Assessing Players' Cognitive Load in Games

Alberto Ramos
Department of Computer Science and Engineering
Instituto Superior Técnico
Lisbon, Portugal
alberto.ramos@tecnico.ulisboa.pt

***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.

***Index Terms*—Cognitive Load (CL), Working Memory (WM), Time-Based Resource Sharing (TBRS), Video Game, Game Development, NASA TLX.**

I. Introduction

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 the 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 [1]. 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.

Assuming it is possible to estimate the duration of time in which the players' 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.

II. Background

The distinction between the nowadays called "Short-Term Memory" (STM) and "Long-Term Memory" (LTM) 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 [2], who appeared to became incapable of either store or retrieve information from the 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 [3].

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* [4].

A. Memory Models

While studying and analysing different Memory Models proposed over the years (*e.g.* Multi-Store Memory Model [3], Working Memory Model (1974 [5] and 2000 [6]), to better understand the core components that allow our Working Memory to properly function, we came across one that particularly grabbed our attention – the Time-Based Resource Sharing (TBRS) Memory Model, initially proposed by Barrouillet and Camos in 2004 [7].

This Memory Model explains how the WM functions, based on four main assumptions:
The *first*, is that both the processing and maintenance of information requires 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*.

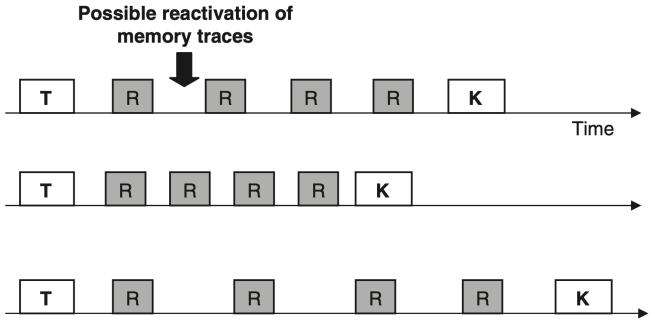

Fig. 1: Time Based Resource Sharing - "Reading digit span task" exercise.

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 (Fig. 1).

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 the 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, *Cognitive Load refers to the total amount of time during which attention was fully captured* and can be formulated as:

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

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:

$$CL = \frac{\overline{a}N}{T} \tag{2}$$

To illustrate the concept of CL, *i.e.* the balance between the competing actions that are processing and maintenance, and

grabbing the example from Fig. 1; 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).

B. 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** [8].

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 machinery. Direct objective measures include examples of dual-task methodologies, eye-tracking or task-invoked pupillary response and brain-activity measures.

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 subject to rate the perceived mental effort or task difficulty in a numerical scale and there are several different types of reports focusing on different problems. One of the great advantages of using self-reports is its simplicity, since it solely requires the appropriate set of questions for the activity that's being implemented on. Additionally, being subjective means that there is no need of using external equipment collecting behavioral data during an activity, making this a non-intrusive approach.

There are two most commonly used techniques for subjectively assessing mental workload [9], the **NASA TLX** and the **SWAT**. They both divide the workload in multiple subscales and are proven to provide quite similar results [9]. 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.

C. NASA TLX

The NASA TLX (NASA Task Load Index), developed in 1981 by Sandra G. Hart of the NASA Ames Research Center [10] 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 [11]. These subscales of the workload are based on the assumption that some combination of these dimensions are likely to represent the "workload" experienced by most people performing most tasks [12]. 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) .

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 task (the weights themselves also provide diagnostic information as the nature of the workload imposed by the task).

To do so, there are 15 possible pairwise comparisons of the six subscales of workload. Each pair (for instance, Temporal Demand vs 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 task. Thus, the respondents are asked to rate each subscale individually from 0 to 10 or 0 to 100 (least to most taxing).

