Tutorial Adaptation based on Working Memory

Miguel Keim
Department of Computer Science and Engineering
Instituto Superior Técnico
Lisbon, Portugal
miguel.nsm.keim@gmail.com

***Abstract*—With the popularity of video games growing exponentially over the years, the complexity and diversity of the genres and their games continues to expand. Certain problems begin to arise, however- with complex mechanics comes a need for good tutorials and proper ways to convey the information the player needs in the most optimal and concise way possible. We believe the key to solving this problem lies in its connection to our memory, the most integral part in the gathering of new knowledge.**

To achieve this, we import concepts from cognitive psychology and test two approaches on participants with different Working Memory capacities. We will test two opposite scenarios, where we apply different types of learning experiences in a Tactical Role-playing Game: (1) we drew inspiration from the Generation Effect to design a tutorial based on exploration and limited information, where the player must seek answers and gather knowledge on their own and (2) a didactic tutorial, where the game teaches in detail the mechanics of the game to the player, and how to tackle the scenario in front of them.

By measuring levels of Working Memory through a test, and then dividing our participants into the different tutorial environments, we aim to evaluate their retention of information in their respective scenario, as well as their subsequent performance in the following levels of the game. This way, we can draw potential conclusions on which tutorial approach is best adapted to take different levels of learning into consideration, or if certain methodologies work better for a specific type of player. While we were unable to gather enough of a sample to make a proper conclusion, we laid the groundwork for future implementations of the idea.

***Index Terms*—Working memory; Long-term memory; Generation effect; Game development; NASA TLX.**

I. Introduction

As the years go by, technology progresses, and with it, so do video games. Gone are the days when simple games like "Pon"[1] and "Tetris"[2] ruled the market, and with each passing generation of both hardware and software, more complex options are available for game designers. Said evolution in technology brings forth more advanced experiences with more in-depth mechanics- and with these ever-growing changes, players must adapt and learn to enjoy themselves properly.

How often do we come across games with complicated menus, with more options than we can account for? This is especially true in the case of slower strategy games, where the player is encouraged to stop, think and check as many details as possible. However, it also applies to more active games as well, such as the Souls Series games like "Dark Souls"[3] and "Sekiro"[4], where the player must learn to master a variety of controls in fast paced action, under highly stressful situations. Such extensive arsenals can be overwhelming, and the responsibility of properly conveying such information lies on the shoulders of tutorials, that many times go unappreciated within the community.

Tutorials are a crucial part of most games, responsible for letting players dip their toes in the complex mechanics of the ever-evolving games of the present market. Game developers often disregard the importance of a good tutorial, instead overwhelming their players with large quantities of information, while not letting them absorb previously given pieces of the puzzle that is their game. Tutorials should be concise and direct to the point, and know how to space new mechanics to allow gamers to practice their newly acquired knowledge, until it is properly retained within their memory.

But not everyone learns at the same pace- that much has become clear with more recent findings in the field of psychology and memory. We each have our own capacity for retaining new information and learning speeds, so how exactly are we supposed to take that into account when planning how to teach our target audience? For us to be able to adapt tutorials, regardless of the individual learning differences we have as people, we must understand how the Human Memory works, a system we all share regardless of such variances. The key lies in our **Working Memory (WM)** [1]- the cognitive system responsible for retaining new information temporarily- which holds the answers to the aforementioned disparity in learning capabilities, due to its varying capacity. Said capacity, besides changing from individual to individual, is also limited, and will fade if not properly trained, like any muscle in our body. In order for the information it stores to be converted into **Long-Term Memory (LTM)**, in other words, for it to become permanent knowledge, it must be repeated and exercised, because the more a certain information is accessed, the stronger the neural network related to it gets.

It stands to reason that tutorials are our starting point as players in video games, and not just because they are the literal beginning of the experience, but because they are the introduction to the game's core mechanics. As such, they are

Supported by Instituto Superior Técnico

[1]Pong, 1972, Atari, Arcade

[2]Tetris, 1988, Mirrorsoft, Electronika 60

[3]Dark Souls, 2011, From Software, PlayStation 3

[4]Sekiro, 2019, From Software, PlayStation 4/Windows/Xbox One

responsible for optimizing the absorption of information to the LTM storage, allowing players to retain new knowledge at an appropriate pace, before jumping into situations that will test what players have learned. However, our capacity for knowledge retention is directly linked to our WM capacity, since it is what allows us to process new information and temporarily contain the data presented for it to be trained. That is what we aim to study through this research- how to adapt tutorials to take into consideration different WM capacities, in order to lessen the impact it holds on a player's experience.

The first step we must establish is how to estimate the individual differences in WM in the first place. Experiments regarding measures of WM have already been developed in the field of psychology [5], and thus can be applied here as well. Suppose we can use them to quantify our WM capacity at the beginning of a video game experience. In that case, it is possible to utilize this information to filter our participants at both ends of the spectrum to different types of tutorials. By dividing both those with high WM capacity and the ones with low WM capacity equally through the scenarios, we can verify the effects WM has in retaining information, as well as test different learning strategies in digital games, for us to establish which provides better retention of the information presented in either type of players.

In order to choose our tutorial environments for this test, we must explore different approaches to the same goal: how do we maximize the retention of information in our players at the start of a digital game experience? Learning is a broad and very explored subject in the area of Education, with studies already reaching conclusions to the question posted above. In the field of education, research has found that self-generated knowledge provides better long term results [9], albeit the exact reason behind such an effect is still unclear to this day. One such approach that focuses on self-generated learning is called the **Generation Effect (GE)**.

