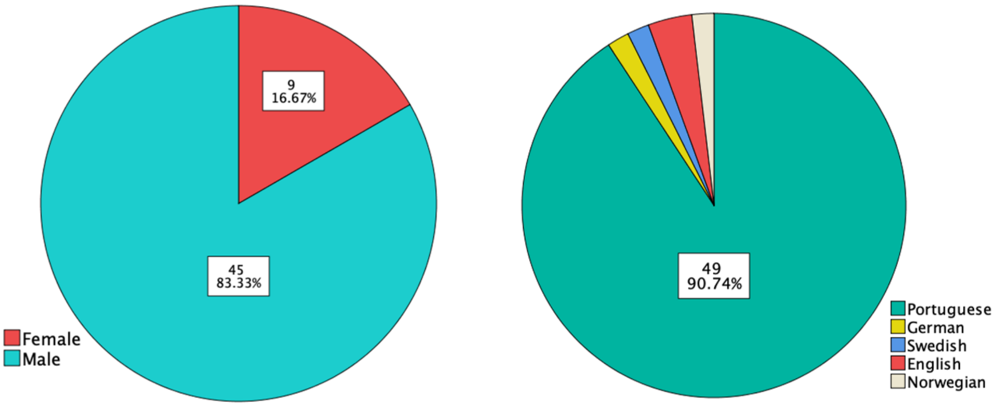
5.3 Sample

In total, we had a convenience sample of 54 participants responding to the questionnaire and, as linearly as possible, playing a version of the game. It is important to emphasise that all tests were done remotely. Hence, the experiment was advertised in multiple social platforms – namely Discord, Facebook and Instagram.

To analyse the obtained results, we used the software **SPSS Statistics (V26)**[1] from IBM; where all the NASA TLX calculations were made and the charts, graphs and tables presented in this chapter were generated. That being said, this section will focus in analysing the demographic answers of the total population of respondents.

[1] https://www.ibm.com/products/spss-statistics

From the 54 participants, 45 (83.33%) identified themselves as males whilst 9 (16.67%) identified as females (Figure 5.2(a)). The majority of our respondents (90.74%) speak Portuguese as their native language while the other 10% speak others (such as English, Norwegian, German and Swedish, as seen in Figure 5.2(b)).



(a) Pie chart with the gender information of 54 users – with 45 male individuals and 9 female answers.

(b) Pie chart with the mother tongue information of 54 users – with 90% native Portuguese speakers and the other 10% speaking others.

Figure 5.2: Pie charts: Gender and Mother Tongue.

Looking to Figure 5.3, we can conclude that the age of our participants ranged from 16 to 44 years old (M=23.63, SD=4.136).

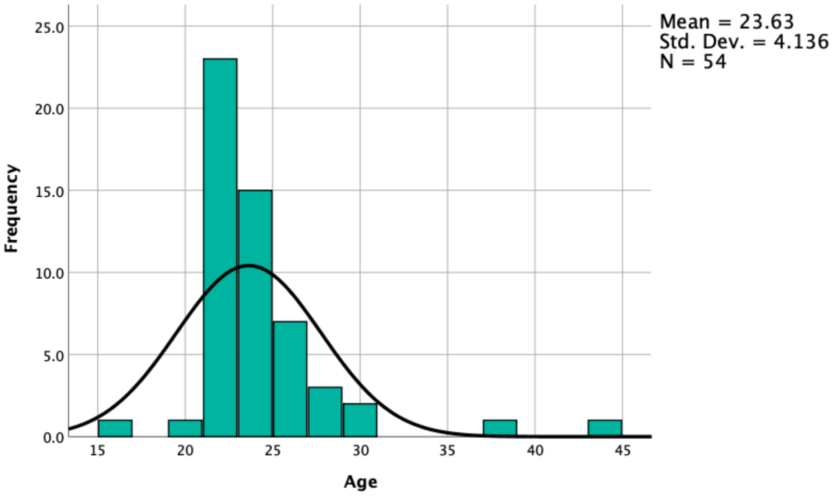
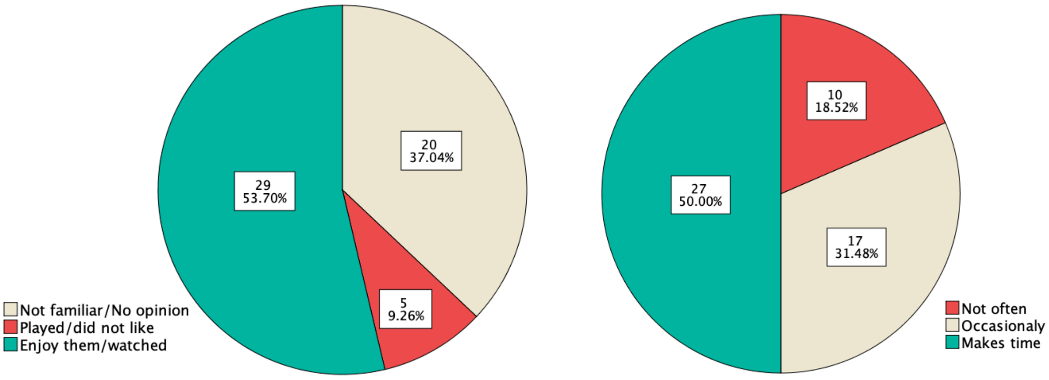


Figure 5.3: Bar chart with the different ages of the total sample of respondents.

When asked how frequently they play video games, half (27 – exactly 50%) responded that they "made some time in their schedule to play video games", 17 respondents answered that they "play occasionally" and 10 "do not play video games often" (Figure 5.4(a)).

Lastly, when asked about how familiar they were with this game genre, 29 (53.70%) of our respondents answered that they "enjoyed and have played/watched others play multiple times", 20 (37.04%) were "not familiar or did not have a formed opinion" and only 5 (9.26%) "did not appreciate these types of games" (Figure 5.4(b)). This data (Figure 5.4) has proven to be useful to justify eventual unexpected gameplay results, as we will look further in the following section.



(a) Pie chart with the answers of 54 respondents, when asked if they enjoyed the type of game that we have developed.

(b) Pie chart with the answers of 54 respondents, when asked if they play video games often.

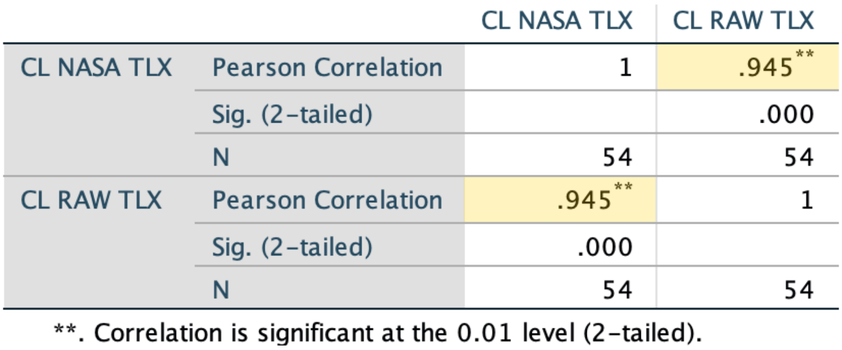
Figure 5.4: Pie charts: "Enjoys Playing" and "Plays Often".

5.3.1 The questionnaire of choice: NASA TLX

To clarify a question brought up during the pilot phase (Section 5.2), we started our analysis by comparing the questionnaire results with and without the pairwise comparisons, *i.e.* by comparing the NASA TLX (Equation (5.2)) with the RAW TLX (Equation (5.3)) – to choose which version of the questionnaire would be more suitable to validate our hypothesis.

We found that there was a high positive correlation between the CL reported by the two versions of the questionnaire, with a nearly perfect *Pearson*[2] correlation of 0.945 (as seen in Figure 5.5).

[2] "The *Pearson* correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r, which can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases."



		CL NASA TLX	CL RAW TLX
CL NASA TLX	Pearson Correlation	1	.945**
	Sig. (2-tailed)		.000
	N	54	54
CL RAW TLX	Pearson Correlation	.945**	1
	Sig. (2-tailed)	.000	
	N	54	54

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5.5: Bivariate Correlation (*Pearson*) between the CL from the NASA TLX and RAW TLX.

Since it is typically more common to use the full version of the questionnaire, and due the high positive correlation observed with its trimmed version, we opted to solely validate the gameplay data with the full version of the questionnaire – the NASA TLX.

5.4 Hypothesis

According to our model, we hypothesise that the CL values will be higher in the Hard versions – A2 and B2 – where, theoretically, more AGEs occur. Additionally, we also want to observe the repercussions in CL when, for the same type of puzzles, the items required for their resolution are scattered around the map (A1 and A2 versus B1 and B2), forcing the player to move more and, consequently, increasing the overall gameplay time. More specifically, we were interested in finding out whether or not the hypothetical increase of gameplay time would affect the CL. Would it increase it, because the players were more consciously focused in shifting attention towards maintenance – to avoid forgetting the relevant and required items? Or would it decrease because the players have more time to process and maintain the information required for the resolution of the puzzles in WM?

The following subsections aim to clarify the questions above and the validity of our hypothesis. We will, therefore, mainly analyse the data that we believe is relevant for that clarification, namely: *total gameplay* and *attention times*; CL from the game; CL from NASA TLX.

5.4.1 Gameplay Time Results

To clarify our hypothesis, we started by analysing **gameplay time** (Figure 5.6) where, interestingly enough, we noticed that both the A versions took, in average, slightly longer to complete than the B versions. We ran a *Kruskal-Wallis*[3] test that showed that there was at least one pair of significantly dif-

[3] "Rank-based non-parametric test that can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable."

ferent groups (H(3) = 18.74 ; $p \leq .001$). The pairwise comparisons[4] with a *Bonferroni* correction showed that the harder versions (A2 and B2) took significantly longer to complete than the easier versions ($p \leq .05$)[5] – comparing A1 with A2 and B1 with B2.

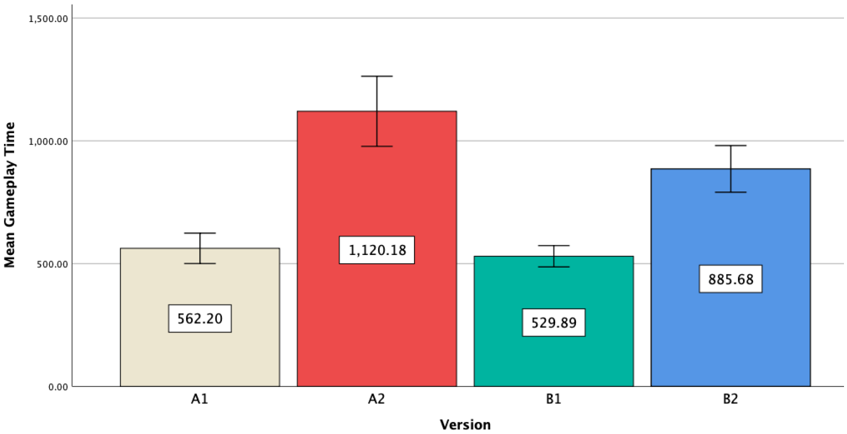
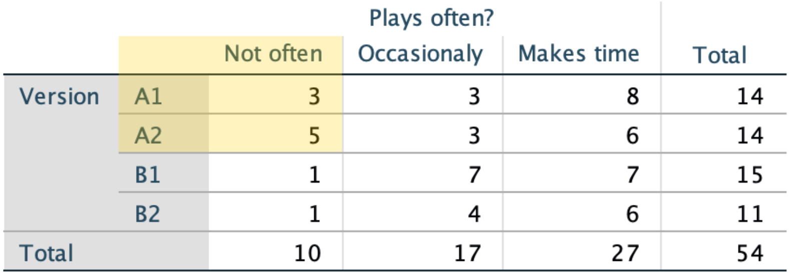


Figure 5.6: Simple Bar Mean of Gameplay Time by the game versions; Error Bars refer to the Standard Error of the Mean (SEM).