The adjusted ratings for each of the six subscales of the workload is computed by multiplying their respective weight with their raw rating (3). For example, if the weight and rating of Temporal Demand was 4 and 50 respectively, its Adjusted Rating would be 4 x 50 = 200.

$$AdjustedRating = Weight * RawRating \tag{3}$$

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) (4).

$$Workload_{NASATLX} = \frac{\sum AdjustedRatings}{15} \tag{4}$$

III. Implementation

To test our hypothesis – whether or not is possible to estimate the player's Cognitive Load based on the attention-shifting principles of the TBRS – 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.

The chosen name for the game was – Way Out. The player plays 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.

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 be explored in-depth in the following subsections.

A. 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 (Formula 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 (AGEs).

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:

- **Object Interactions:** The time spent interacting with interactive objects (*e.g.* Fig. 2).

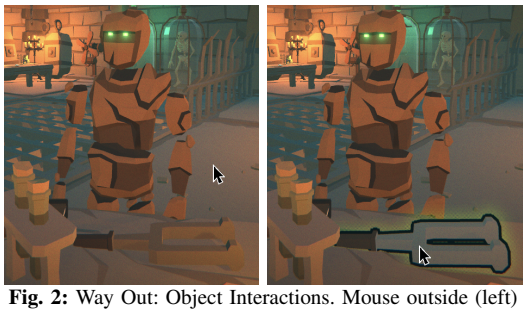

Fig. 2: Way Out: Object Interactions. Mouse outside (left) / inside (right) an interactive object.

- **Interface Interactions:** The time spent with the Inventory, Instructions, Notebook and Sphere placeholder interface opened (*e.g.* Fig. 3).
- **Notifications:** Time in which notifications were shown on screen (*e.g.* Fig. 4).

However, the overlap of events would interfere with the formula, since it would mean that the player's attention would

Fig. 3: Way Out: Interface Example (Inventory).

Fig. 4: Way Out: Notification Example.

be shifted towards processing multiple events at once. An example of this happening can also be observed in Fig. 4, where the player is interacting with an object whilst a notification is simultaneously 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 (Fig. 5).

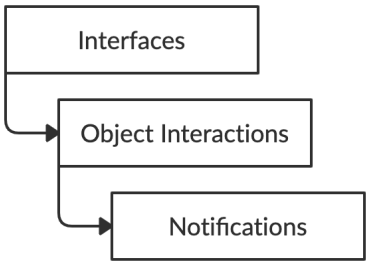

Fig. 5: Attention-Grabbing Events (AGEs) hierarchy.

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.

B. 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 (III-A), we consider object interactions, notification and interface 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 [9], we started by creating two versions of the 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 [7]. 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 (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 I: Game Versions.

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

Thus, as seen in Table I, 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 repercussions that the additional movement induced on the player has on the CL results. More specifically, the effects that additional 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 I. This will theoretically increase the overall gameplay time and, consequently, using the TBRS CL formula, decrease the CL.

C. Game Puzzles

The puzzles implemented[1] were designed to verify the effects that the variations in AGEs and movement had on the players' CL. For that purpose, we have developed two main puzzles that slightly vary between the four versions of our 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 difference in the players movement through the versions (A1 vs B1 and A2 vs B2), *i.e.* more or less gameplay time, we changed the items disposition between the versions of the puzzles.

For instance, the first puzzle of our game – The Lever Puzzle – 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. 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 that the movement of each lever has on the other 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. This leads to a theoretical higher number of interactions – and AGEs – since the solution is not as straight forward as the opposite versions.

Fig. 6: Way Out: Lever Puzzle – Versions A (left) and B (right).

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 Fig. 6.

In both the A versions (Fig. 6 left), 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

[1]A full walk-through of all the game's puzzles and versions is available at: https://www.youtube.com/watch?v=95j85Add1Rg&ab_channel=AlbertoRamos

versions (Fig. 6 right), 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.

D. Data Gathering and Cognitive Load calculation

To follow the principles of the TBRS memory model and in order to use its formula [7], we need to collect relevant gameplay data that estimates the players' attention time during the game. Therefore and, as mentioned on a previous section (Section III-A), beyond the *Total Gameplay Time*, we will mainly collect data related with the duration of the multiple AGEs that occur throughout the game (listed in Section III-A).