This method tells us that by providing the tester with incomplete information, as opposed to giving them all of the details, we can enhance their retention of the knowledge provided, since they will be forced to complete said information and exercise the created memory. This effect has been proven effective, albeit slower in the beginning when compared to more didactic approaches, especially in complex scenarios. Such theories have, to our knowledge, never been tested in the field of digital games, and we believe it holds unexplored potential to further enhance learning experiences in video games, especially since research has shown its connection to the use of mental resources [10]. We wish to apply it in a video game of our own creation, and test its relationship with WM when compared to a didactic tutorial, where all the information is provided to the player, akin to a lecture in a classroom.

After the tutorials, participants will be tasked to play additional levels of the game we will provide, where no more help will be given and they will have the opportunity to utilize the knowledge previously obtained in the learning stage. This is when we will be able to see the performance of the different levels of WM, when applied to the different learning approaches, in a real scenario. To finalize the process, we will also measure our participants' **Cognitive Load (CL)** through the NASA TLX [15] survey, since CL refers to our usage of WM resources to perform tasks, and will help us further understand the impact both tutorial approaches have in individuals with different WM capacities [18].

We hypothesize that, by utilizing the information gathered through this experiment, we will be able to understand how we could adapt tutorials to fit our players while taking into account their WM capacity. We believe the WM capacity should influence how quickly our players will familiarize themselves with the mechanics presented to them, while the GE affects how well one retains the new knowledge they have acquired [9], even if initially slower in results when compared to a more lecture-like tutorial, depending on the mental effort required [10]. We also expect players with high WM capacity to better adapt themselves to the GE tutorial, considering they should show more capacity for retention of new information [4], while the low WM ones should instead show better results in an environment best suited for their lower capacity for retention, like the didactic approach, which provides more guidance. Regardless, we aim to verify how both types of players behave in both types of environments in order to draw more concrete conclusions.

II. Background

A. Long-Term Memory

LTM was first defined by Atkinson-Shiffrin [1] in his multi-store memory model, back in 1968. The LTM was responsible for the retention of information and skills for long periods of time- so long, in fact, that said memories are believed to potentially last a lifetime. It is said that our LTM holds unlimited capacity, and the main constraint people feel with the passage of time links instead to its accessibility. In order for information in the WM to become permanent, it must be rehearsed multiple times, each time strengthening its connection in the LTM, and likewise the longer it stays in short-term storage, the stronger its roots within our memory becomes. The transferring process between the two memory storages is called Synaptic Consolidation [2].

After training new information in our WM, it is integrated in our LTM within structures called Schemas- either stored in existing ones if associated with them, or by creating new ones. These Schemas are created by our brain in order to ease the burden of our WM: through the process of association, we can more easily understand new knowledge if we already have a Schema related to it. Through the same logic, novel information applies more of a burden in our WM, since we lack any previous information related to it. And after all this, it is still necessary to train our brain in order to prevent the forgetting process- even if our LTM is unlimited, it does not mean all the information will be kept without any maintenance. This is usually done through rehearsal, where

we recall certain knowledge and practice it to clarify it within our memory.

B. Working Memory

Prior to the second half of the 20th century, there was no concept of separation between two different types of memory storage. It was only in 1957, when William Beecher Scoville and Brenda Milner published an article by the name of "Loss of recent Memory after Bilateral Hippocampal Lesions" [3], that it was finally considered. The duo found out through examination of patients with hippocampal lesions that, even if they were unable to add new information to their long-term memory, it was still possible for them to process immediate input on the short-term. This discovery was then further explored by Richard Atkinson and Richard Shiffrin [1], where they deepened the differences between the **Short-Term Memory (STM)** and LTM, which eventually created the first **Memory Model (MM)**- a representation of the inner workings of memory within our brain- known as the Multi-Store Memory Model. Other MMs have since been theorized, which to this date are still discussed among researchers. We will be touching upon the most pertinent MM to our research shortly.

It was Richard Atkinson and Richard Shiffrin's studies that gave birth to the concept of WM- also known as STM, albeit some scientists debate that the two should be separate. Our work will, however, focus on the WM as its own separate concept. This type of memory storage is a cognitive system that allows us to hold information temporarily. It is an integral part of our brain, responsible for handling a lot of our decision making. However, it is limited, as opposed to the aforementioned Long-Term Memory, which holds theoretically unlimited permanent knowledge capable of lasting for decades. The bigger one's WM capacity, the greater number of pieces of information they can hold and process at a time, and it is believed most adults can store between 5 and 9 items at a time within it. This information must then be rehearsed and practiced, as we have explained before, as to be converted into LTM and stored within one of the aforementioned Schemas, or if lacking any previous association to other stored memories, creating its own Schema.

C. Working Memory Capacity Measures

The most widely utilized process to evaluate one's Working Memory Capacity comes in the form of **Complex Span Task (CST)**- exercises that mix memory tasks, such as remembering a sequence of objects, with interleaved secondary processing tasks, like judging the correctness of equations. These tasks are thus designed to engage multiple aspects of our working memory by targeting different brain regions, while limiting one's ability to utilize mnemonics through demanding processes. By utilizing the CST, researchers have been able to establish that the average of items lies around 4, when previously with Simple Span Tasks the number registered was 7 [7].