Due to the game versions implementation (described in Chapter 4, Section 4.2.4), these results were unexpected – since we tried to implement the game in a way that the B versions would result, in average, in a higher gameplay time than the A versions. Therefore, we analysed two main things: The first was whether or not the A versions had a higher number of participants that did "not play often" than the B versions. As seen in Figure 5.7, we found that to be indeed true.



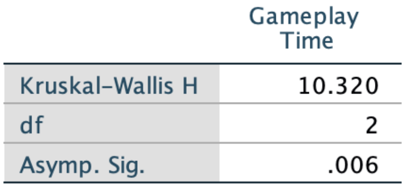
		Plays often?			
		Not often	Occasionaly	Makes time	Total
Version	A1	3	3	8	14
	A2	5	3	6	14
	B1	1	7	7	15
	B2	1	4	6	11
Total		10	17	27	54

Figure 5.7: Table showing the distribution of the "Gameplay Frequency" groups across all four versions of the game.

The second, was to analyse if there was, in fact, a significant difference in gameplay time between the different "gameplay frequency" groups (*i.e.* "Does not play often", "Plays occasionally", "Makes time to play"). We ran a *Kruskal-Wallis* test that confirmed that there was, indeed, a significant difference; H(2) =10.320, p = .006 (Figure 5.8(a)).

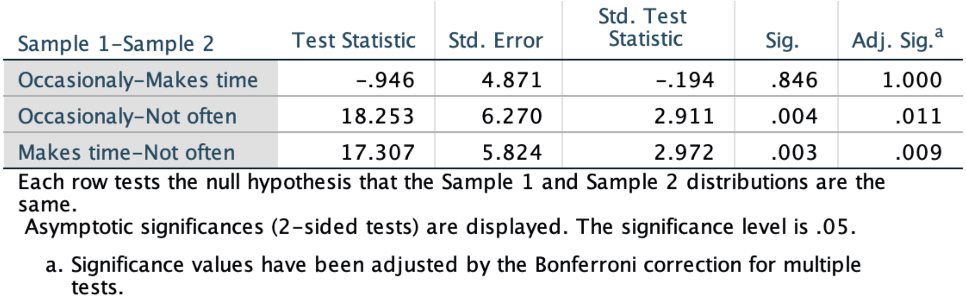
[4] "Pairwise comparisons refer to a statistical method that is used to evaluate relationships between pairs of means when doing group comparisons."

[5] p (sig.) $\leq$ 0.05 means that there is at least a pair of average gameplay times that is significantly different; whilst, on the other hand, p-value $>$ 0.05 tells us that there is no significantly different pair. The closer this value gets to 1, the higher the correlation is.



	Gameplay Time
Kruskal-Wallis H	10.320
df	2
Asymp. Sig.	.006

(a) *Kruskal-Wallis* test results with "Plays Often" as the grouping variable.



Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Occasionaly-Makes time	-.946	4.871	-.194	.846	1.000
Occasionaly-Not often	18.253	6.270	2.911	.004	.011
Makes time-Not often	17.307	5.824	2.972	.003	.009

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

(b) Gameplay Time across "Plays Often": Independent-Samples *Bonferroni* Test Summary.

Figure 5.8: *Kruskal-Wallis* test results and *Bonferroni* test summary.

The pairwise comparisons with a *Bonferroni* correction (Figure 5.8(b)) showed that there is a significant difference in gameplay time between the groups "Occasionaly - Not Often" with *p* = .004, and "Makes time - Not Often" with *p* = .003.

To confirm whether these results were due to the groups distribution, we decided to analyse them without the 10 respondents that answered "Not Often" – considering them, in this specific analysis (Figure 5.9), as outliers.

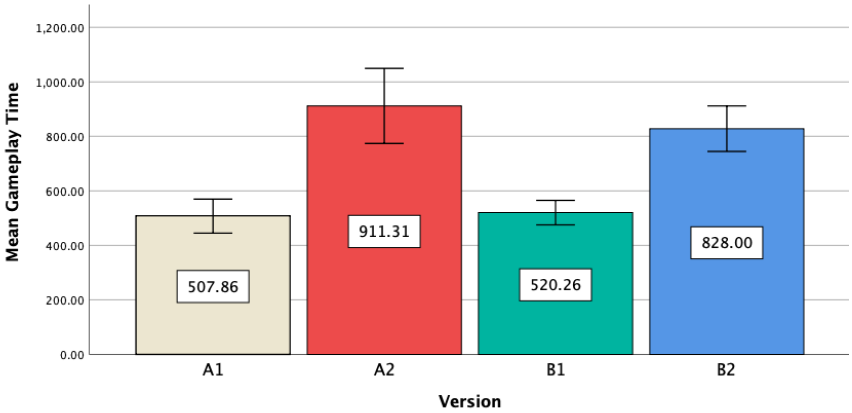


Figure 5.9: Simple Bar Mean of Gameplay Time by the game versions – without the group "Not Often"; Error Bars refer to the Standard Error of the Mean (SEM).

As seen in Figure 5.9, discarding the respondents that answered “Not Often”, the average gameplay time of the A versions decreased much more when compared with the B versions – A1 went from 562.20s to 507.86s and A2 from 1120.18s to 911.31s, while version B1 went from 529.89s to 520.26s and B2 from 885.68s to 828.00s. However, although closer, the average gameplay time between the A and B version was still very similar, which was not intended when designing the game.

This led us to conclude that our manipulation of the “Additional Movement” (B) versions was unsuccessful. Meaning that we were unable to answer one of the questions we initially had: “For two puzzles with a similar intrinsic difficulty, how would the variations in gameplay time affect the player's CL?”

5.4.2 Attention Time Results

Following the gameplay time, we analysed the other factor that, according to the TBRS, also influences the CL of a task – the **attention time**. Again, very briefly, for each player, the attention time results from the sum of the duration of every AGE during the gameplay. We ran a *Kruskal-Wallis* test to analyse the distribution of attention time across the different game versions (H(3) = 23.12; $p \leq .001$). The pairwise comparisons with a *Bonferroni* correction showed the same pattern found in gameplay time: A1 demanded significantly less attention time than A2 (p = .002) and B1, less attention than B2 (p = .011). As expected, the versions of the game with a higher difficulty (A2 and B2) had also, on average, a higher attention time (Figure 5.10).

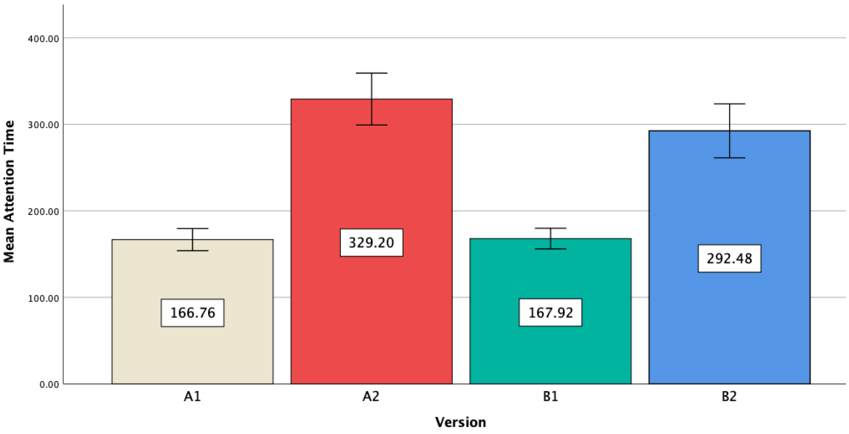


Figure 5.10: Simple Bar Mean of Attention Time by the game versions; Error Bars refer to the Standard Error of the Mean (SEM).

For statistical purposes – and to better understand our AGE approach to estimate the player's attention during the game – we decided to analyse the influence that each AGE had in the total attention time. Observing the pie chart of Figure 5.11, we can conclude that three AGEs took almost 80% of the total attention time; with Object interactions having 29.27%, Notebook interface with 27.54% and the notifications displayed taking 22.02% of the players total attention time.

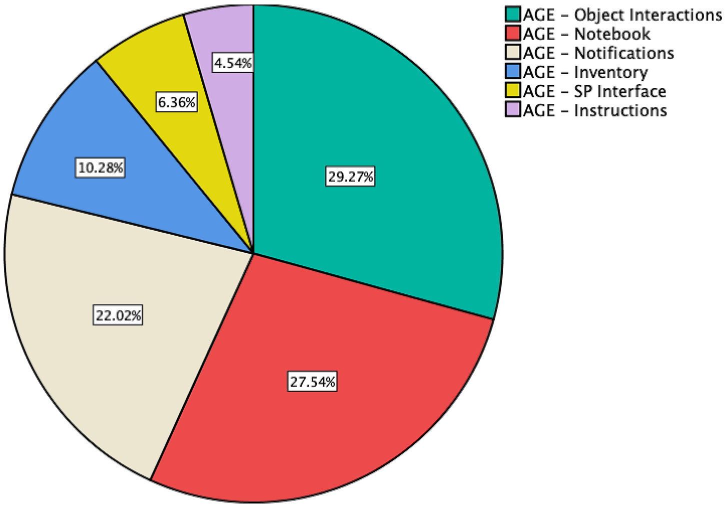


Figure 5.11: Pie chart with the distribution of AGEs that most influenced the attention time.

5.4.3 Resulting CL from the Game

Onto the actual **CL values** reported from the game (Figure 5.12), we can conclude they were very similar in every version (around 32%). We ran a *Kruskal-Wallis* test to analyse the distribution of the calculated CL (using the TBRS formula) across the different game versions (H(3) = .842; *p* = .839), and found that there were no statistically significant differences between the medians. These results, however, do not reflect the differences noticed in terms of gameplay and attention time; *meaning that, perhaps, the adapted TBRS CL formula, in its current form, is not sensitive enough to detect the variations across the versions.*

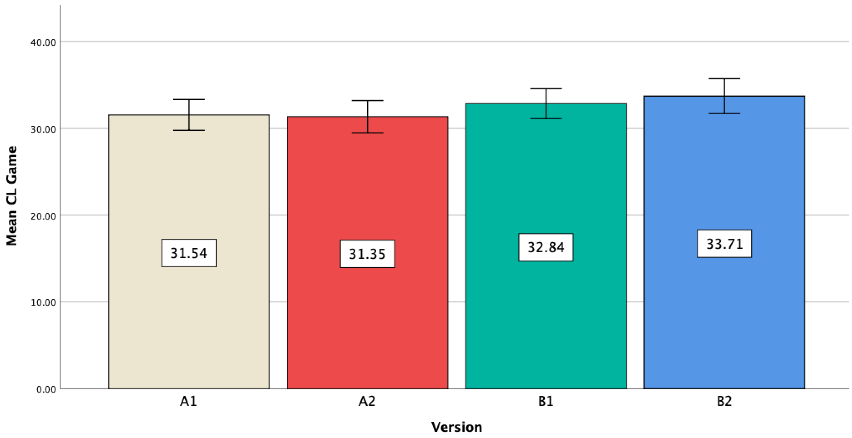
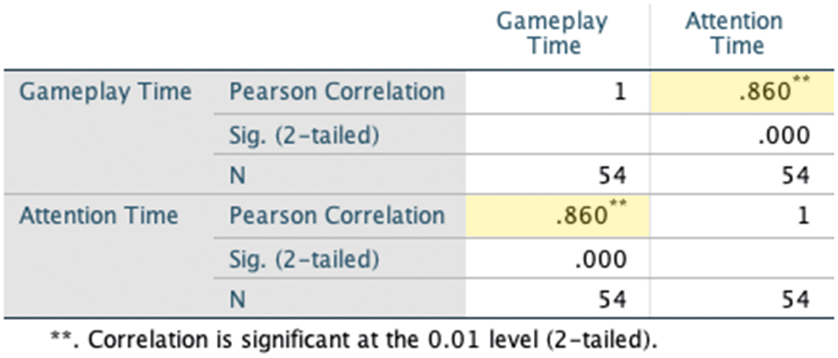


Figure 5.12: Simple Bar Mean of the players CL percentages by the game versions (using the TBRS CL formula); Error Bars refer to the Standard Error of the Mean (SEM).