The total sum of AGEs will return the **Total Attention Time** (5) 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 6.

$$TotalAttentionTime = \sum_{i=1}^{N} AGE \tag{5}$$

$$CL = \frac{TotalAttentionTime}{TotalGameplayTime} \tag{6}$$

Additionally, 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 will be stored on a online Google Sheets document for further analysis.

IV. Procedure and Results

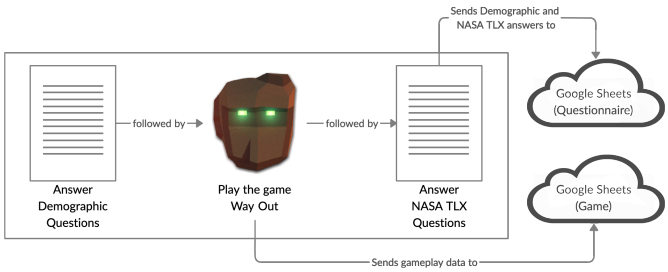

Fig. 7: Procedure to acquire data.

In short, the structure of the followed procedure is summarized in Fig. 7.

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 section (Section III). After playing and finalising the game, a random code name is generated and provided to the player, so it can be pasted 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 analysed in depth by Galy, Cariou and Mélan (2011) [9].

Therefore, after playing the game, the participants were asked to answer a few questions related with their overall workload during the game.

For that purpose, we will use the **NASA TLX** questionnaire [11] which was explained in detail in a previous section (Section II).

A. 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". Discarding the pairwise comparisons is another way of using the questionnaire and often called – RAW TLX.

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.

B. 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)** from IBM; where all the NASA TLX calculations were made and the charts, graphs and tables presented in this section were generated.

From the 54 participants, 45 (83.33%) identified themselves as males whilst 9 (16.67%) identified as females. 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).

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".

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".

C. The questionnaire of choice: NASA TLX

To clarify a question brought up during the pilot phase (Section IV-A), we started our analysis by comparing the questionnaire results with and without the pairwise comparisons, *i.e.* by comparing the NASA TLX with the RAW TLX – 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* correlation of 0.945 (as seen in Fig. 8).

		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).

Fig. 8: 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.

D. Hyphothesis

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?

To clarify our hypothesis, we started by analysing **gameplay time** (Fig. 9) 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* test that showed that there is at least one pair of significantly different groups ($H(3) = 18.74$; $p \leq .001$). The pairwise comparisons with a *Bonferroni* correction showed that the harder versions (A2 and B2) took significantly longer to complete than the easier versions ($p \leq .05$) – comparing A1 with A2 and B1 with B2.

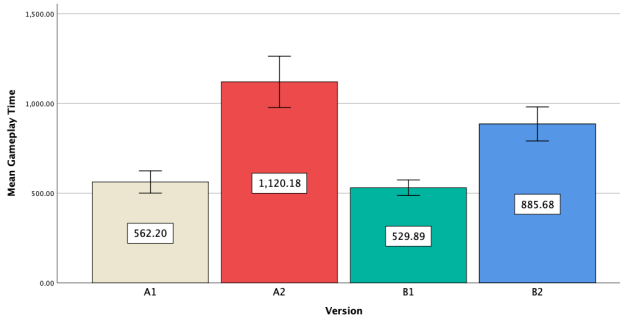

Fig. 9: Simple Bar Mean of Gameplay Time by the game versions; Error Bars refer to the Standard Error of the Mean (SEM).

Due to the game versions implementation, 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 Fig. 10, 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

Fig. 10: 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; $H(2) = 10.320$, $p = .006$. The pairwise comparisons with a *Bonferroni* correction 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 (Fig. 11), as outliers.