Many forms of CSTs have been explored throughout the years, with researchers aiming to further understand the mechanics behind such phenomena, as well as find out more accurate measures. The first example of a CST dates back to 1980, with the "Reading Span" by Daneman and Carpenter [5], where they combined the storing process of a Simple Span Task (for example, remembering a sequence of words) with a secondary task like reading the phrase presented. Other examples of relevant tests that were later developed based on this research would include the Operation Span by Turner and Engle in 1989 [6], where participants had to solve arithmetic equations while remembering a list of unrelated words, which proved that the processing content of the CST didn't need to be similar to measure WM; and the Symmetry Span in 1996 by Shah and Miyake [8], in which participants made symmetry judgments and remembered spatial locations, which demonstrated a verbal-spatial distinction when compared to the Reading Span in the prediction of spatial abilities.

D. Cognitive Load Theory

The **Cognitive Load (CL)** corresponds to the amount of resources used from our WM at a time, as defined by John Sweller in 1980 [14]. The Cognitive Load Theory was developed with the intent of helping instructors optimize teaching designs to lower the toll on the CL of their learners, since those with lower WM capacity would struggle to keep up with new information when under heavy load. This theory, in essence, proposed that quality of instructional design would be increased if the limitations of WM were taken into consideration during the process.

Sweller divided CL into three different types: the Intrinsic Load, the Extraneous Load and the Germane Load. The **Intrinsic Load** refers to the load associated with the inherent difficulty and complexity of a given task, varying depending on the degree and number of complex concepts needed for it to be processed; the **Extraneous Load** is the load utilized for processing information that is dissociated from the learning of the current task, including the distractions that otherwise hinder it; and the **Germane Load**, the type of load which process the stored information in order to create the Schemas we have previously touched upon when talking about the LTM.

By combining these three varieties one would then be able to get the CL currently being used, an amount that shouldn't surpass the WM capacity of the individual in question, as that would impede the proper processing of newly acquired information. The most widely used measures of CL are subjective in nature, based on ratings of perceived mental effort and task difficulty, and not objective calculations. Some examples would include the NASA Task Load Index [15] and the 9-point Linkert scale utilized by Paas [16].

E. Generation Effect

Memory researchers have been investigating effective mnemonics- techniques developed to enhance the storage and retrieval of information from our memory- for several years, hoping to understand how such strategies can promote better memorization. Not only are they useful in our daily lives for remembering names and tasks, but they also hold a proven advantage in the educational field. These researchers hope that, by understanding the mechanics behind such strategies, we will further master our studies of memory as a whole.

One such strategy that has shown promise is known as the GE, first explored by Slamecka and Graft in 1978 [9], a phenomenon where information is better remembered if self-generated, as opposed to reading. This type of mnemonic has shown results through many experiments, often using word lists to further study different benefits of self-generated memories. These tests often presented the participants with a list of stimuli, usually pairs of words. Half of the participants would be given intact target words, while the other half would have to self-generate the pair with incomplete words instead[5]. Later stages of the experiment would find better memorization in the section of the group that had to self-generate their answers. That being said, even though many theories have been developed around this effect, there hasn't been a single hypothesis that has been able to satisfy all the questions regarding the mechanics behind the GE.

Furthermore, one factor within this effect has been explored in the hopes of justifying the strategy- known as the **Generation Constraint (GC)**- which refers to the amount of information an individual is given that limits what can be produced in such generated tasks. In essence, these constraints serve to filter the amount of possible answers a participant can give to the problem they face, working as clues to funnel them towards the expected answers. However, there have been works that have disproved the utility of such tactics, showing that testers with fewer constraints can generate better memory benefits. Regardless, studies keep testing various constraint levels to verify the extent to which the GC can influence both item and context memory.

1) Mental Effort Theory: Among the many theories revolving around the GE, there exists one most relevant to our studies. In 1979, Tyler et. al. [10] found a potential relationship between the GE and Mental Effort. Their theory suggested that self-generated information requires more mental effort than other methodologies, i.e., a larger amount of cognitive resources must be utilized to perform a given task. Their results showed that, for instance, high-effort self-generated information led to better retention than low-effort scenarios. Given that mental effort is, by definition, associated with our cognitive resources, this would imply a direct connection to our WM, and as such, this theory could serve as a backbone to prove a logical connection between the GE and our WM capacity which we aim to research in our paper. However, a lack of universal measures and manipulation of mental effort has held this theory back in use, since there is no reliable way to quantify the data it wishes to evaluate, and it has so far relied on subjective measurements instead.

[5]e.g. Intact pair of keywords: Above-Below; Incomplete pairs: Open-C...

III. Implementation

A. Approach

Our goal with this work is to find the best approach to tutorials, as in verifying which method is capable of better adapting to multiple levels of knowledge retention in players- their WM capacity [1]. After all, in video games we often have to learn new mechanics and apply them in complex scenarios, with our WM constantly engaged at least to some degree- the better we retain the information, the quicker and more efficiently can we apply what we have learned. But before anything else, we must determine if the gathered evidence supports our hypothesis, which aims to verify the influence of our WM capacity in a video game environment, as well as compare different approaches to tutorials in order to best adapt for said differences in players.

In order to utilize the desired WM capacity tests, the CSTs [6], not a lot needs to be changed beyond the usual approach to them, since in our situation we can simply apply the existing methodology directly to our project. The only difference it will have compared to the usual CST test relates to where exactly they will be created and applied- in this case, in the context of our Game Engine of choice. Changing its structure and rules would require us to verify its validity, which is our case is an unnecessary step to take.