Observing both the **gameplay time** (Figure 5.6) and **attention time** (Figure 5.10) bar charts, a noticeable pattern can be seen – a higher gameplay time appears to result in a higher average of attention time. To clarify this, we made a *Pearson* correlation between these two variables (Figure 5.13) and we ended up observing a high positive correlation of 0.860. This means that, whenever the gameplay time increases, there is a high chance that the attention time will also follow that path.



		Gameplay Time	Attention Time
Gameplay Time	Pearson Correlation	1	.860**
	Sig. (2-tailed)		.000
	N	54	54
Attention Time	Pearson Correlation	.860**	1
	Sig. (2-tailed)	.000	
	N	54	54

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5.13: Bivariate Correlation (*Pearson*) between the gameplay time and attention time.

If both the dividend and divisor have a high positive correlation (*i.e.* in the equation, when one increases/decreases the other also follows that path) – the resulting CL will always be similar regardless of the times spent in the game – which justifies the results obtained in Figure 5.12. A possible way to mitigate this problem would be by significantly restricting the gameplay time and, for instance, by asking the player to complete as many tasks as possible in the time limit. However, since our goal was to generalize our hypothesis to any game type, we opted not to add a time restriction in our implementation.

5.4.4 Resulting CL from NASA TLX

Onto the **NASA TLX scores**, there is a noticeable CL variation across the game versions (Figure 5.14), leading us to observe two main things:

- According to the NASA TLX, as predicted, the players that played the more challenging versions of the game (A2 and B2), reported higher values of CL during the game (comparing A1 and B1 with A2 and B2).

- The “Additional Movement” (B) versions appear to have induced a slightly higher percentage of CL when compared with their respective “Normal Movement” (A) versions. This inclines us to assume that the distance between crucial items for the game appears to, in some way, affect the CL (comparing A1 with B1 and A2 with B2). Nevertheless, as discussed previously (Section 5.4.1), the “Additional Movement” (B) versions were not successfully manipulated – preventing us from concluding anything concrete related to this topic.

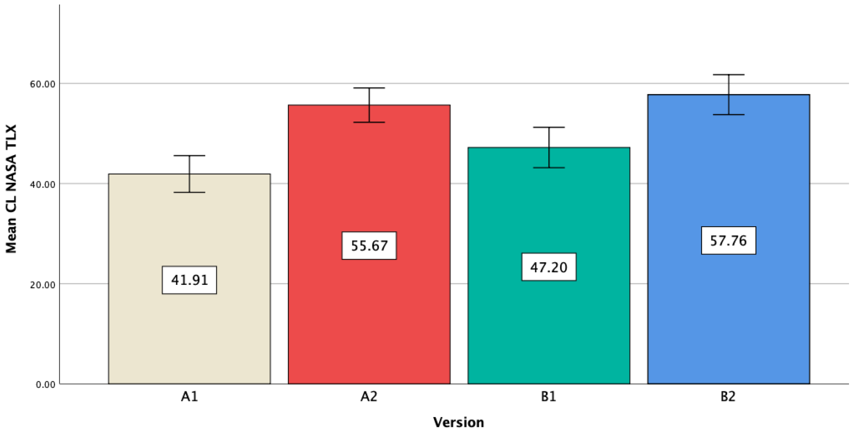


Figure 5.14: Simple Bar Mean of the NASA TLX's CL by the game versions; Error Bars refer to the Standard Error of the Mean (SEM).

Considering that only one of our manipulations was successful and that the "Additional Movement" (B) versions ended up with similar gameplay times than the "Normal Movement" (A) versions, we decided to merge both the "Easy" and "Hard" versions of the game, and test the NASA TLX scores for both groups. To clarify, in this specific analysis, the "Easy" group corresponds to the aggregate of versions A1 and B1; while the "Hard" group results from merging versions A2 and B2 (Figure 5.15).

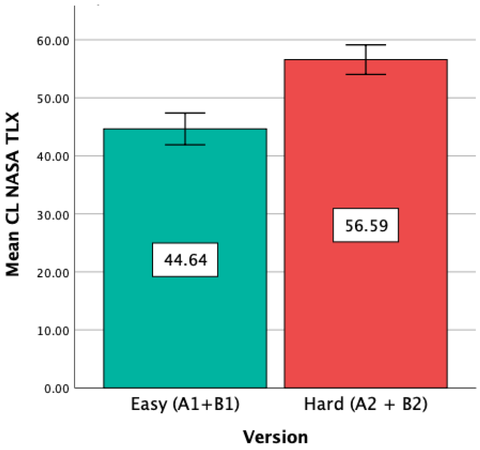


Figure 5.15: Simple Bar Mean of the NASA TLX's CL by the game versions – in this specific analysis, groups A1 and B1 were merged into the "Easy" group; and groups A2 and B2 were merged the into "Hard" group; Error Bars refer to the Standard Error of the Mean (SEM).

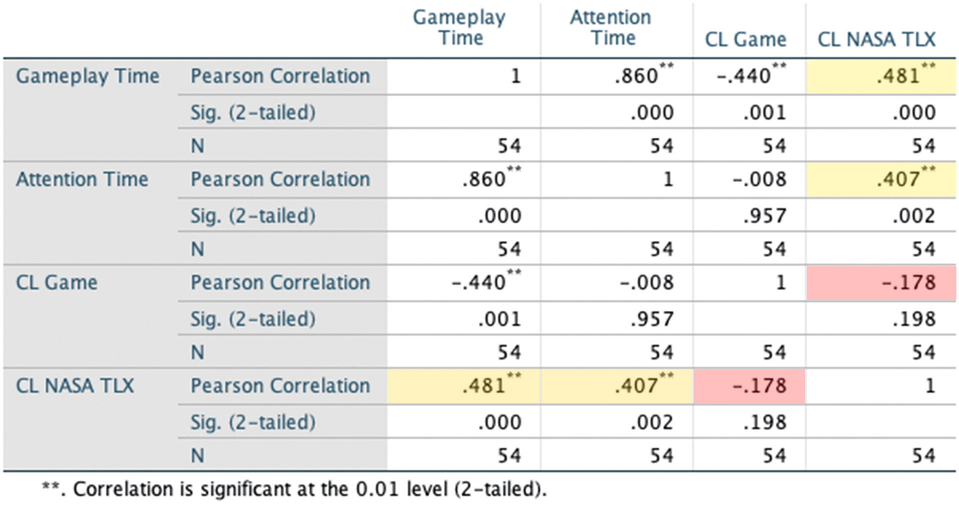
We ran a *Mann-Whitney*[6] test to compare the NASA TLX scores in both groups and found that respondents in the "Hard" group reported significantly higher CL (M = 56.58; SEM = 2.55) than participants in the "Easy" group (M = 44.64; SEM = 2.73) (U = 192.5; p = .003).

Comparing the NASA TLX scores with the CL obtained from the game, we notice that there is no

[6] "The Mann-Whitney test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed."

correlation (Figure 5.16 highlighted with red). This can be justified by the same reason why the average CL reported from the game rounded the 32% for every version (Section 5.4.3).

However, we also wanted to observe if there was a correlation between the NASA TLX scores and both the individual dimensions that, according to the TBRS memory model, affect the CL – gameplay time and attention time. Even though not perfect, as seen in Figure 5.16 (highlighted in yellow), there is a positive *Pearson* correlation between the NASA TLX scores with both the game times. This makes sense because the same pattern has been observed across the previous results: *The Hard versions (A2 and B2) resulted in significant longer game times (both total gameplay and attention) and higher NASA TLX scores; while the opposite was observed in the Easy versions (A1 and B2)*.



		Gameplay Time	Attention Time	CL Game	CL NASA TLX
Gameplay Time	Pearson Correlation	1	.860**	-.440**	.481**
	Sig. (2-tailed)		.000	.001	.000
	N	54	54	54	54
Attention Time	Pearson Correlation	.860**	1	-.008	.407**
	Sig. (2-tailed)	.000		.957	.002
	N	54	54	54	54
CL Game	Pearson Correlation	-.440**	-.008	1	-.178
	Sig. (2-tailed)	.001	.957		.198
	N	54	54	54	54
CL NASA TLX	Pearson Correlation	.481**	.407**	-.178	1
	Sig. (2-tailed)	.000	.002	.198	
	N	54	54	54	54

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5.16: Bivariate Correlation (*Pearson*) between the gameplay time, attention time, CL from the game and CL from the NASA TLX.

5.5 Summary

This chapter started by explaining the structure of our procedure, dividing it in 3 parts: The *first* consists of obtaining demographic answers of our respondents; the *second* consists of playing the actual game; and the *third* asks the participants questions related with their mental workload while playing the game (using the NASA TLX Questionnaire).

A consistent feedback received during the pilot tests was that the pairwise comparisons, at the end of the NASA TLX, were somewhat confusing. Therefore, the first thing we analysed was the correlation between the two versions of the questionnaire (NASA TLX versus its trimmed version, the RAW TLX); with the aim of deciding which would later be used to validate the game data. Since it is typically more common to use the full version of the questionnaire, and due the high positive correlation observed with its trimmed version (Figure 5.5) we opted to validate the gameplay data with the complete questionnaire, the NASA TLX.

After, we analysed the game data that, according to the TBRS memory model, directly affects the CL – both the gameplay and attention time of the four versions of our game.

While analysing the gameplay data, we noticed some unexpected results – the gameplay time of the "Additional Movement" (B) versions of the game appeared to be very similar (or even lower) than the "Normal Movement" (A) versions. We explored some different factors that may have affected these results, *but ended up concluding that our manipulation of the B versions was unsuccessful.*

On the contrary, the total attention time results were expected. Versions A1 and B1 reported significant less attention time when compared with versions A2 and B2 – *confirming that we have successfully implemented the distinction between the Easy and Hard versions of our game.*

The CL data obtained from the game was similar through all versions of the game. However, this can be justified by the positive correlation observed between the gameplay and attention times. If both the dividend and divisor have a high positive correlation, the resulting CL will always be similar regardless the times spent in the game. *This led us to conclude that the adapted TBRS CL formula, in its current form, is not sensitive enough to detect the variations across the versions.*

Onto the NASA TLX CL results, there was a noticeable variation in every version of the game. Both the Hard versions (A2 and B2) resulted in higher levels CL. Furthermore, we noticed a positive correlation between the CL from NASA TLX and the game times (both the total gameplay and attention times); meaning that, in average, the *respondents who scored higher levels of CL in the NASA TLX, also spent more time playing and had a higher number of AGEs through the game.*

6

Conclusion

Contents

6.1 Summary of Work

It is unquestionable that the video game industry is doing a proper job in keeping up with the exponential technological growth. Each passing year, thousands of games are launched with complex mechanics and challenges that, if not dealt with properly, can easily defy the limitations of the players WM. This work hypothesised that it was possible to assess the players CL based on their gameplay behaviours – and figuring out a way to accomplish it was our motivation.

The approach we took consisted of applying the attention-shifting principles of the TBRS Memory Model in the game Way Out (a game we have developed from scratch). Based on the model's principles, we formulated the idea of Attention-Grabbing Events (AGE) – which are periods of time during the gameplay in which the player's attention is most likely being grabbed. In Way Out, we considered the following events as attention-grabbers: object interactions, actively interacting with the game's UIs and display notification times. Having the total gameplay time and player's attention time, it would be possible to apply a formula similar with the one from TBRS to estimate the player's CL.