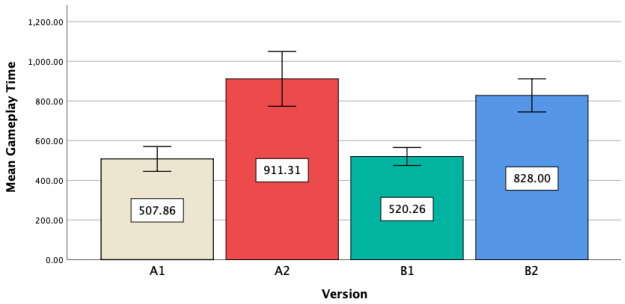

Fig. 11: Simple Bar Mean of Gameplay Time by the game versions – without the group "Not Often"; Error Bars refer to the Standard Error of the Mean (SEM).

As seen in Fig. 11, 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?".

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 (Fig. 12).

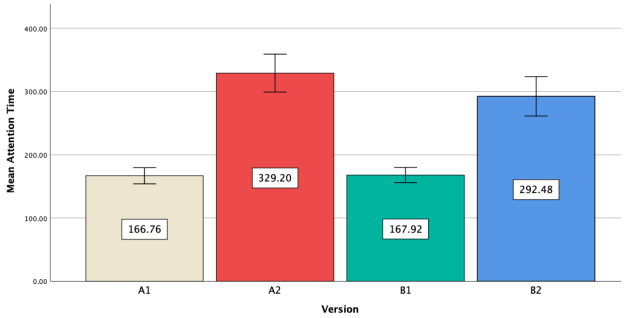

Fig. 12: Simple Bar Mean of Attention Time by the game versions; Error Bars refer to the Standard Error of the Mean (SEM).

Onto the actual **CL values** reported from the game (Fig. 13), 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.*

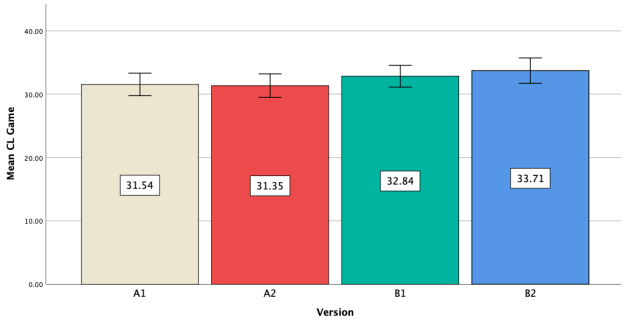

Fig. 13: Simple Bar Mean of the players CL percentages by the game versions (using the TBRS CL formula); Error Bars refer to the Standard Error of the Mean (SEM).

Observing both the **gameplay time** (Fig. 9) and **attention time** (Fig. 12) 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 (Fig. 14) 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).

Fig. 14: 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 Fig. 13. 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.

Onto the **NASA TLX scores**, there is a noticeable CL variation across the game versions (Fig. 15), 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, the "Additional Movement" (B) versions were not successfully manipulated – preventing us from concluding anything concrete related to this topic.

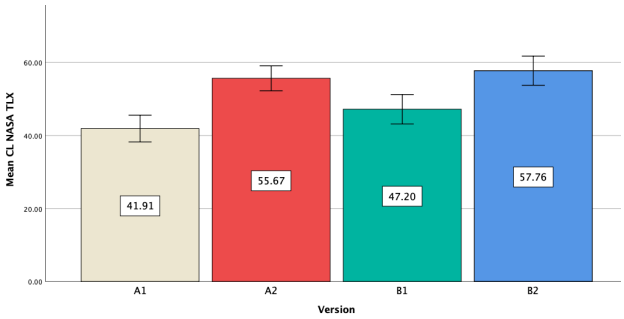

Fig. 15: Simple Bar Mean of the NASA TLX's CL by the game versions; Error Bars refer to the Standard Error of the Mean (SEM).

Comparing the NASA TLX scores with the CL obtained from the game, we notice that there is no correlation (Fig. 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 (Fig. 13).

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 Fig. 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).

Fig. 16: Bivariate Correlation (*Pearson*) between the gameplay time, attention time, CL from the game and CL from the NASA TLX.

V. 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. Even though we were unable to directly estimate the player's CL, we believe that our 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.

VI. 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.

References

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

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

[3] 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.

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

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

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

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

[8] 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.

[9] 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.

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

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

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