As far as the approach to the tutorials that will be developed in the scope of this project, we have decided to make use of polar opposite scenarios, utilizing the common disparity between didactic and self-generated learning- or the GE [9]- in the context of a video game, often debated in the education field. While this theory has not been tested in digital games, tutorials are as much of a learning experience as any other given in a classroom, where the teacher is the game itself, and the student the player. Our brain is engaged in similar ways, and information is gathered through the same sections. The Mental Effort Theory [10] we have established before also points to a connection between the GE and our CL, or Mental Effort, which is the usage of our WM resources. As such, we believe this comparison between different teaching methodologies has unexplored potential in our field that has enough evidence to justify it.

With the previous studies we have explored before, we have the necessary backing for the reasoning behind our experiment, and will now explain our approach. As mentioned above, we aim to utilize CSTs to determine the capacity of our participants' WM [5]- we wish to test the influence of this data when applied to real scenarios, opposite to each other. The didactic tutorial approach should provide immediate results when acquiring new knowledge, but fall behind in the long run, while the GE approach has been proven to wield different results depending on the complexity of the tasks,

which would imply an influence in WM levels of the testers GE. Granted, we do not expect immediate results just on the tutorial stage- especially since it has been established that complex tasks take longer to be absorbed through the GE methodology, while wielding better results in later stages of experiments. Because of that fact, we require a longer experiment than usual, with a few initial levels to further test the retention of new knowledge perceived in our participants.

B. Scenario: Nick of Time

1) Concept: In order to test our hypothesis, it was necessary to create its environment first. To accomplish this, a game was designed from scratch, as that would allow us to shape it from the ground up to fit the requirements we needed for this study. This so called environment, the video-game created to house our hypothesis, was titled **"Nick of Time"**, named after one of its core mechanics and overall main appeal. Its original concept saw it becoming a 2D Tactical Role-playing Game (TRPG for short), a video game genre in which we combine the core gameplay elements of a Role-playing Game, in which players control one or multiple characters and immerse themselves in the world the game presents, with strategy elements like turn-based strategy.

The chosen platform to be utilized in the creation of the game for this study was Unity- a cross-platform game engine capable of supporting development of 2D and 3D games, popular as a starting point for game developers, and also widely used among the indie side of the industry. It has been chosen over its competitors due to the considerable amount of accessible assets online to support the creation of this project, given its popularity, and has housed well known and graphically complex titles like "Ori and the Blind Forest"[6], "Cuphead"[7] and "Subnautica"[8], only to name a few.

Tactical Role-playing Games are usually portrayed in grid maps, easily compared to a game of chess, where the player controls their characters like pieces in a board game with a given goal in mind, which includes but is not exclusive to eliminating all the pieces in the enemy's side of the field. One of the most iconic examples of the genre would be the "Fire Emblem"[9] franchise, where you control an army of characters in a medieval fantasy setting, notorious for its permanent death of units combined with their personification to develop an attachment to said characters. **"Nick of Time"** takes inspiration on such titles, but after careful consideration on the time available for our participants to test our game, was further defined to also have a Puzzle game aspect to it- that is to say, it turns the maps seen in Tactical Role-playing Games into a short problem to be solved by our players. In order to do so, certain elements from TRPGs had to be removed, especially given the desirable quick nature of tests like these. Notably, maps are much shorter, there is no permanent death system and the narrative structure was removed, as it simply couldn't afford to be the focus.

In exchange, the maps become shorter and thus quicker to complete. To emphasize that, **"Nick of Time"** has a limited amount of turns available for its players to complete levels, which varies depending on the current stage. The players are given a selection of characters, each with their own attributes distribution and unique abilities, which are given an action and a movement option per turn in order to tackle a challenge in the format of a Tactical Role-playing Game's grid map. Players must make a selection of which units to bring in order to solve the puzzle at hand, which in the game's current version, always means eliminating all enemies before the turn counter reaches zero. Each turn, following the standard rules from the genre associated with it, has a player side and an enemy side, which must be taken into account when formulating a strategy to complete the levels.

However, the game had to deviate somewhat from other tropes of the genre, like specific weapons often seen and associated with a given game element, in order to minimize the influence any previous experience in these types of games might have over our players. Because of this, we chose to avoid the medieval fantasy or futuristic settings seen in the more notable entries of the genre, and instead focused on a mixed fantasy scenario, where all types of creatures from different timelines can be present, thus allowing us more creativity and a broader range of options that will not be immediately understood through association.

2) Architecture: Moving onto the environment itself, Nick of Time's introduction takes the form of a CST test, more specifically a version of the Operation Span Task [12] which was already scientifically proven to work in other aforementioned studies. This version, however, is fully implemented in the game itself in a digital format. In order to avoid potential incoherence with its pen and paper counterpart, the test was kept mostly accurate to its original form, including the timing between the arithmetic questions (Fig. 1) and the sequence input of letters (Fig. 2) that constitute an Operation Span Task. In this case, we opted to do a sequence of 6 operations and 6 letters to remember, the tests akin to this one choosing from 4 to 8. Note that the arithmetic operations are meant to be used only to validate the attempt, and as such any participants who do not meet a quota of at least one operation correctly answered will have their attempt nullified, regardless if they get all the sequences right. This is utilized to make sure participants don't just focus on one side of the test, as the real quantified data will be the sequence to remember. While not directly linked to the rest of the gameplay, this section will be relevant in the data gathering, which we will detail soon. Upon finishing the test, our participants will then be split in the tutorials.

[6]Ori and the Blind Forest, 2015, Moon Studios, Windows/Xbox One

[7]Cuphead, 2017, Studio MDHR, Windows/Xbox One

[8]Subnautica, 2018, Unknown Worlds Entertainment, Windows

[9]Fire Emblem is a Tactical Role Playing Game video game franchise developed by Intelligent Systems and published by Nintendo, originally released on the Famicon, where your units are individual characters that can die and be lost permanently

Fig. 1. An example of an arithmetic operation in the CST test.