We implemented four versions of the game to manipulate two variables, each with two levels (a 2x2 factorial design): we manipulated the number of AGEs to analyse the repercussions that more or less AGEs had on the player's CL (versions A1 and A2); and we also manipulated how much players had to move around the map, aiming to see the effects that a longer gameplay time had on their CL (versions B1 and B2).

To validate our results, we opted to use the NASA TLX Questionnaire – a subjective approach that assesses the mental workload experienced during a task. The experiment was advertised across multiple social media platforms, and we ended up with a convenience sample of 54 participants. It consisted of answering a few demographic questions; followed by playing the game Way Out; and ended with the NASA TLX questionnaire.

The main variables we wanted to analyse across all game versions were: the total gameplay and attention times, the CL experienced by the players during the game (using the TBRS formula) and the resulting CL from NASA TLX (the players NASA TLX scores). While analysing the gameplay data – namely the gameplay time – we ended up with some unexpected results. The "Additional Movement" (B) versions took, in average, less time to complete than the "Normal Movement" (A) versions. Leading us to conclude that our manipulation of the B version was unsuccessful; and preventing us from answering a question we initially had: "For two puzzles with a similar intrinsic difficulty, how would the variations in gameplay time affect the player's CL?"

On the contrary, the game data indicated that our manipulation of the intrinsic difficulty of the puzzles was successful – the players that played the harder versions (A2 and B2) spent more time interacting with objects and playing the game, when compared with the players that played the easier versions (A1 and B1).

Using the TBRS CL formula to calculate the CL experienced by the players during the game, we noticed that it was nearly the same across all the game versions (around 32%). However, we also noticed that there was a high positive correlation between the gameplay and attention times; and, since the formula we used to calculate the CL results from the division of these two variables – the similar percentages of CL can be justified by this positive correlation. Nevertheless, we concluded that the TBRS CL formula, at least in its current form, is not sensitive enough to directly measure the player's CL in a gameplay scenario.

Finally, we analysed the NASA TLX scores, aiming to compare them with the game data. We noticed that the players that played the harder versions (A2 and B2) scored higher percentages of CL when compared with the ones that played the easier versions (A1 and B1). This led us to conclude that, although the TBRS formula does not appear to be sensitive enough to directly assess the player's CL, there was a positive correlation between the game times (both total gameplay and attention time) and the NASA TLX scores, meaning that – more AGEs and gameplay time resulted in higher scores of CL using the NASA TLX.

This was the first study that tried to assess the player's CL, in an automatic non-intrusive way, while playing a video game. Although we were unable to directly estimate the player's CL, we believe that this work was a step forward towards achieving that goal. Based on the TBRS attention-shifting principles, the amount of AGEs and gameplay time, when compared with the NASA TLX scores, seem to be a good indicator of CL levels; however, the TBRS CL formula, in its current form, does not appear to be reliable when directly applied in a general gameplay scenario – at least following the approach we did.

6.2 Limitations and Future Work

In order to strengthen our conclusion, a larger sample of players should be gathered – ideally with the same amount of participants for each different version and with similar gaming experience.

Directly following our work, it would be interesting to verify whether intrinsic time pressure in a similar game, using the TBRS adapted CL formula, would return more reliable results. In other words, would the direct division of the total attention time by the total restricted gameplay time, return similar CL values to the ones reported in a valid questionnaire (for instance, NASA TLX).

In addition, it would also be interesting to answer one of the questions that we initially had, but were unable to answer due to the unsuccessful manipulation of the "Additional Movement" (B) versions: How would the items disposition affect the player's CL? More specifically, how would the CL vary if the player had to memorize something crucial for the gameplay, but no AGEs happen for an extended period of time? For instance, the player retains a code sequence in WM that is written in a room, but that information is only useful after the player follows a long trail.

Even though our initial goal was to support game designers (especially during the testing phase) – by providing them with a tool-set that measured the CL percentage experience by the players, while playing a video game – this work could be expanded in a broader set of fields. For instance, when designing and implementing autonomous agents; where human-like behaviours, based on the available cognitive resources, could be improved by using the principles of the TBRS and attention-shifting in an approach similar to ours. In this scenario, game designers would also be the ones defining the AGEs, taking in consideration the environment in which the agents were situated.

Bibliography

[1] R. C. Atkinson and R. M. Shiffrin, “Human memory: A proposed system and its control processes.” *In K. W. Spence and J. T. Spence (Eds.), The Psychology of learning and motivation: Advances in research and theory*, 1968.

[2] A. Baddeley and G. Hitch, *Working memory*. Academic Press, 1974.

[3] A. Baddeley, “The episodic buffer: A new component of working memory?” *Trends in cognitive sciences*, December 2000.

[4] P. Barrouillet and V. Camos, “The time-based resource-sharing model of working memory,” *The Cognitive Neuroscience of Working Memory*, June 2007.

[5] E. Granholm, R. Asarnow, A. Sarkin, and K. Dykes, “Pupillary responses index cognitive resource limitations,” *Psychophysiology*, August 1996.

[6] R. Mayer and R. Moreno, “Nine ways to reduce cognitive load in multimedia learning,” *Educational Psychologist*, March 2003.

[7] E. Alemdag and K. Cagiltay, “A systematic review of eye tracking research on multimedia learning,” *Computers Education*, June 2018.

[8] J. J. G. Van Merrienboer, O. Jelsma, and F. Paas, “Training for reflective expertise: A four-component instructional design model for complex cognitive skills,” *Educational Technology Research and Development*, June 1992.

[9] R. Ferdig, W. Huang, and T. Johnson, *Instructional Game Design Using Cognitive Load Theory*. IGI Global, January 2009.

[10] J. Sweller, J. J. G. Van Merrienboer, and F. Paas, “Cognitive architecture and instructional design,” *Educational Psychology Review*, September 1998.

[11] R. Holden and J. Passey, “Social desirability,” *Handbook of individual differences in social behavior*, January 2009.

[12] W. Scoville and B. Milner, “Loss of recent memory after bilateral hippocampal lesions,” *Journal of neurology, neurosurgery, and psychiatry*, February 1957.

[13] F. Paas, A. Renkl, and J. Sweller, “Cognitive load theory and instructional design: Recent developments,” *Educational Psychologist*, June 2010.

[14] G. A. Miller, “The magical number seven, plus or minus two: Some limits on our capacity for processing inforation,” *The Psychological Review*, March 1956.

[15] S. Sala, A. Baddeley, C. Papagno, and H. Spinnler, “Dual-task paradigm: A means to examine the central executive,” *Annals of the New York Academy of Sciences*, December 2006.

[16] R. H. Logie, “Retiring the central executive,” *Quarterly Journal of Experimental Psychology*, January 2016.

[17] D. Kahneman, *Attention and effort*. Prentice-Hall, 1973.

[18] K. Oberauer and S. Lewandowsky, “Modeling working memory: A computational implementation of the time-based resource-sharing theory,” *Psychonomic bulletin review*, February 2011.

[19] J. Sweller, P. Ayres, and S. Kalyuga, *Cognitive Load Theory*, ser. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies. Springer New York, 2011.

[20] G. Hossain and M. Yeasin, “Understanding effects of cognitive load from pupillary responses using hilbert analytic phase,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, September 2014.

[21] F. Paas, J. Tuovinen, H. Tabbers, and P. Van Gerven, “Cognitive load measurement as a means to advance cognitive load theory,” *Educational Psychologist*, March 2003.

[22] F. Paas, “Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach,” *Journal of Educational Psychology*, December 1992.

[23] E. Galy, M. Cariou, and C. Mélan, “What is the relationship between mental workload factors and cognitive load types?” *International journal of psychophysiology : official journal of the International Organization of Psychophysiology*, October 2011.

[24] S. Hart and L. Staveland, *Human Mental Workload*. Elsevier Science, 1988.

[25] A. Cao, K. Chintamani, A. Pandya, and R. Ellis, “Nasa tlx: Software for assessing subjective mental workload,” *Behavior research methods*, March 2009.

[26] S. Hart, “Nasa-task load index (nasa-tlx); 20 years later,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, October 2006.

[27] W. F. Moroney, D. W. Biers, F. T. Eggemeier, and J. A. Mitchell, “A comparison of two scoring procedures with the nasa task load index in a simulated flight task,” in *Proceedings of the IEEE 1992 National Aerospace and Electronics Conference*, June 1992.

[28] F. Paas, P. Ayres, and M. Pachman, *Assessment of Cognitive Load in multimedia learning theory, methods and Applications*. Information Age Publishing, November 2008.

[29] P. Ayres and J. Sweller, “The split attention principle in multimedia learning,” *The Cambridge handbook of multimedia learning*, January 2014.

[30] T. Chong, W. Munassar, and W. Yahaya, “Redundancy effect in multimedia learning: A closer look.” Australasian Society for Computers in Learning in Tertiary Education, December 2010.

[31] M. Just and P. Carpenter, “Eye fixations and cognitive processes,” *Cognitive Psychology*, October 1976.

[32] R. Mayer, *Cognitive theory of multimedia learning*. The Cambridge handbook of multimedia learning, 2005.

[33] J.-L. Kruger and S. Doherty, “Measuring cognitive load in the presence of educational video: Towards a multimodal methodology,” *Australasian Journal of Educational Technology*, December 2016.

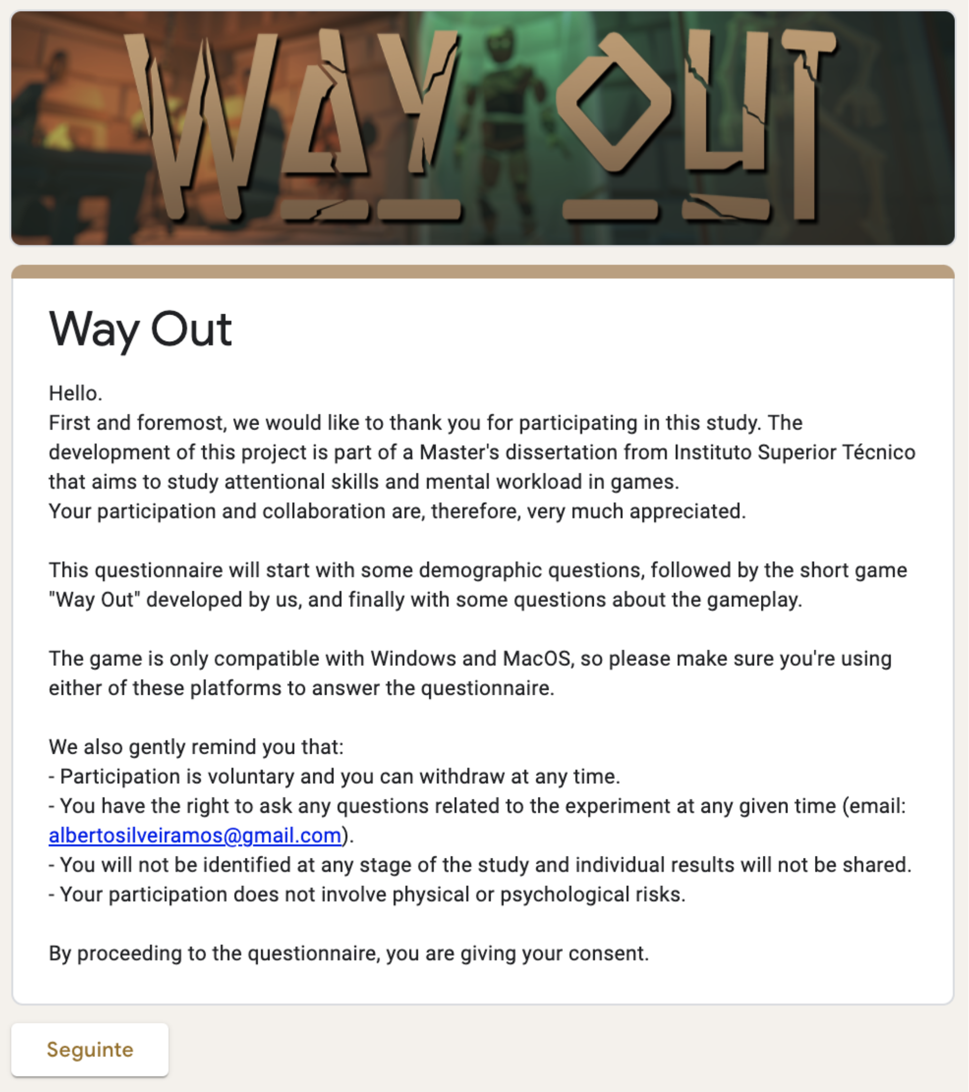
[34] S. Kalyuga and J. Plass, *Evaluating and Managing Cognitive Load in Games*. IGI Global, January 2008.