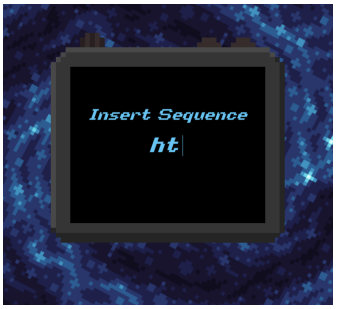

Fig. 2. An example of the sequence input in the CST test.

Fig. 3. A screenshot of the Generation Effect tutorial, where the player skip each section at their leisure, as well as an example of the character selection.

The tutorials we have mentioned, the GE tutorial and the didactic tutorial, have a few select differences when compared to regular levels: they lack a turn limit, allowing our participants to have as much time as they desire to experiment in the training area before they jump onto the challenges, and players will be able to revisit the character selection screen as many times as they please without having to restart the level (see the example in Fig. 3). These tutorials will share the same map and dummy enemies, their main difference lying instead in how new information is provided to the players. Our testers will be divided between a GE focused tutorial and a more didactic one, where the GE tutorial will only give a minimal amount of information, thus only explaining the very basic of each mechanic and letting the players complete the missing pieces by themselves, while the didactic tutorial will portray the same elements extensively, guiding the players step by step. What we aim for here is to in one case give the players total freedom on how they learn, while in the other scenario they are given everything they must know. To drive the point across, the GE players will always be able to skip ahead of each tutorial section, while the didactic participants must first follow the lesson before giving the option to continue.

The difference will be most notable when the core mechanic of the game is revealed: the “Sync” action. Characters will each have their unique set of abilities, and through teamwork, the player will be able to combine these units and enhance their techniques in order to beat increasingly more difficult challenges that the individual characters wouldn’t be able to on their own- or at least, not as efficiently. To further exemplify the differences in tutorials, the GE method will only mention how to access these skills and what they are, while the didactic one will detail more information and take players step by step through each subcategory of the action type.

After the tutorials, all participants will converge into a sequence of three levels (see Fig. 4 as an example), with the difficulty increasing after each one. Each individual map will have its own layout, both of terrain structure and enemy positioning, as well as the turn limit previously mentioned. Part of the gameplay loop includes studying these changes and adapting, which is an important step on what we aim to study in our research. Common to all maps, however, is how it starts: with the character selection menu. Since players must only pick a few of the units available to them, they are given the option to view the map and inspect enemies before selecting said units, organizing them in the starting positions available to them. Once the player is ready, they can begin the challenge when they see fit, and restart if they believe they have done something wrong whenever they please. All the maps available will start with the player’s side, moving onto the enemy turn once all units have finished their actions, and then back onto the player’s turn until the counter at the bottom reaches zero or a player manages to clear out all enemies present in the map.

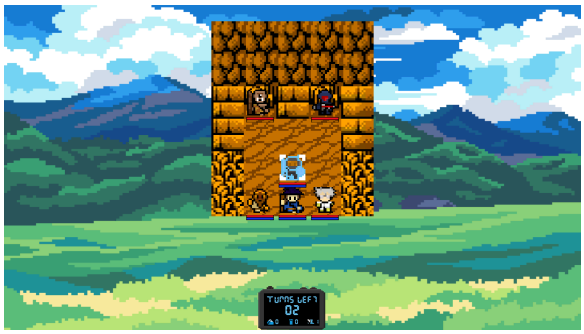

Fig. 4. A screenshot of one of the levels, a shorter map with two very strong units. They each can defeat any of player unit on their own, but have exploitable weaknesses using the ”Chain” action.

C. Collected Data

Our collected data will be separated in three different groups. First, participants will answer demographic questions, used to ensure we do not make incorrect conclusions with the data collected based on aspects outside of our experiment, such as their experience with the genres of video-games utilized. Following it, we will have scripts running during the playtime of Nick of Time, which will collect data on real-time based on the players’ performance, like the time spent on each level and the amount of tries needed to complete a level, as well

as their results in the CST test. Finally, we will add another section in which players must answer subjective questions based on the workload felt during the experiment. Most of these questions will take on the form of a simplified version of NASA's Task Load Index [17], or NASA TLX for short, which is a assessment tool used to rate perceived workload in a given task, developed by the Human Performance Group at NASA's Ames Research Center. This part of the survey will help us verify the quality of our learning environment, as the workload refers to the amount of CST resources used [18].

IV. Evaluation

A. Procedure

Our procedure can be divided in three main segments, all of which were done online, which followed the setup by the "Collected Data" section; an initial demographic section, the short video-game Nick of Time in which the main tests will be setup and a follow-up section with subjective questions regarding workload. To accomplish this specific structure, we decided to make use of a Google Forms, where we present all three of these stages. Our first page consist mainly of a short explanation of our study and its intended purpose, as well as serving as a consent form for the participation of the remaining sections. The participants that chose to continue were presented with the demographic questions, previously explained in the last chapter, and only once it is concluded will they gain access to the second section, where a short explanation of Nick of Time, together with a drive link to download it, are presented.