[35] J. J. G. Van Merrienboer, R. Clark, and M. Croock, “Blueprints for complex learning: The 4c/id-model,” *Educational Technology Research and Development*, June 2002.

[36] A. Diamond, “Executive functions,” *Annual Review of Psychology*, September 2012.

A

Questionnaire



Way Out

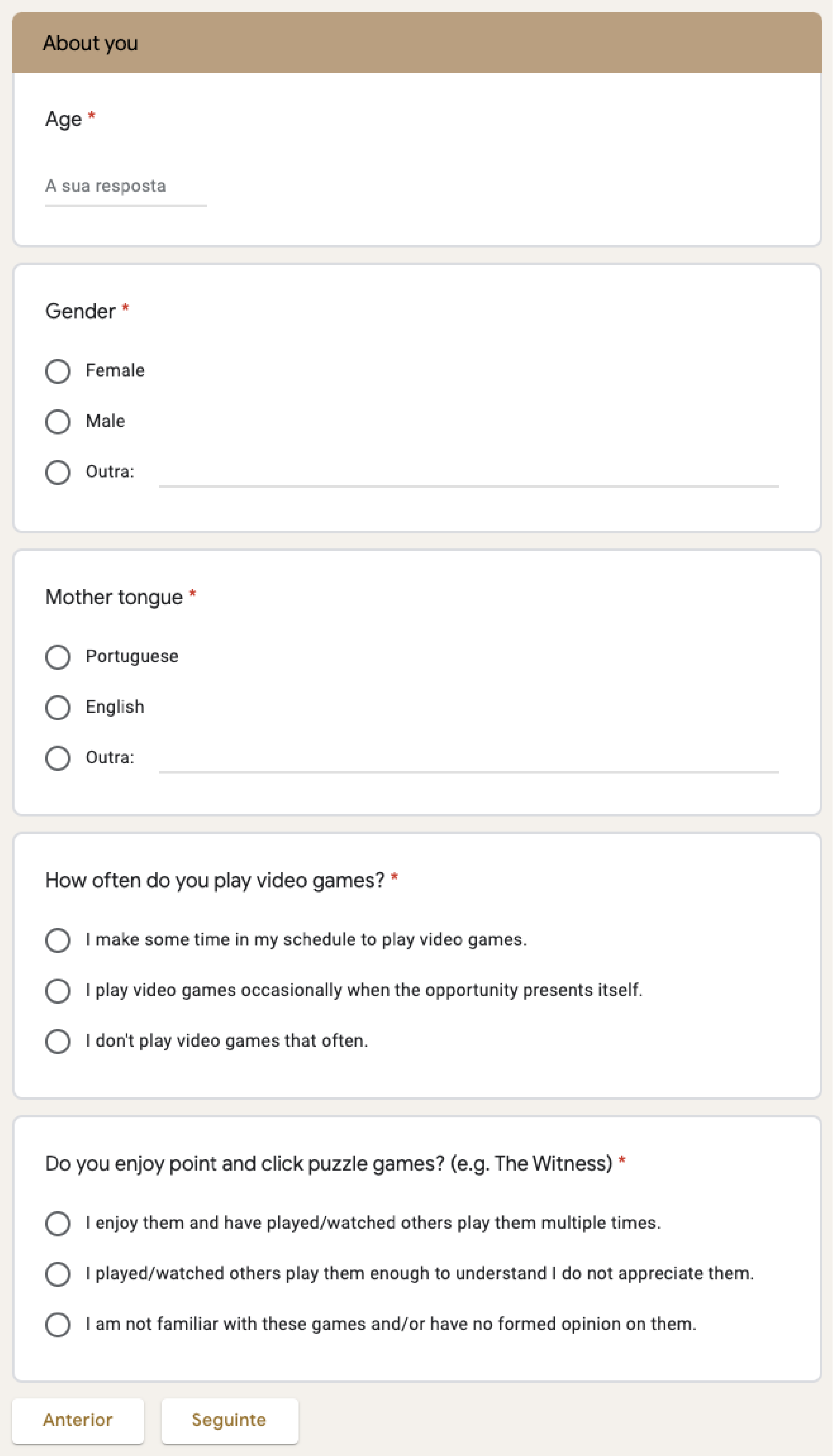
Hello.
First and foremost, we would like to thank you for participating in this study. The development of this project is part of a Master's dissertation from Instituto Superior Técnico that aims to study attentional skills and mental workload in games.
Your participation and collaboration are, therefore, very much appreciated.

This questionnaire will start with some demographic questions, followed by the short game "Way Out" developed by us, and finally with some questions about the gameplay.

The game is only compatible with Windows and MacOS, so please make sure you're using either of these platforms to answer the questionnaire.

We also gently remind you that:
- Participation is voluntary and you can withdraw at any time.
- You have the right to ask any questions related to the experiment at any given time (email: albertosilveiramos@gmail.com).
- You will not be identified at any stage of the study and individual results will not be shared.
- Your participation does not involve physical or psychological risks.

By proceeding to the questionnaire, you are giving your consent.



About you

Age *

A sua resposta

Gender *

○ Female

○ Male

○ Outra: ____________

Mother tongue *

○ Portuguese

○ English

○ Outra: ____________

How often do you play video games? *

○ I make some time in my schedule to play video games.

○ I play video games occasionally when the opportunity presents itself.

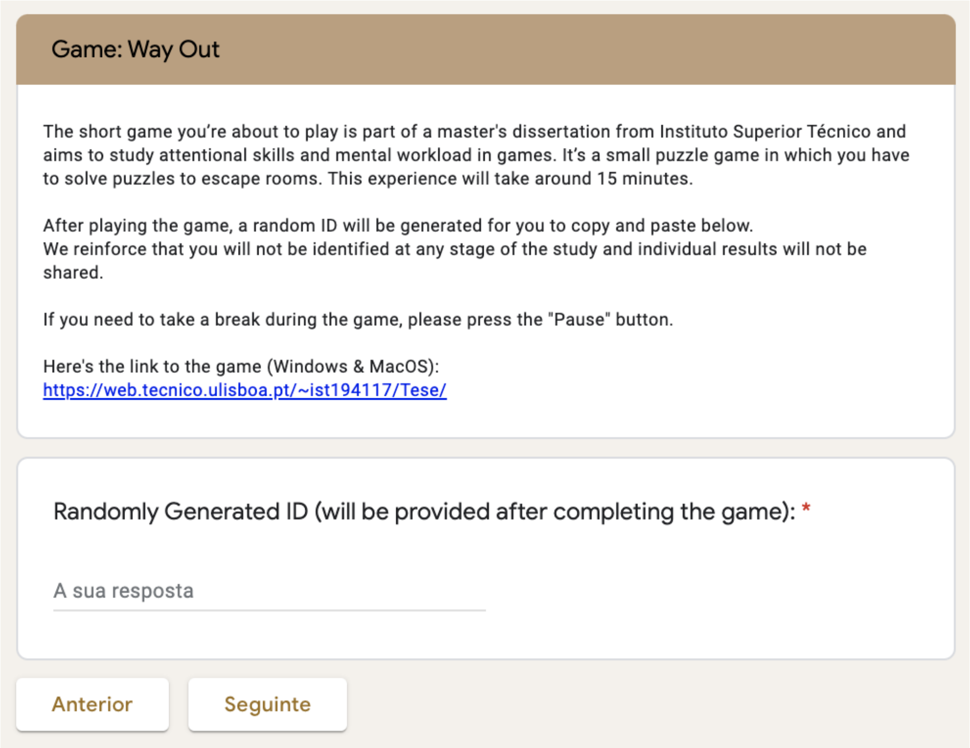
○ I don't play video games that often.

Do you enjoy point and click puzzle games? (e.g. The Witness) *

○ I enjoy them and have played/watched others play them multiple times.

○ I played/watched others play them enough to understand I do not appreciate them.

○ I am not familiar with these games and/or have no formed opinion on them.



Game: Way Out

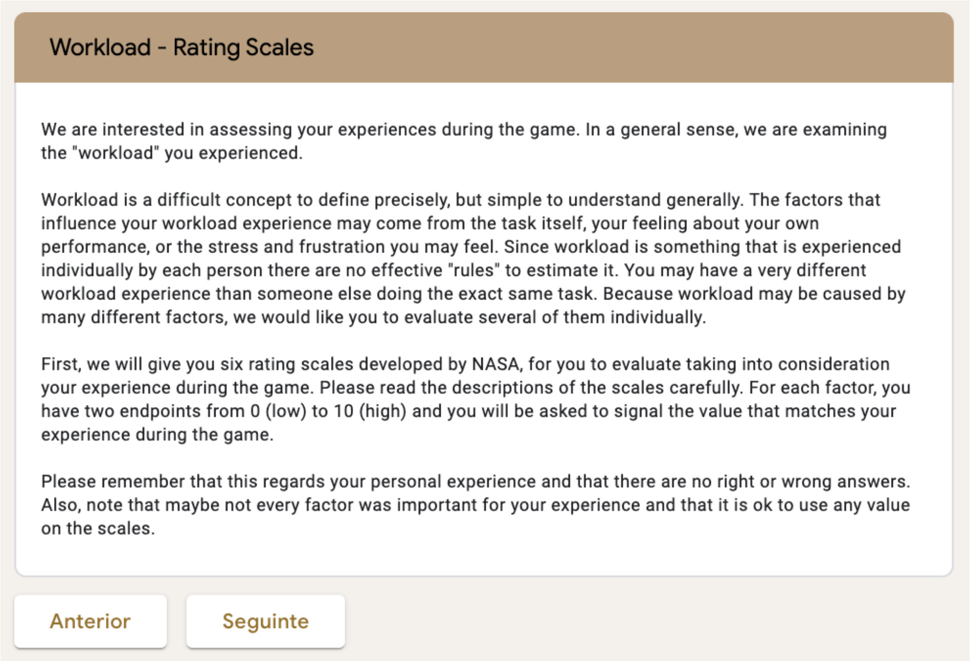
The short game you're about to play is part of a master's dissertation from Instituto Superior Técnico and aims to study attentional skills and mental workload in games. It's a small puzzle game in which you have to solve puzzles to escape rooms. This experience will take around 15 minutes.

After playing the game, a random ID will be generated for you to copy and paste below.
We reinforce that you will not be identified at any stage of the study and individual results will not be shared.

If you need to take a break during the game, please press the "Pause" button.

Here's the link to the game (Windows & MacOS):
https://web.tecnico.ulisboa.pt/~ist194117/Tese/

Randomly Generated ID (will be provided after completing the game): *



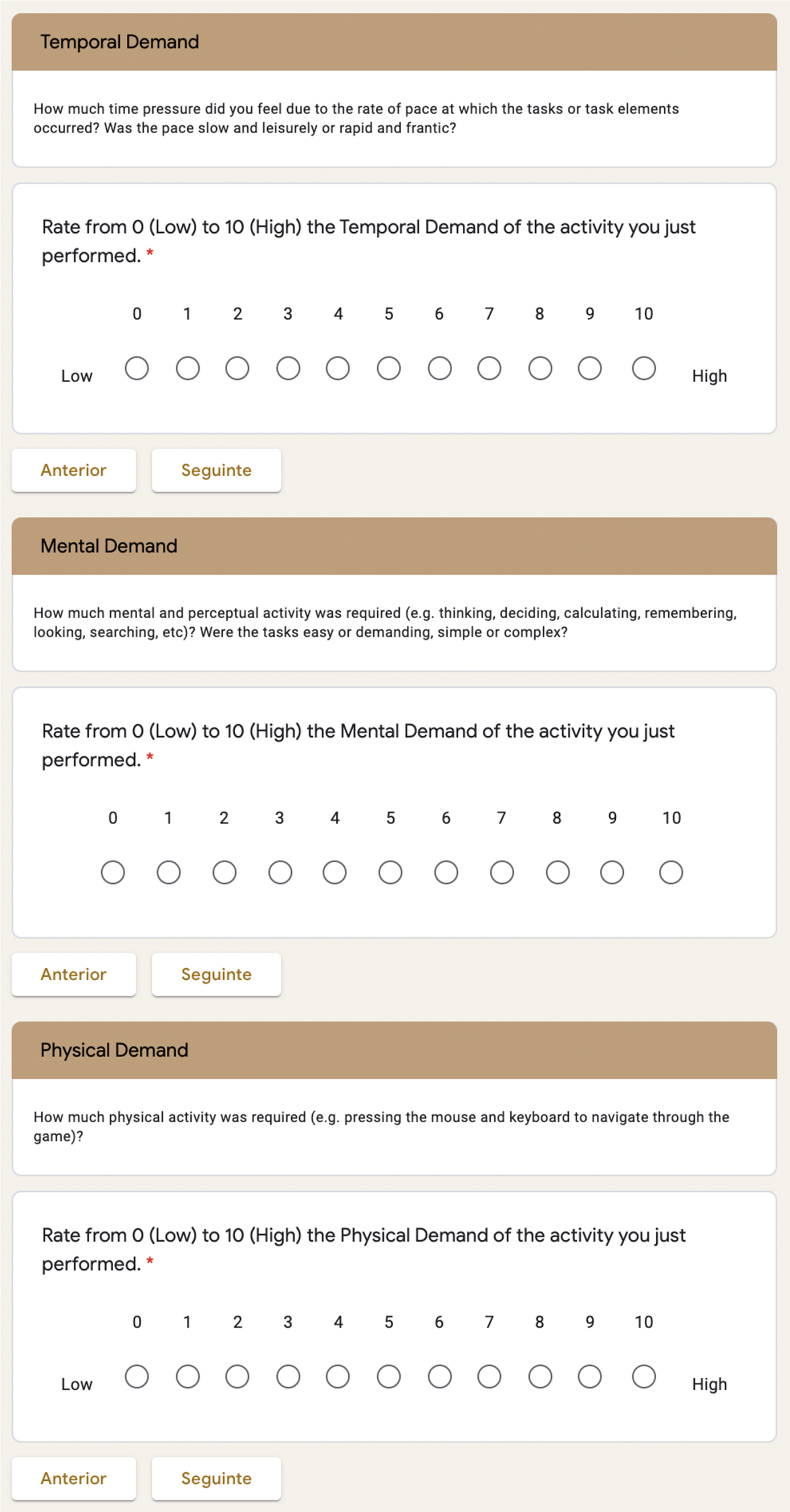
Workload - Rating Scales

We are interested in assessing your experiences during the game. In a general sense, we are examining the "workload" you experienced.

Workload is a difficult concept to define precisely, but simple to understand generally. The factors that influence your workload experience may come from the task itself, your feeling about your own performance, or the stress and frustration you may feel. Since workload is something that is experienced individually by each person there are no effective "rules" to estimate it. You may have a very different workload experience than someone else doing the exact same task. Because workload may be caused by many different factors, we would like you to evaluate several of them individually.

First, we will give you six rating scales developed by NASA, for you to evaluate taking into consideration your experience during the game. Please read the descriptions of the scales carefully. For each factor, you have two endpoints from 0 (low) to 10 (high) and you will be asked to signal the value that matches your experience during the game.

Please remember that this regards your personal experience and that there are no right or wrong answers. Also, note that maybe not every factor was important for your experience and that it is ok to use any value on the scales.



Temporal Demand

How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

Rate from 0 (Low) to 10 (High) the Temporal Demand of the activity you just performed. *

	0	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	○	High

Mental Demand

How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc)? Were the tasks easy or demanding, simple or complex?

Rate from 0 (Low) to 10 (High) the Mental Demand of the activity you just performed. *

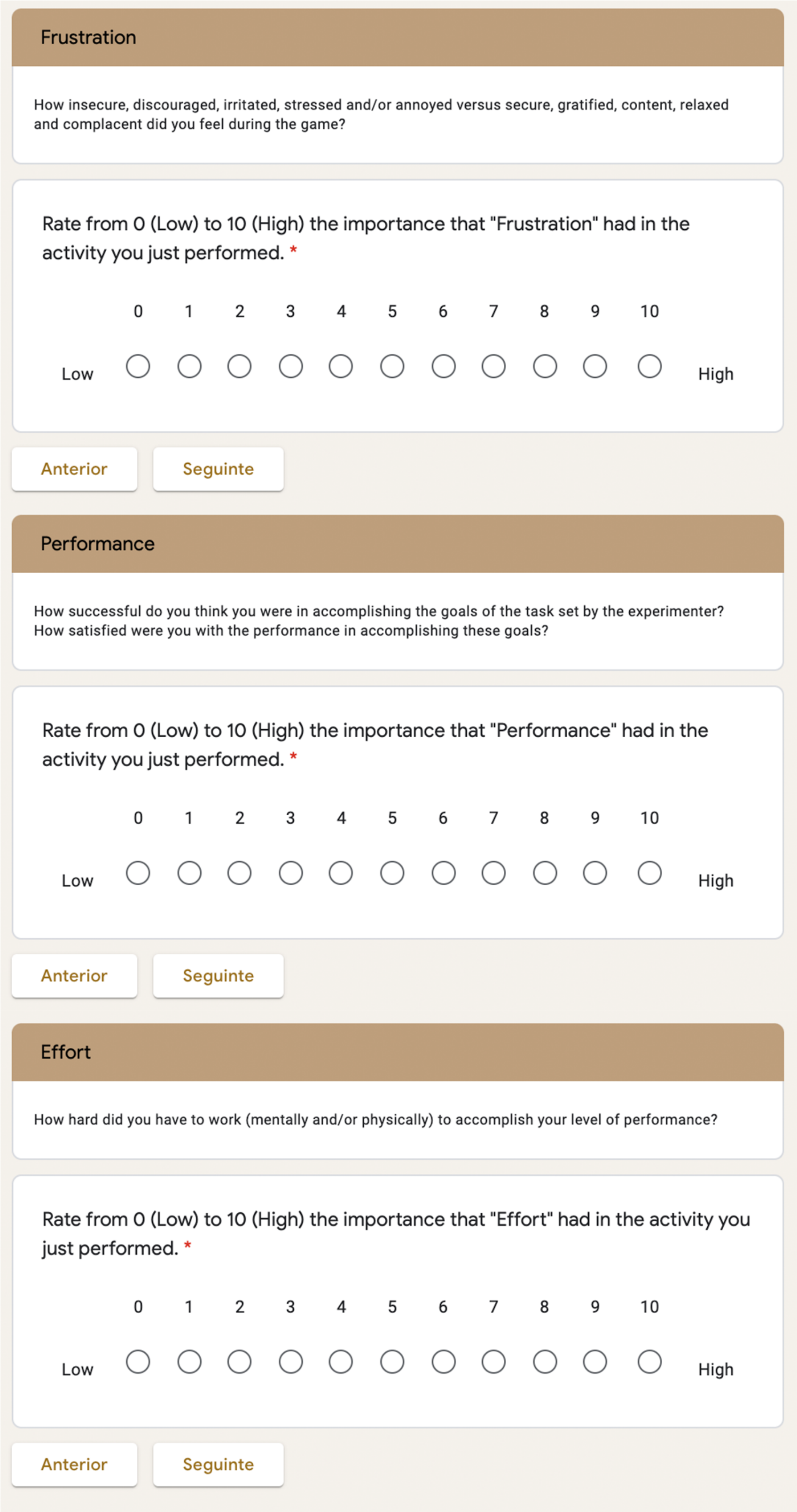
0	1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○	○

Physical Demand

How much physical activity was required (e.g. pressing the mouse and keyboard to navigate through the game)?

Rate from 0 (Low) to 10 (High) the Physical Demand of the activity you just performed. *

	0	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	○	High



Frustration

How insecure, discouraged, irritated, stressed and/or annoyed versus secure, gratified, content, relaxed and complacent did you feel during the game?

Rate from 0 (Low) to 10 (High) the importance that "Frustration" had in the activity you just performed. *

	0	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	○	High

Performance

How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with the performance in accomplishing these goals?

Rate from 0 (Low) to 10 (High) the importance that "Performance" had in the activity you just performed. *

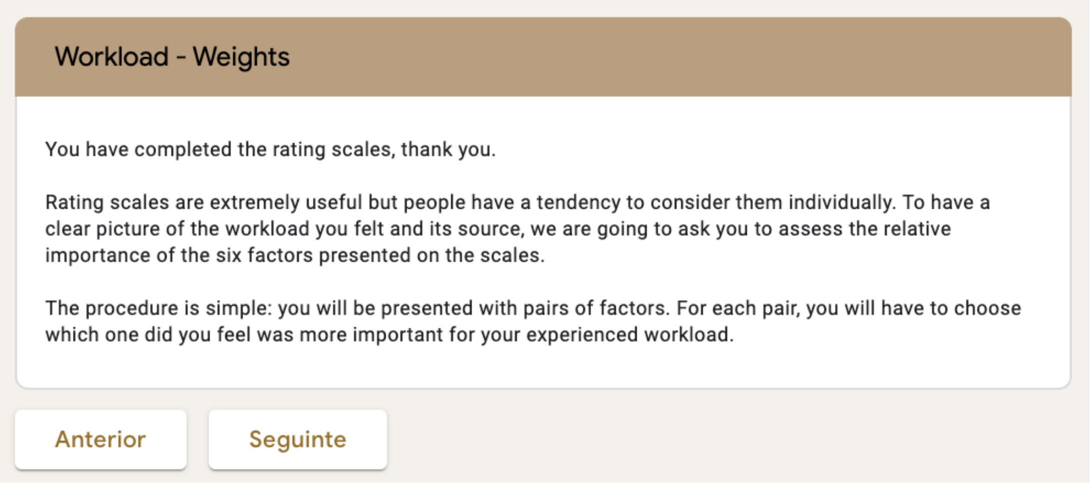
	0	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	○	High

Effort

How hard did you have to work (mentally and/or physically) to accomplish your level of performance?