The player will then jump onto the game, which will begin the CST test. Considering the nature of the test, we provided our participants with an explanation of how it operates, including a free trial of a short version of it, where the participant can retry as many times to get used to the short timers. Only the player feels ready, they can begin the real trial. Once it is finished, the game will take the player to one of two tutorials, the GE version or the didactic one, which will follow the behavior explained previously. Following the tutorial, the players will be moved onto the first of three levels, and will be tasked to complete as many of them as possible. Considering the difficult nature of the challenges presented, they were given the option to back out once they felt like giving up. While we understand that doing so might persuade players to throw in the towel early, we couldn't afford to simply frustrate the participants beyond an unnecessary level. Regardless, we believe their attempts and early surrender will give us valuable information regarding the tutorials and their CST levels.

Once the player concludes the video-game, be it by finishing the last level or giving up somewhere through the levels, they must then return to the Google Forms from earlier. There, they will be requested to submit a file generated by the game, containing all their collected data throughout the video-game experience in Nick of Time, as well as provide feedback on the tutorial they had received.

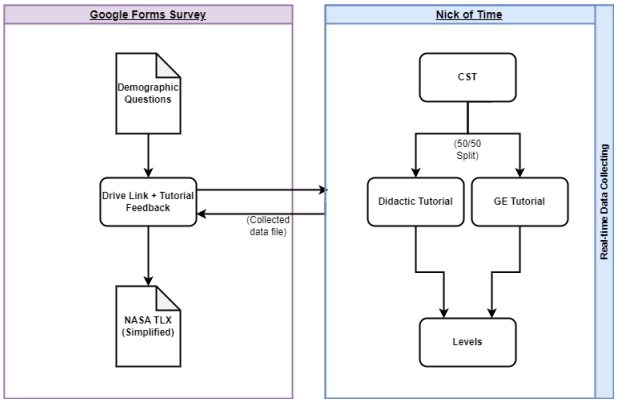

Fig. 5. The procedure utilized in this research.

After doing so, the final section will be available, containing the simplified version of the NASA TLX, constituted by six simple questions regarding the workload felt during the experiment, regarding: Frustration, Temporal Demand, Physical Demand, Mental Demand, Performance and Effort, which must be scored from 1 to 10, as opposed to 1 to 20, due to Google Forms' restriction on scale questions. This section will conclude the participant's task, and the overall structure we followed throughout the experiment described here can be seen in the figure down below (Fig. 5), for the sake of clarity.

B. Sample

To start off, we ran a short pilot in order to ensure the process could be done without any major hiccups that were not considered or tested previously. The pilot, much like the real run of the experiment, was done online through the use of the aforementioned Google Forms and Google Drive link. We asked 5 participants to help solidify the experiment, and thanks to it, we were able to discern some significant bugs and, more importantly, the difficulty of the levels. The issue lied in the order in which the levels were presented, and as such they were correctly shuffled to better suit their difficulty. Any bugs found during this short pilot were also taken care of.

The sample itself, however, only amounted to 17 new participants after the original test run, and as such, does not possess enough data to draw any proper conclusion. Given this fact, we will treat the results presented in the remainder of this document as a pilot study, helping future implementations of this theory to have a base defined for them. That being said, we will still be analyzing the collected data and verifying our hypothesis, based on the limited data we gathered. The data presented was extracted from the Google Sheet generated by the Google Forms and another extracted from the data generated by the game and delivered by our participants in the questionnaire.

To start us off, we will analyze the demographic section. From our sample, 14 (82,4%) participants identify themselves

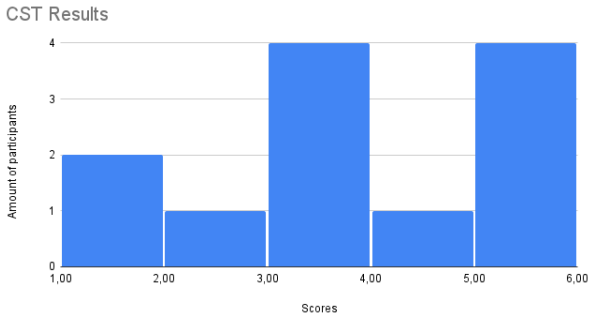

Fig. 6. Bar chart with the different participants' CST scores.

as male, with only 3 (17,6%) identifying as female, while the age average collected rounds to 25, with ages ranging from 21 to 34 years old.

Out of the 17 participants, all but one confirmed they make time in their schedule to play video-games regularly, with the outlier only playing them socially. As far as the genres presented are concerned, a majority of our participants enjoy (41,2%) or favor (35,3%) Tactical Role-playing games, with the remaining percentage either not having a formed opinion (11,8%) or not favoring them (11,8%). Puzzle games ended up having more defined opinions, with close to half enjoying them (47,1%), but about a quarter not favoring the genre (23,5%), the remaining participants favoring Puzzle games as one of their favorites.

Moving onto the data collected through the scripts in Nick of Time, we can verify that our CST results have a median score of 3.5 on a scale of 1 to 6, meaning that it lands somewhere between 3 to 4 correct letter sequences. However, 3 out of the 17 participants did not get enough inquiries correct to qualify, be it from a lack of correct arithmetic operations or in general, reducing our real data sample to 14. Out of those 14 participants, 9 ended with the GE tutorial and averaged 7,8 minutes spent in it, the remaining 5 getting the didactic tutorial with an average of 16 minutes spent in it. Those with above average scores in the CST spent more time in their tutorials, around 11,3 minutes, while those with below average scores only spent 7 minutes in their respective tutorials.