Rate from 0 (Low) to 10 (High) the importance that "Effort" had in the activity you just performed. *

	0	1	2	3	4	5	6	7	8	9	10	
Low	○	○	○	○	○	○	○	○	○	○	○	High

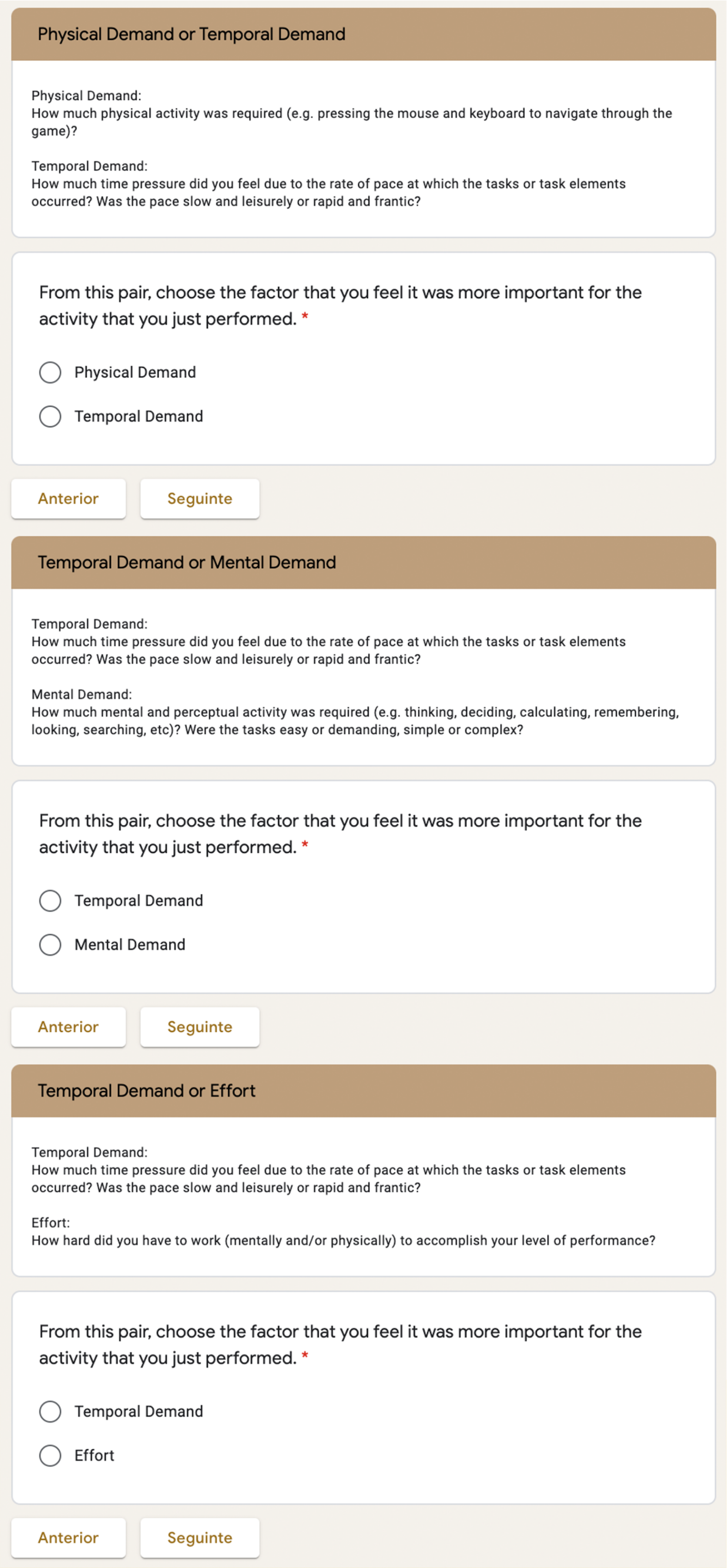


Workload - Weights

You have completed the rating scales, thank you.

Rating scales are extremely useful but people have a tendency to consider them individually. To have a clear picture of the workload you felt and its source, we are going to ask you to assess the relative importance of the six factors presented on the scales.

The procedure is simple: you will be presented with pairs of factors. For each pair, you will have to choose which one did you feel was more important for your experienced workload.



Physical Demand or Temporal Demand

Physical Demand:
How much physical activity was required (e.g. pressing the mouse and keyboard to navigate through the game)?

Temporal Demand:
How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Physical Demand

○ Temporal Demand

Temporal Demand or Mental Demand

Temporal Demand:
How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

Mental Demand:
How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc)? Were the tasks easy or demanding, simple or complex?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Temporal Demand

○ Mental Demand

Temporal Demand or Effort

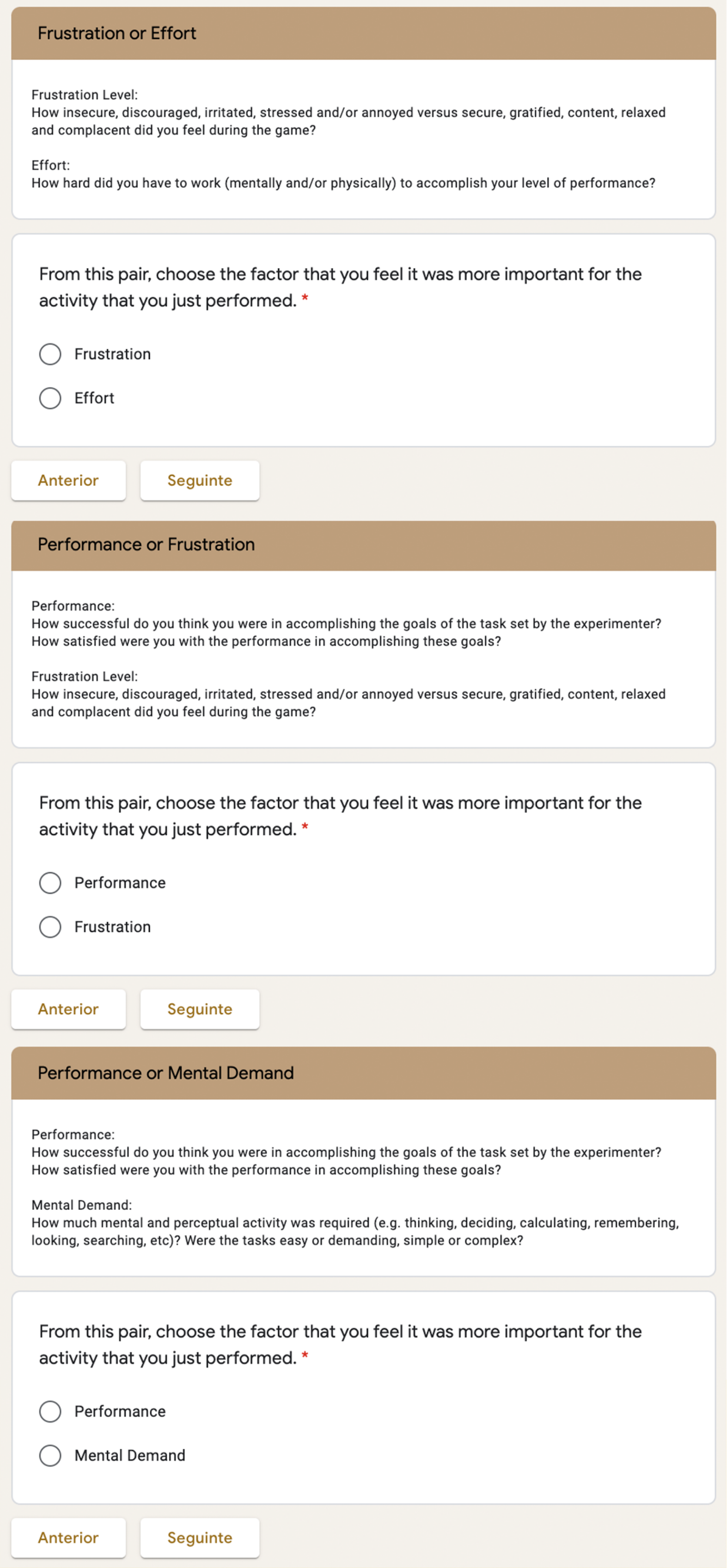
Temporal Demand:
How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

Effort:
How hard did you have to work (mentally and/or physically) to accomplish your level of performance?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Temporal Demand

○ Effort



Frustration or Effort

Frustration Level:
How insecure, discouraged, irritated, stressed and/or annoyed versus secure, gratified, content, relaxed and complacent did you feel during the game?

Effort:
How hard did you have to work (mentally and/or physically) to accomplish your level of performance?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Frustration

○ Effort

Performance or Frustration

Performance:
How successful do you think you were in accomplishing the goals of the task set by the experimenter?
How satisfied were you with the performance in accomplishing these goals?

Frustration Level:
How insecure, discouraged, irritated, stressed and/or annoyed versus secure, gratified, content, relaxed and complacent did you feel during the game?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Performance

○ Frustration

Performance or Mental Demand

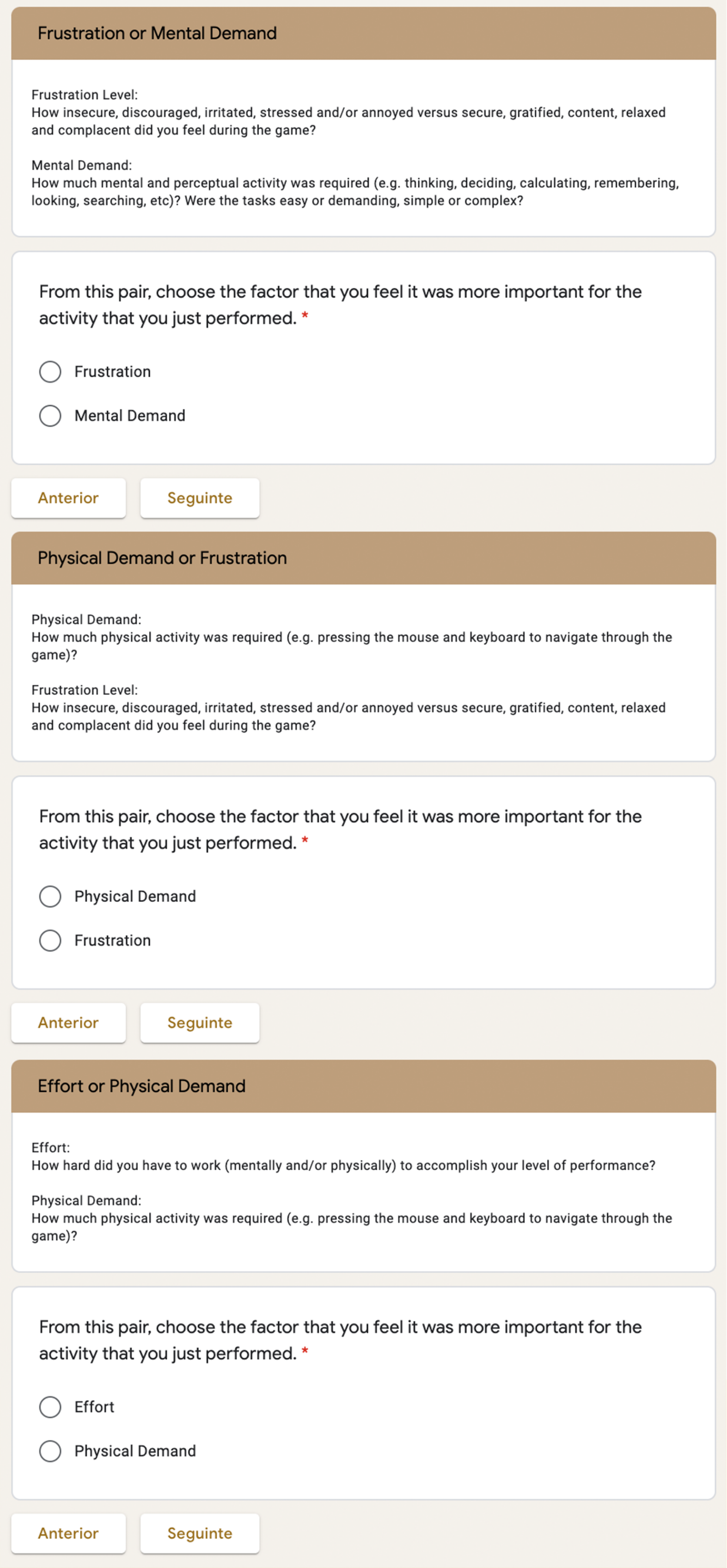
Performance:
How successful do you think you were in accomplishing the goals of the task set by the experimenter?
How satisfied were you with the performance in accomplishing these goals?