On the levels, participants needed an average of 3 tries to beat the first level, with 5 participants giving up on it and only one having a low CST score. On the second level the tries required increased to an average of 8 with 2 more participants giving up, the remaining ones moving onto the final level with an average of 11 tries and 3 more quits. This means that only 4 participants truly concluded Nick of Time, with only 1 of them taking the didactic tutorial in the beginning, and only 1 of them having a low CST score, albeit barely with a 50%. It is also noteworthy to mention that all of those from the GE tutorial that finished were experienced players in the genre, while the only player that beat Nick of Time from the didactic tutorial only had some experience with Tactical Role-playing Games, but did not enjoy Puzzle games. Out of the didactic tutorial participants, 3 quit right from the first level, meaning the results from that tutorial were cut shorter than expected.

When looking at the second part of the Google Forms survey, we can note that the didactic tutorial scored an average of 6 out of 10 from the subjective opinion of our participants, while the GE tutorial scored 5,5 with a bigger sample of testers. In the NASA TLX section, Mental Demand averaged on 7,6 out of 10, Physical Demand only averaged on 3 out of 10, Temporal Demand scored 5,9 out of 10, Performance on 5,3 out of 10, Effort with 7,5 out of 10 and finally Frustration with 6,5 out of 10. That means that on total the task our participants performed was rated a 35,8 out of 60, on average.

C. Results

Before we theorize with our results, it must be clarified again that the sample collected is not large enough to determine anything concrete, and as such we will analyze the data from the point of a view of a pilot instead. We had hypothesized that participants with lower WM levels, i.e., that scored lower in the CST test, would tend to perform better in a didactic environment, since it would guide them slowly through what they had to learn, while those with higher levels would perform better in the GE tutorial, since their higher capacity would allow them to understand mechanics quicker with the freedom provided from a more exploratory approach.

First, we can verify that participants in the didactic tutorial stayed in it for considerably more time than the participants in the GE tutorial stayed in their respective environment. That is to be expected, considering how the didactic approach takes its time to guide the player through the mechanics in detail, but the fact that the average time ended up being almost doubled when compared to the GE approach may indicate that when a player is given the chance to dictate their own rhythm, it will lead to them potentially investing less time in learning or trimming out information they might already be familiar with, as we will soon discuss.

Most of those that concluded all the levels originated from the GE tutorial, and managed to complete said levels in less turns than the one participant from the didactic side. Additionally, we can verify that most of the participants that finished Nick of Time were, indeed, those with higher WM presented. In fact, the only participant from the didactic tutorial that finished the game had an average to low score in the CST, meaning that the present data does not contradict our original theory, and those with higher WM do perform better in an exploratory approach like the GE teaching method, while those with lower levels perform better with didactic teaching methods.

Another point to bring up is how those that beat all

the stages in the GE tutorial were players that favored the genres present in Nick of Time, while the one participant on the didactic side only enjoyed Tactical Role-playing Games, but not Puzzles. The GE approach does favor experienced participants in video-games, even more so those with experience in the genres selected, due to it allowing players to select which areas to explore and dedicate time to. Meanwhile, those that lack in experience might do better in the didactic tutorial, since it provides more information on topics the experienced player might already be familiar with. This could imply that different approaches favor experience instead of WM levels.

Unfortunately, we do not have enough data to support either theory just yet, even if the data prove favorable thus far. There are too many dimensions to consider, most of which share conditions and are interconnected, which makes it difficult to distinguish the real cause with a small sample. On that note, it should be mentioned that both tutorials scored similarly on the survey, so no concrete evidence could be derived from that data thus far. This may be due to the questions' vague nature that could lead those answering it to be rating it from a critical stand point, and not from their personal preference. The average score the task received through the simplified NASA TLX was relatively middling, meaning that the video-game environment was successful enough not to overwhelm our players' memory capacity [18].

V. Conclusion

A. Summary of the Work

Video games grow more and more complex as years pass by, especially considering how quickly it has done so. However, it creates a problem derived from such an exponential growth: more complex games bring more complex mechanics, and it creates a need for good tutorial environments that can properly convey the information in a clear and easy to retain way. Video games do not take into account the different learning growths, and that pushes part of the player-base away- but we hypothesize that the key to solving this problem lies outside our industry, and instead connects to our memory, which is the one who is truly responsible for gathering new knowledge.

To accomplish this, we imported a variety of concepts and theories from cognitive psychology in order to properly understand how to analyze it in players and derive conclusions. First, we had to comprehend how it is believed our memory is structured, and learnt that it can be divided in two major components: the LTM and the STM [1]. The LTM is responsible for storing long-term information and skills, with a potentially unlimited capacity and for a long time. The more new information is trained, the longer it tends to stay in our memory. Meanwhile the STM, also known as WM, is responsible for holding information temporarily with a limited capacity, which varies from person to person, and is responsible for a lot of our decision making, and is the first step towards the retention of new knowledge. The bigger our WM, the bigger the amount of information we can process at a time.

The key to our theory lied in the conversion between the two, and if we can optimize it in our tutorials regardless of our players' WM capacity. To test it, we needed to measure this capacity, and watch how different levels behaved in different learning environments, in order to understand if there was a best approach to memory retention. For us to measure the WM capacity, we made use of the CST test [5], exercises that mix memory tasks, such as remembering a sequence of objects, with interleaved secondary processing tasks, like judging the correctness of equations, with the objective of targeting different brain regions. In our case, we made use of the Operation Span Task [6], that makes its participants remember an increasingly larger sequence of consonants while mixing in arithmetic operations between each sequence. We then turned our attention to the tutorial environments we wished to study. That is when we landed on two opposite approaches, often compared in the education field, the didactic and GE [9] approach. The latter, not yet applied to the video-game industry to our knowledge, made use of incomplete information to prompt students to complete the missing pieces, and thus retain the information better. We believed that such an exploratory approach, where the player must find out the details of each mechanic on their own, could prove to be a useful comparison to a relatively common tutorial like the didactic one, where detailed information is slowly given to the players.

We believed that the didactic tutorial approach should provide immediate results when acquiring new knowledge, but fall behind in the long run, while the GE approach has been proven to wield different results depending on the complexity of the tasks, which would imply an influence in WM levels of the testers GE, with the potential to provide better results with higher levels of WM. Studying its influence in both approaches was set to be our goal, and in order to do so, we required a testing environment for our hypothesis- thus, we created a video-game for the purposes of this work by the name of "Nick of Time", a Tactical Role-playing game combined with a Puzzle game with an unique approach to both genres and a moderately difficult learning curve, constructed that way to ensure the differences between the learning methods and different WM levels would be noticeable in a shorter experiment. The game was structured in three main sections: first, the CST test, which we converted into an introduction to the game itself, then one of the two tutorials selected at random, where the players were taught how to play the game using different methodologies, and finally a sequence of three levels in increasing difficulty.

With the blueprint set, we then decided on how to gather the necessary data for our research. First, we needed to get the demographic information of our sample, to ensure any outside influence such as experience in games, familiarity with the genres and age would be taken into account when analyzing the information. Then, we set up in the game itself scripts that would collect data on the players' performance

in each section, from collecting their results on the CST test, to which tutorial they got and how long they spent on it, to how many turns one took to complete a level and which actions and units were selected. Additionally, we decided we should reinforce our data with a subjective approach, making use of a simplified NASA TLX version [17] to measure the workload felt by our participants.

Our procedure was then separated in the same three stages, all setup in a Google Forms, where participants were first asked to fill out the demographic questions, then be given a link to download and play "Nick of Time", as well as deliver the resulting file generated by the game, and finally answer a section with an additional feedback question regarding the tutorial they got as well as the simplified NASA TLX. First, we ran a pilot to correct any major flaws in our setup, and then proceeded with the real test run. Unfortunately, we did not manage to gather enough participants in the remaining time, and thus analyzed our data as a pilot run as opposed to the full procedure.

From the gathered results, we could note that participants in the didactic tutorial took almost double the amount of time, on average, to conclude it when compared to the GE tutorial. This could mean that when given the chance, players will dictate their own rhythm and conclude the tutorial earlier. Since players were given the option to back out during the levels, only a select few participants were able to finish all three levels presented, but interestingly, most of those that finished were GE participants. On top of that, those with more experience in the genres presented saw better results on the GE approach, potentially due to the fact that its exploratory approach allows them to be selective on what needs to be learned, while those without experience did better on the didactic tutorial, perhaps thanks to how it teaches things slowly and in detail. While not conclusive with the amount of data collected, the results gathered do not contradict our hypothesis thus far, and instead built upon what we believed to be true. Additionally, the average score the task received through the simplified NASA TLX was relatively middling, meaning that the video-game environment was successful enough not to overwhelm our players' memory capacity [18].

B. Limitations and Future Work

Ideally, a future work should be attempted with a much larger sample of players, as our test run was not significant enough in size to make conclusions. We believe that the main cause of the low amount of participants might be due to its time-consuming nature, taking from 30 to 60 minutes to undertake. When doing it online, a test run should take less time as to avoid any participants being dissuaded from taking it, an issue that is less prevalent on site.

Unfortunately, we do not believe shortening the game would be possible, as something of this genre of video-games takes time to explain their mechanics, and simplifying it further would make it considerably easier to understand and complete, which would make the data collected overall higher than normal. It could be possible to lower the overall difficulty of some of the levels, especially the last one, as to avoid so many of the participants quitting earlier, since the difficulty itself might have lost us some potential testers. There were also quite a few complaints in regards to the controls chosen for the game, and with more time, we would have liked to convert it to mouse only, or simply change the setup controls to something more intuitive, especially since issues with the placement of keys in foreign keyboards were felt by some of the participants.

REFERENCES

[1] Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Psychology of learning and motivation (Vol. 2, pp. 89-195). Academic Press.

[2] Clopath C. (2012). Synaptic consolidation: an approach to long-term learning. Cognitive neurodynamics, 6(3), 251–257.

[3] Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. Journal of neurology, neurosurgery, and psychiatry, 20(1), 11.

[4] Baddeley, A. D., & Hitch, G. (2001). Working memory in perspective. Psychology Press.

[5] Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. Journal of verbal learning and verbal behavior, 19(4), 450-466.

[6] Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent?. Journal of memory and language, 28(2), 127-154.

[7] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81–97.

[8] Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: an individual differences approach. Journal of experimental psychology: General, 125(1), 4.

[9] Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. Journal of experimental Psychology: Human learning and Memory, 4(6), 592.

[10] Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. Journal of Experimental Psychology: Human Learning and Memory, 5(6), 607.

[11] Nacke, L. E., Bateman, C., & Mandryk, R. L. (2014). BrainHex: A neurobiological gamer typology survey. Entertainment computing, 5(1), 55-62.

[12] Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. European Journal of Psychological Assessment, 28(3), 164.

[13] Green, M. C., Khalifa, A., Barros, G. A., & Togellius, J. (2017, September). " Press Space to Fire": Automatic Video Game Tutorial Generation. In Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference.

[14] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), 257-285.

[15] Hart, S. G. & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload. Amsterdam: North Holland Press.

[16] Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. Journal of educational psychology, 84(4), 429.

[17] Hart, S. G. (1986). NASA task load index (TLX).

[18] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), 257-285.