Mental Demand:
How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc)? Were the tasks easy or demanding, simple or complex?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Performance

○ Mental Demand



Frustration or Mental Demand

Frustration Level:
How insecure, discouraged, irritated, stressed and/or annoyed versus secure, gratified, content, relaxed and complacent did you feel during the game?

Mental Demand:
How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc)? Were the tasks easy or demanding, simple or complex?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Frustration

○ Mental Demand

Physical Demand or Frustration

Physical Demand:
How much physical activity was required (e.g. pressing the mouse and keyboard to navigate through the game)?

Frustration Level:
How insecure, discouraged, irritated, stressed and/or annoyed versus secure, gratified, content, relaxed and complacent did you feel during the game?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Physical Demand

○ Frustration

Effort or Physical Demand

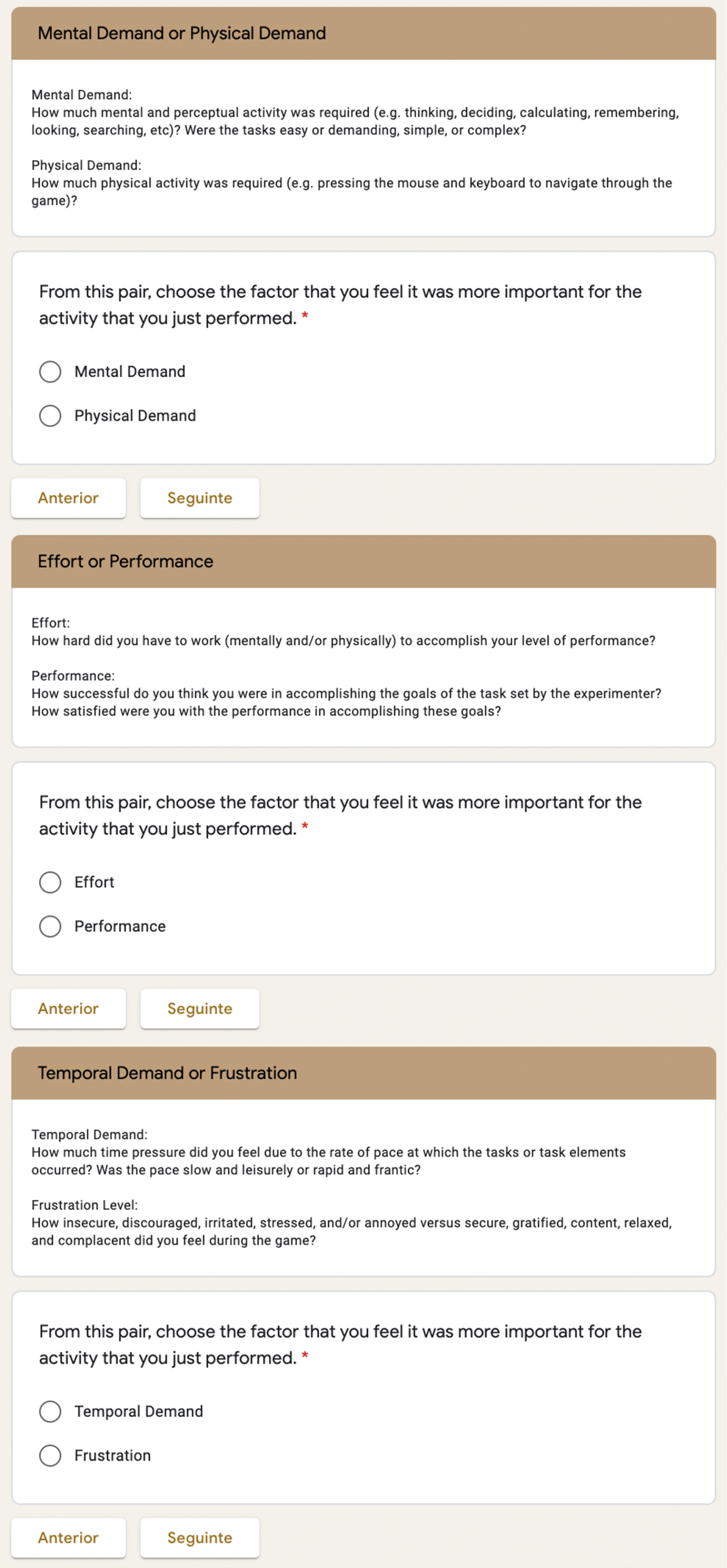
Effort:
How hard did you have to work (mentally and/or physically) to accomplish your level of performance?

Physical Demand:
How much physical activity was required (e.g. pressing the mouse and keyboard to navigate through the game)?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Effort

○ Physical Demand



Mental Demand or Physical Demand

Mental Demand:
How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc)? Were the tasks easy or demanding, simple, or complex?

Physical Demand:
How much physical activity was required (e.g. pressing the mouse and keyboard to navigate through the game)?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Mental Demand

○ Physical Demand

Effort or Performance

Effort:
How hard did you have to work (mentally and/or physically) to accomplish your level of performance?

Performance:
How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with the performance in accomplishing these goals?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Effort

○ Performance

Temporal Demand or Frustration

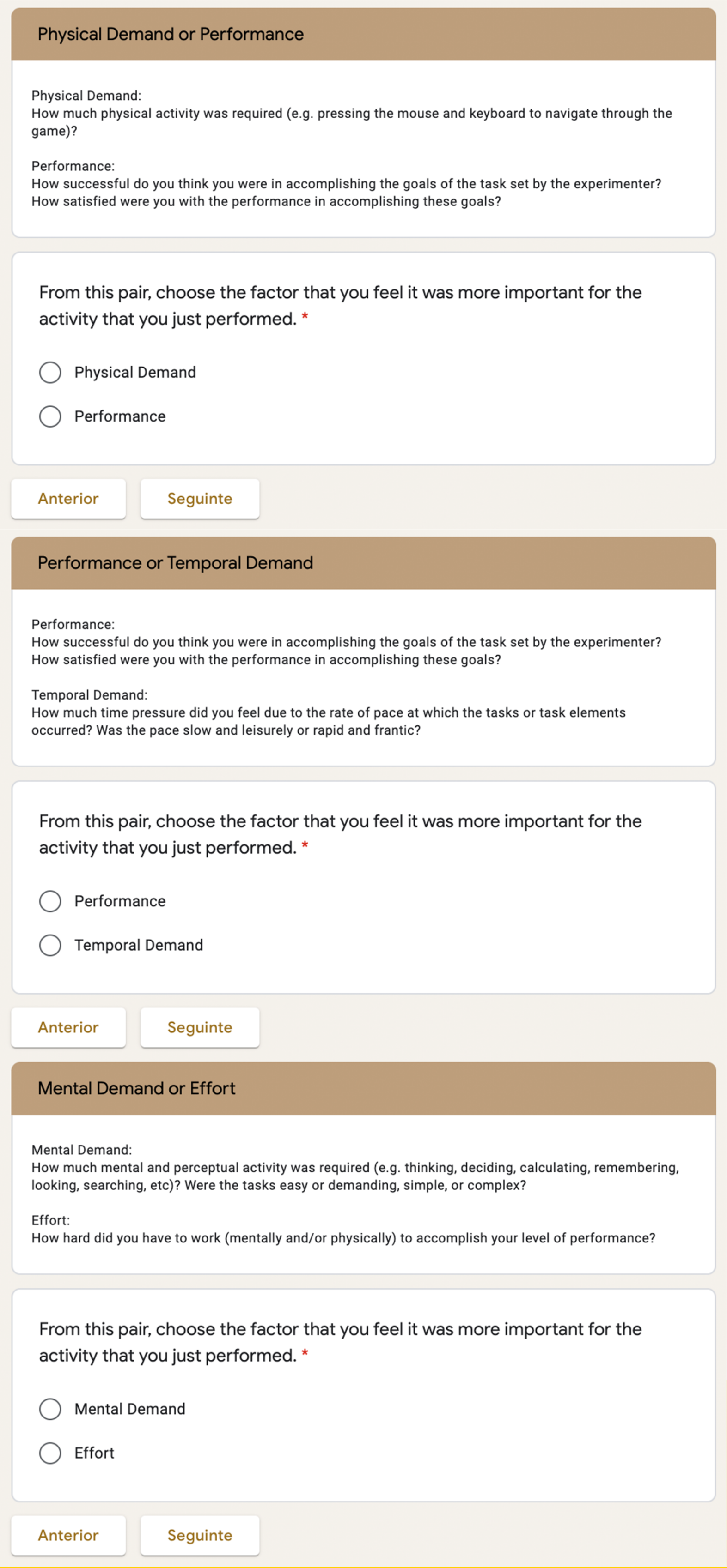
Temporal Demand:
How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

Frustration Level:
How insecure, discouraged, irritated, stressed, and/or annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the game?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Temporal Demand

○ Frustration



Physical Demand or Performance

Physical Demand:
How much physical activity was required (e.g. pressing the mouse and keyboard to navigate through the game)?

Performance:
How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with the performance in accomplishing these goals?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Physical Demand

○ Performance

Performance or Temporal Demand

Performance:
How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with the performance in accomplishing these goals?

Temporal Demand:
How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Performance

○ Temporal Demand

Mental Demand or Effort

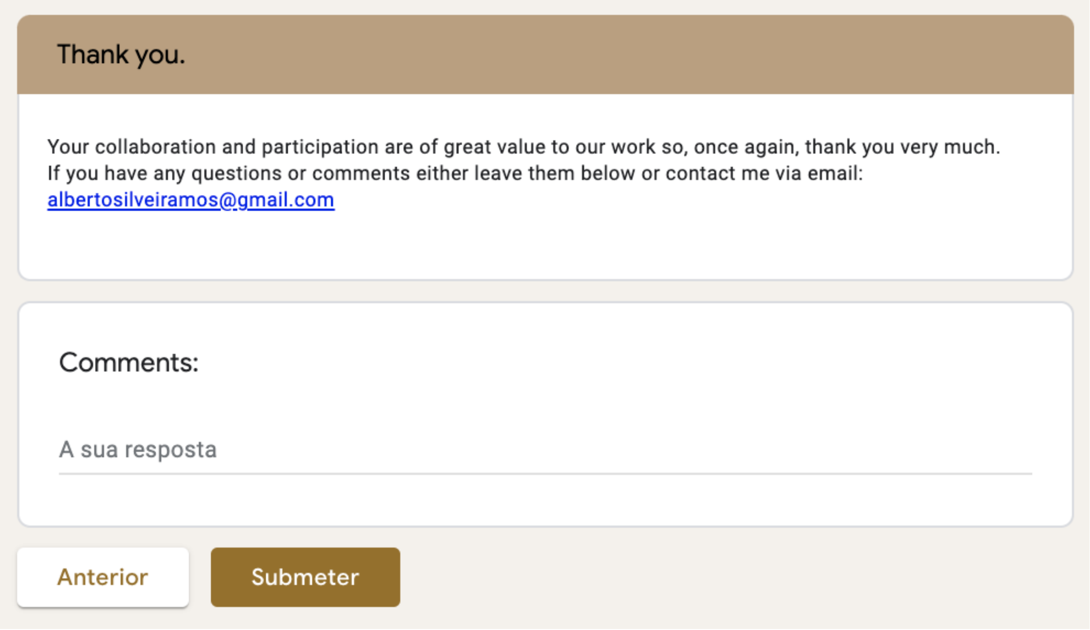
Mental Demand:
How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc)? Were the tasks easy or demanding, simple, or complex?

Effort:
How hard did you have to work (mentally and/or physically) to accomplish your level of performance?

From this pair, choose the factor that you feel it was more important for the activity that you just performed. *

○ Mental Demand

○ Effort



Thank you.

Your collaboration and participation are of great value to our work so, once again, thank you very much.
If you have any questions or comments either leave them below or contact me via email:
albertosilveiramos@gmail.com

Comments: