Mikael Nylund

An assessment on video game tutorial learnability from a cognitive load framework

An eye tracking study on Baba Is You and Terra Nil

School of Marketing and Communication
Master's thesis in Communication Studies
Master's Programme in Technical Communication

Vaasa 2025

VAASAN YLIOPISTO
Markkinoinnin ja viestinnän akateeminen yksikkö
Tekijä: Mikael Nylund
Tutkielman nimi: An assessment on video game tutorial learnability from a cognitive load framework : An eye tracking study on Baba Is You and Terra Nil
Tutkinto: Filosofian maisteri
Oppiaine: Teknisen viestinnän maisteriohjelma
Viestintätieteet
Työn ohjaaja: Niina Nissilä & Rebekah Rousi
Valmistumisvuosi: 2025 **Sivumäärä:** 81

TIIVISTELMÄ:
Tutkielmassani selvitin, mitkä osatekijät vaikuttavat videopelitutoriaalien opittavuuteen pelaajan kognitiivisen taakan näkökulmasta. Tutkielman tavoitteena on tutkia, miten tutoriaalit voivat auttaa parantamaan videopelien opittavuutta pelaajien kognitiivisen taakan näkökulmasta optimoimalla pelaajien kognitiivisen taakan. Tutkielman aineistona ovat testaamalla saatu katseenseurantadata fiksaatio- ja sakkadimäärinä, työmuistin kapasiteetin arvio, subjektiiviset työtaakka-arviot, sekä osa-aineistona osallistujien demografiset tiedot ja pelit itse. Fiksaatio on fysiologinen ilmiö, jossa katse pysyy kiinnittyneenä kohteeseen. Sakkadi on fysiologinen ilmiö, jossa katse vaihtelee kahden tai useamman pisteen välillä nopeasti. Tutkimuksen 27 osallistujalle selitettiin ensiksi tutkimussuunnitelma, ja he saivat aikaa tutustua ympäristöönsä. Osallistujien fiksaatioiden ja sakkadien määrät mitattiin tutkimuksessa katseenseurantalaitteella heidän pelatessaan pelin tutoriaalia läpi. Näin pyrittiin selvittämään fysiologisia mittareita hyödyntäen, miten videopelin tutoriaali vaikuttaa kognitiivisen taakan optimointiin. Pelihetken aikana osallistujat reagoivat ääniärsykkeeseen, jolla mitattiin sekundaarisen tehtävän vastausajalla työmuistin kapasiteettia. Osallistujat jaettiin kahteen ryhmään näiden tulosten perusteella, jotta niiden pohjalta pystyttäisiin arvioimaan työmuistin kapasiteetin vaikutuksia kognitiivisen taakan optimointiin. Pelihetken jälkeen osallistujat vastasivat työtaakkakyselyyn. Kaikki arvot validoitiin tilastollisin metodein, tarkastellen mahdollisia eroja ryhmien ja pelien välillä opittavuuden ja kognitiivisen taakan näkökulmasta. Pelien opittavuus on keskeinen tekijä sekä uuden pelaajan houkuttelussa ja hauskanpidossa että tekijä, joka vaikuttaa myös kokeneempien pelaajien pelikokemukseen. Opittavuuteen vaikuttaa pelaajan kognitio ja täten myös hänen kognitiivinen taakkansa. Osallistujat pelasivat heille annettua peliä 15–30 minuuttia, minkä aikana he vastasivat työmuistia mittaavaan toissijaisen tehtävän vastausaika -testiin. Katseenseurantadataa kerättiin samaan aikaan. Osallistujien suoriutumista pelissä mitattiin. Suoriutumisen mittarina toimi aika, mitä osallistuja käytti tason suorittamiseen. Pelihetken jälkeen osallistujat vastasivat työtaakan arvioinnin kyselyyn. Tulokset osoittavat, että subjektiivinen työtaakka ei korreloinut fysiologisten kognitiivisen taakan markkereiden kanssa, ja tutoriaaleilla oli osittainen kognitiivista taakkaa optimoiva efekti. Tästä huolimatta tutoriaalien vaikutukset kokeneempiin pelaajiin on epäselvä, sillä suoriutumiset tutoriaaleissa olivat hyvin samankaltaisia molemmilla ryhmillä. Työtaakka yleisellä tasolla oli kohtalaisen matalaksi arvioitu, mutta tilastollisesti merkitsevä ero oli huomattavissa ryhmien välillä. Suoriutuminen pelihetkestä oli ryhmien välillä samanlainen, vaikkakin suuremman työmuistin ryhmään kuuluvat osallistujat näyttivät vähemmän kognitiivisen taakan fysiologisia markkereita. Tulosten mukaan voidaan päätellä, että tutoriaalit vaikuttaisivat auttavan kognitiivisen taakan hallinnassa ja parantaneen opittavuutta osittain, mutta lisätutkimus näiden vaikutusten määrästä on tarpeen.

Avainsanat: video games, games, learnability, cognitive load, tutorial, playability, human-computer interaction

Contents

1 Introduction 7

1.1 Aims 8

1.2 Materials 9

1.3 Methods 16

2 Cognitive load, its components, and learnability in video games 20

2.1 Cognitive load and its components 20

2.2 Learnability 24

3 Cognitive load theory and learnability in video game tutorials 27

4 Explanation of the experimental design 31

5 Analysis and results 36

5.1 Integrity of the dataset and participant performance evaluations 36

5.2 NASA-TLX -findings 43

5.3 Findings from eye tracking 47

6 Discussion and limitations 63

7 Conclusion 67

References 68

Appendices 72

Appendix 1. Consent form page 1 72

Appendix 2. Consent form page 2 73

Appendix 3. Data protection notice page 1 74

Appendix 4. Data protection notice page 2 75

Appendix 5. Textbook instructions in the beginning of Terra Nil 76

Appendix 6. Instructional pop-up in Terra Nil 77

Appendix 7. Baba Is You's stage 2 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line 77

Appendix 8. Baba Is You's stage 3 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line 78

Appendix 9. Baba Is You's stage 4 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line 78
Appendix 10. Baba Is You's stage 5 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line 79
Appendix 11. Baba Is You's stage 6 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line 79
Appendix 12. Baba Is You's stage 7 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line 80
Appendix 13. Terra Nil's fixations per minute plotted against the difficulty coefficient with a linear relationship line 80
Appendix 14. Terra Nil's saccades per minute plotted against the difficulty coefficient with a linear relationship line 81

Figures

Figure 1. On the left, the first stage of Baba Is You. On the right, the second stage of Baba Is You, where the rules affect the player character. **12**

Figure 2. On the upper left-hand corner, first task in Terra Nil. On the bottom left-hand corner, climate control tutorial in Terra Nil. On the right, a notification on fauna appearing in Terra Nil. **13**

Figure 3. Terra Nil's difficulty settings. **13**

Figure 4. NASA Task Load Index. **17**

Figure 5. Stage zero linear regression. **41**

Figure 6. Stage one linear regression. **42**

Figure 7. All stages linear regression. **43**

Figure 8. Example of eye tracking features in Baba Is You. **52**

Figure 9. Example of eye tracking features in Terra Nil. **53**

Figure 10. Relationship of stage difficulty coefficient and fixations per minute in Baba Is You. **55**

Figure 11. Relationship of stage difficulty coefficient and fixations per minute in Baba Is You. **56**

Tables

Table 1. Participants' demographic breakdown. **10**

Table 2. Types of data gathered in the experimental design. **15**

Table 3. Tutorial impact framework. **33**

Table 4. Tutorial impact framework applied. **34**

Table 5. Successful STRT-test results. **36**

Table 6. Analysis groups based on working memory evaluation. **37**

Table 7. Baba Is You performance data. **38**

Table 8. Terra Nil performance data. **39**

Table 9. ANOVA values on fixations and saccades per stage. **40**

Table 10. NASA-TLX Weighted Ratings. **45**

Table 11. Fixation counts for Baba Is You. **48**

Table 12. Saccade counts for Baba Is You. **49**

Table 13. Fixation counts for Terra Nil. **50**

Table 14. Saccade counts for Terra Nil. **51**

Table 15. Fixations per minute per stage in Baba Is You. **57**

Table 16. Saccades per minute per stage in Baba Is You. **58**

Table 17. Fixations per minute per stage in Terra Nil. **59**

Table 18. Saccades per minute per stage in Terra Nil. **60**

Table 19. P-values on Baba Is You's fixations and saccades per minute between Group 1 and 2. **61**

Table 20. P-values on Terra Nil's fixations and saccades per minute between Group 1 and 2. **61**

1 Introduction

Instructions and tutorials serve an important role in any usage context where they may be needed. A good set of instructions can help guide a user to an acceptable level of skill within the usage context or help deepen their understanding and usage capabilities within the usage context. The importance of good guidance is also prevalent in computer technologies, which bring a whole new dimension outside of the physical into the matter, adding complexity by the sheer possibilities of depth that computer programs can have.

The field of human-computer interaction (HCI) considers the interaction between humans and computer technology and what effects it can have. The cognitive load, or the current load on the working memory, of a user is another factor that greatly affects the use of any system (Lee & Heeter, 2017; Chang et al., 2017; Sevcenko et al., 2023). Design choices made can help in easing the cognitive load of a user or give tools to facilitate better interactions with technology (Li et al., 2024). Usability of any system can also factor into the ease of using it, and as such the usability of a tutorial can affect how much information the user is able to process. This goes into the learnability of technology, and how design choices can improve tutorial usage.

Video games as rich multimedia often associates usability with playability. Playability can be defined as the “quality of a video game in terms of its rules, mechanics, goals and design” (Sanchez et al., 2012, p. 3). One aspect of playability is learnability. Learnability as defined in Poretski and Tang’s (2022, p. 2) paper considers the aspects of the software, in this case a video game, that aid the user in learning and improving their usage of the software. The definition of learnability used in this thesis mimics Pretorious et al. (2010, p. 2), who define it according to Dix et al. as “the ease with which users can enter a new system and reach a maximal level of performance.” Similarly, as Poretski and Tang (2022, p. 2-3) do in their paper, this thesis looks at the specific aspects of approachability and usage efficiency, the latter of which is referred in literature as accessibility. Tutorials are one extremely common way of improving game learnability. Learnability in games can be realized in an array of teachings, ranging from giving a new user information on basic

functions, to showing nuanced interactions or possibilities to an experienced one. Tutorials can also range in their forms, such as contextual tips, levels that act as guided tours, or giving options for experimentation in a sandbox level (Poretski & Tang, 2022, p. 6).

Video games can be defined as computer-based multimedia that consists of interactive elements, set goals and rules which to follow (Chang et al., 2017, p. 2). Video games offer us an interesting look into how learnability can be affected by design choices. Not only do video games span different genres while still maintaining mostly the same mechanical devices, but they also often offer a complex set of interactions from physics to design elements, which need to be conveyed to players. Games by nature are very rich in multimedia content. Despite this, research is more focused on educational or serious games. Games broadly as a form of entertainment function as inspiration for developing the more serious alternatives. A video game tutorial can be defined as instructive content in that aims to familiarize the players with the game and its rules and mechanics (Cao & Liu, 2022, p. 1-2). How a tutorial is designed can affect user engagement and motivation to continue as well (Cao & Liu, 2022; Andersen et al. 2012, p. 8-9). We can also apply some of the principles observed from how video game tutorials aid in improving learnability to other computer programs as well, especially ones comparably rich in multimedia content.

1.1 Aims

The aim of this thesis is to explore how tutorials can aid in improving video game learnability for players from the framework of cognitive load theory by optimizing player cognitive load. In the context of this thesis, Pretorious et al. (2010, p.2) maximal level of performance will be used to measure a players understanding of the necessary functions of the game to the maximum effect. This thesis seeks to answer two research questions:

1. How do video game tutorials affect a player's cognitive load?
2. How does cognitive load affect the learnability of the game?

This thesis considers a few hypotheses, explained in detail in chapter 1.3. The theoretical framework of this thesis centers around video game learnability research and cognitive load theory. The thesis also discusses usability and user experience (UX) briefly as aspects of HCI, which is also related to the research of cognitive loads in technological contexts. As the topic and materials of the thesis focus on games, game studies are also present in various ways in the theoretical framework, usually combined with the aforementioned disciplines or related disciplines such as pedagogy, studying matters such as cognitive loads through an educational gamification perspective. Learnability is also a concept that is multidisciplinary, connecting HCI with usability and UX to psychology, pedagogy and other fields.

1.2 Materials

The participants of this study include gamers and non-gamers from ages 16-55. As the aim of this study does not necessitate gaming experience, knowledge or the lack of those traits, a wider sampling of participants was chosen. The main requirement of the participants was proficiency or comfort with using computers. The participants were acquired through convenience sampling, utilizing school and workplace personal connections to recruit participants. Demographic information gathered included participants' age, if they had any prior experience playing the game they were allocated, as well as a subjective rating of their experience in playing video games. This breakdown is shown in Table 1. Information in the table is sorted in age from youngest to oldest. Average age of participants was 32, with the median age group being in their 20s. Participants were split into two groups based on their working memory capacities, with Group 1 denoted by an asterisk (*). Groupings are explained in depth in Table 6.

Table 1. Participants' demographic breakdown.

Participant	Age	Prior Knowledge	Gaming Expertise
P25*	16	None	4
P16	18	None	3
P23	20	Heard of it a little	4
P10*	20	Via YouTube videos	5
P13*	20	None	4
P11*	21	None	5
P18*	21	None	5
P15	21	None	4
P14	23	None	5
P8*	23	Moderate experience	5
P1*	24	None	4
P5*	24	Played a demo, much	4
P6*	24	None	4
P7	25	None	4
P3*	26	None	2
P19	26	None	2
P22	36	None	1
P2*	38	None	2
P21	39	None	4
P17	47	None	2
P12	47	None	3
P4	49	None	3
P24	49	None	3
P9*	50	None	1
P20	51	None	2
P26	54	None	0
P27	55	None	2,5

The materials used in this thesis can be split into main materials and sub-materials. The main materials consist of data gathered from the participants in the form of eye tracking, subjective workload assessment via survey, secondary task response time -test (STRT)-results. The sub-materials include the games themselves, as they are bound to affect the nature of the experiment. This thesis will look at two games from different genres. This was done to observe possible differences between the games' design choices that could affect the players' cognitive load (Poretski & Tang, 2022, p. 5). The games were accessed by downloading them onto a USB-stick to be played locally at the VME Interaction Design Environment. The games were chosen based on a set of criteria. Firstly, the game should be easily obtainable and usable to the researcher. Secondly, the game should be relatively effortless to run in the experimental setting. Some games can be resource intensive, which can increase the set-up time for the experimental design or decrease the availability of participants per day to run the experiment on. Lastly, the game should require fairly low mechanical ability or gaming knowledge. This is so that the effects on cognitive load are affected as minimally as possible by mechanical requirements.

The games chosen for the thesis are Baba Is You (Hempuli Oy, 2019) and Terra Nil (Free Lives, 2023). Baba Is You is an independently developed puzzle game, published by Hempuli Oy in 2019. The puzzles in the game involve manipulating sentence structures found on a grid to formulate a solution. An example of this can be found in Figure 1. The game is fairly lightweight in its requirements and has no mechanical ability requirements from the player. Only a reasonable level of basic English grammar is required, as solutions are offered via alternating sentence structures. The level is low enough so that it should not manifest any significant knowledge level issues that would be needed to account for in the experimental design.

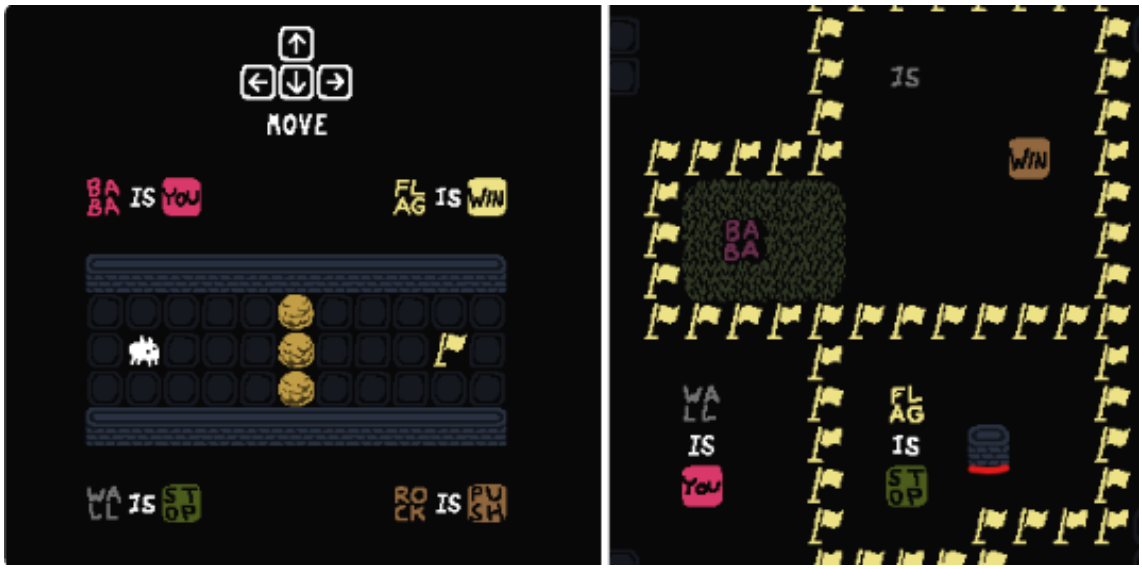


Figure 1. On the left, the first stage of Baba Is You. On the right, the second stage of Baba Is You, where the rules affect the player character.

Terra Nil is a nature-preservation -themed strategy game developed by Free Lives and published by Devolver Digital in 2023. The game is based on the city-builder and top-down- strategy game genre. A top-down strategy game utilizes a birds-eye view of a game board that is usually based on a grid. The player's mission is to restore a barren landscapes ecosystem to a flourishing level by placing objects such as irrigators, wind turbines, etc. around the map. Using these objects deducts points, whilst their effects give them in proportion to the amount of nature restored to the landscape. The gameplay begins with basic aspects such as energy, soil, and water and moves onto more precise climate restoration such as humidity of the area or flora and fauna, shown in Figure 2. More examples of instructional messages are found in Appendices 5 and 6.

The games were played until a pre-determined time limit or progression stage was met. The time limit for both games was the same, between 15-30 minutes. Progression stage for Baba Is You was the end of the first block of stages, from the tutorial through to stages 01-07. This was due to the stages taking enough time to fit the time limit as well as teaching the core aspects of Baba Is You's gameplay. For Terra Nil, the progression stage was the end of the tutorial level itself. The tutorial was played on gardener-difficulty,

presented in Figure 3, which is the second easiest difficulty as of writing this. This difficulty was chosen to make the game more accessible for the average player, limiting the resource-management aspects to a lower level or removing them, but without removing the difficulty altogether.

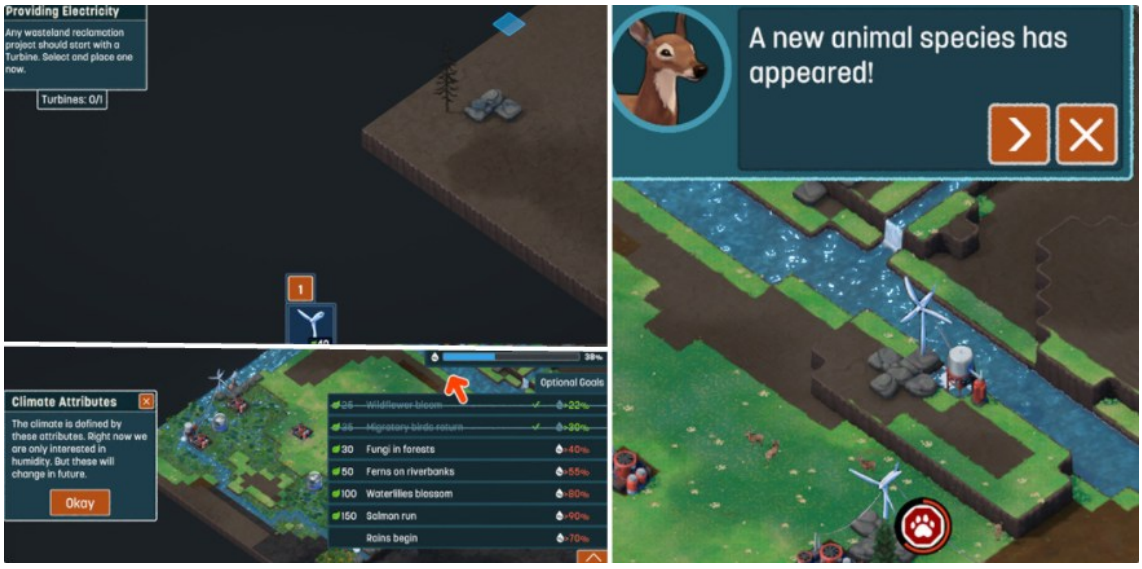


Figure 2. On the upper left-hand corner, first task in Terra Nil. On the bottom left-hand corner, climate control tutorial in Terra Nil. On the right, a notification on fauna appearing in Terra Nil.

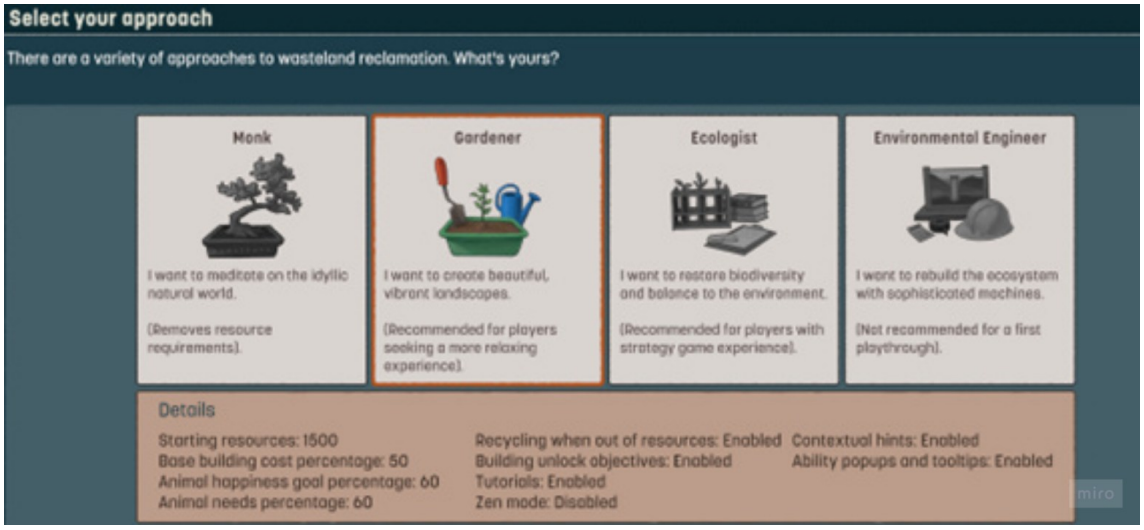


Figure 3. Terra Nil's difficulty settings.

This thesis defines a tutorial by combining aspects of Poretski and Tang's (2022, p. 6, p. 8-14) listing of design strategies of game learnability. Poretski and Tang (2022, p. 6) already define a tutorial as "explicitly teaching required functionality in well-defined stages using structured delivery of information", but the definition does not include implicit tutorials or other methods, which are listed their framework. For example, a sandbox level is defined as its own design strategy, although it can be viewed as a tutorial. As such, this thesis expands the definition of a tutorial to be a design strategy that explicitly or implicitly teaches the player relevant information in an appropriate context, applied to fit gameplay or narrative needs.

This definition is further validated by Andersen et al. (2012) article studying tutorials on games of varying complexity. In their article, they consider different aspects of tutorials, such as context sensitivity, freedom of playing, and availability of help (Andersen et al., 2012, p. 1-3). Whereas Poretski and Tang (2022, p. 6) define tutorials as explicitly structured learning materials that teach required knowledge, Andersen et al. (2012, p. 2) include varying context-dependencies that tutorials can have. For example, Poretski and Tang (2022, p. 6) outline Just-In-Time reminders as their own category, whilst by Andersen et al.'s (2012, p. 2) definition they could be included in the definition of a tutorial.

The games were first analysed using the tutorial impact assessment framework presented in Benvenuti et al. (2023, p. 5), which provides heuristics that can be used to evaluate different playability aspects such as mechanical playability. The tutorial impact assessment framework allows for certain traits that each game holds to be evaluated, and any possible effects of those traits on the data can be analysed further.

As a part of the experiment, the biometric data of the participants will be gathered for research purposes. This was done via eye tracking carried out during the gameplay session. Types of data gained is explained more in detail alongside other data in Table 2. The data was processed anonymously, with only respect to certain demographic

information i.e. age and prior knowledge level gathered prior to the experiment. The consent form for the participants is present in the appendices in Appendices 1 and 2. The data privacy form is present in Appendices 3 and 4. After the data was analysed, it was destroyed accordingly to University of Vaasa guidelines. Participants are also briefed on their informed and continuous consent, meaning that they reserve the right to withdraw their data at any point. No additional research permits were required, as the participants were not from protected groups, nor the data collected was of sensitive nature.

Table 2. Types of data gathered in the experimental design.

Demographic information	**Participant performance rating**	**Subjective workload assessment via Task-Load Index (NASA-TLX)**	**Working memory capacity assumptions via secondary task reaction time test (STRT)**	**Fixation data**	**Saccade data**
Age, Knowledge, Expertise	Time taken to complete stages	Weighted rating, statistically validated	Milliseconds taken to respond to a stressor	Fixation count, statistical validation	Saccade count, statistical validation

The types of data gathered in the experiment include both quantitative and qualitative data. Qualitative data includes the subjective workload assessment from NASA's task load index (TLX) survey, which can be then statistically analysed. Quantitative data includes participant performance in time taken to complete stages, STRT-test response times in milliseconds, eye tracking results in fixation counts as well as saccade counts. All of the quantitative data can be statistically validated. Participant demographics are also gathered.

1.3 Methods

Qualitative data on task load and learnability of the game is gained from an on-location survey after the experiment, conducted electronically. The data was electronic to better facilitate the analysis process of it without having to convert paper data into electronic formats. Survey at hand did not also necessitate a certain form of survey taking, so the most convenient one can be chosen. Eye tracking data was also gathered via devices available in the VME Interaction Design Environment. Exact types of data are explained further in chapter 5.

An experiment was performed to measure tutorial learnability and effects on user cognitive load. Learnability and cognitive load were measured quantitatively using a NASA-TLX -survey. NASA-TLX survey is a survey developed in the 1980s to measure qualitative workload of a person engaging in a task with a machine (NASA, 2022). The survey is commonly used in HCI research as well (e.g. Shankhwar & Smith, 2022). Part of the survey is shown in Figure 4. The procedure outlined for the survey includes familiarizing the participant with the survey instructions and implementing a few control runs while they tackle different tasks. Following the task of playing the game, the participants answer 15 pairwise comparisons, choosing a theme amongst the ones in Figure 4 that most applied to their task, i.e. choosing mental demand over physical demand. This gives a weighting for each theme, ranging from 0-5. The participants will then answer the questions shown in Figure 4. The scale presented is from 0-20. Both the pairwise comparison and the subjective ratings for each theme were presented to the participants electronically. Following the ratings, the TLX score was calculated using the responses.

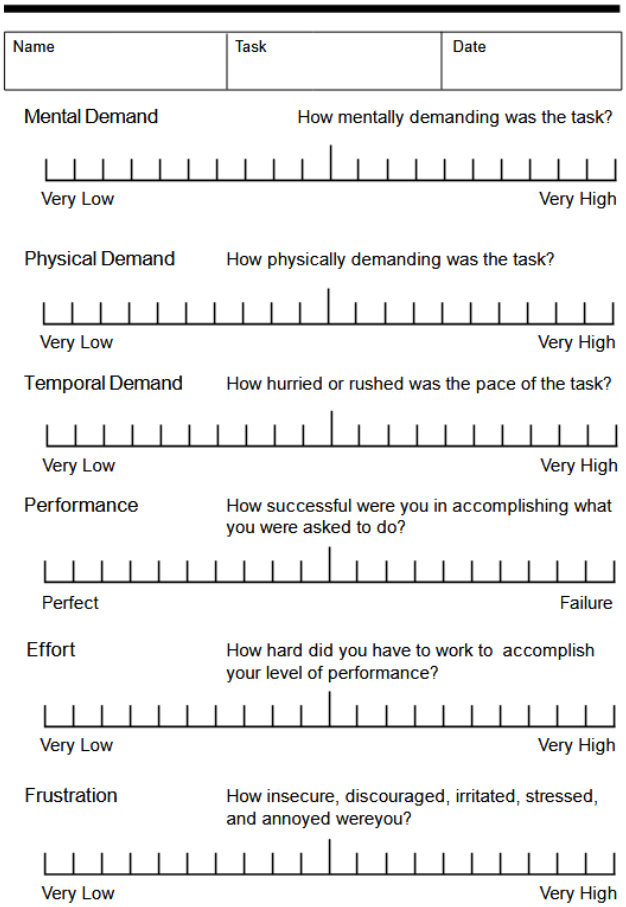


Figure 4. NASA Task Load Index.

Additionally, the games were evaluated using the heuristic list by Benvenuti et al. (2023), present in Table 3. This allows for categorization of the tutorial to assess its core traits and is not an evaluation that correlates with the quality of the tutorial. The list is described in more detail in chapter 5.

Cognitive load can be assessed using the NASA-TLX -survey, but as the data is based on participant self-assessment, it can be somewhat unreliable. The survey was thus combined with physiological measures, gathering eye tracking data in the form of fixations and saccades to note possible challenges the participants faced during the experiment. Task performance of the participants was also tracked via time taken to complete tutorial segments (Orru & Longo, 2019, p. 8). Sevcenko et al. (2023, p. 3-4) showcases a variety of factors analysable via eye tracking technologies that can

showcase cognitive functions of participants, such as fixation and saccade count. This framework was utilized to evaluate participant cognitive load in this thesis as well.

Statistical analysis was also conducted to further assess the results of the experiment. For the NASA-TLX -survey the data was evaluated with a paired t-test to assess the statistical validity of the findings. Post-hoc testing was not conducted, as this research is more considerate towards false negatives than false positives. Eye tracking data was also validated with statistical analysis, using either t-tests, ANOVA, regression analysis, or fitted lines wherever fit.

To ensure proper physiological measures, the participants were first briefed on the nature of the experiment once more and given time to get comfortable in the environment. A calibration test was run to validate the eye tracking device's reliability. Although the researcher was not to guide the participants in any way, small talk with either the participant or with the lab assistant was fostered if a participant seemed nervous to hopefully foster a more comfortable environment. This should aid in the ecological validity of experimental design.

Factors such as the psychological capabilities of participants could also be controlled by applying appropriate tests, including Lee and Heeter's (2017, p. 5) STRT-test to evaluate participants' working memory capacities. Estimates of participants' working memory capacity were obtained to observe possible differences the tutorial can bring to people with differing capacities and also acknowledge possible outlier participants and analyse the data set with and without them.

Additionally, the knowledge levels of the participants may affect how they experience the learnability of the tutorial and how the cognitive load is affected by tutorials. Morin et al. (2016, p. 6) found that newer players need tutorials and that the presence of them would indicate further playing. This effect was not found in the experienced players' group, as experienced players had similar scores with or without tutorials. Similarly, Lee

and Heeter (2017, p. 8) found that experienced players' available working memory did not correlate with improved results. As such, the knowledge levels of the participants will be recorded via a self-assessment, and the games allocated to each will aim to be as new as possible for them.

This thesis also proposes a few hypotheses:

H1. More difficult stages showcase a higher level of cognitive load on the participants

H2A. Group 1 showcases eye tracking markers consistent with lower levels of cognitive load when compared to Group 2

H2B. Group 2 showcases a higher subjective rating of workload than Group 1

The first hypothesis H1 posits that the more difficult stages showcase higher cognitive load markers such as completion time, fixation count, and saccade count (Sevcenko et al., 2023, p. 14-15;Seyderhelm & Blackmore, 2023, p. 22-23). H2 posits that the participants who performed better on the STRT-test showcase fixations and saccades consistent with a more optimized cognitive load. As the STRT-test measures the time taken for the participant to respond to secondary cues outside their main objective, the results appear as time taken to complete the secondary task. Performing better would thus manifest in a shorter time to complete said secondary task. Upon the STRT-test encountering issues, described more in detail in chapter 6.1, this hypothesis was broken down into H2A and H2B. H2A posits that Group 1 can properly optimize their cognitive load when encountering the tutorial, showcasing lower saccade counts (Sevcenko et al., 2023, p. 10). H2B posits that Group 2 shows higher subjective ratings of workload due to a strained cognitive load via not applying proper schemas or experiencing a more complex task itself (Sweller et al., 2019, p. 16;18-20; Chen, Kalyuga, Sweller, 2016).

2 Cognitive load, its components, and learnability in video games

2.1 Cognitive load and its components

Cognitive load refers to the effort required to perform cognitive tasks (Orru & Longo, 2019, p. 5). Cognitive load theory divides cognitive load into three main sections: intrinsic, extraneous, and germane. Intrinsic loads occur from element interactivity, which Sweller (1994, p. 9-10) argues to be inherent in the context itself, and thus cannot be affected by instructional design. Sweller et al. (2019, p. 4) furthers the definition as involving both the difficulty of the content to be learned, as well as knowledge level of the learner.

Extraneous load then again is the excess load that an individual must overcome after the intrinsic load, and which can be affected by instructional design (Orru & Longo, 2019, p. 6; Sweller, 1994, p. 9-10). Although original definitions of extraneous load does not include element interactivity (Sweller, 1994, p. 9-10), current research suggest that instructional design can also affect element interactivity (Sweller et al., 2019, p. 4; Orru & Longo, 2019, p. 6).

Germane load is defined as an excess load similarly to extraneous loads that is built off of effective instructional design, but that aids in the learning process of a person and to develop schemas (Orru & Longo, 2019, p. 7). It does not necessarily contribute to the total load but rather makes the current use of cognitive functions more efficient by i.e. streamlining working memory usage (Sweller et al., 2019, p. 4).

Sweller et al. (2019, p. 16-20) note several cognitive load effects found in current research, dividing them into compound and non-compound effects. Element interactivity is a compound cognitive load effect that consists of cognitive complexity of material (Sweller et al., 2019, p. 16; Chen, Kalyuga, Sweller, 2016). For example, the

element interactivity of the following equation $y^2 = x + \frac{1}{3}$ can be seen as relatively high, as the learner would have to not only know each symbol provided, but also how to process and interpret it. We can lower the element interactivity of the previous equation by writing it instead as $y \times y = 3 + \frac{1}{3}$.

Other compound cognitive load effects include the expertise reversal effect, guidance-facing effect, transient information effect, and self-management effect (Sweller et al., 2019, p. 17). Expertise reversal effect functions where shifts in learner knowledge and expertise change the effects of element interactivity on cognitive load (Sweller et al., 2019, p. 17; Chen, Kalyuga, Sweller, 2017). For example, most of us should be able to process the previous equations provided, as knowledge on how to do so has already entered our long-term memory, and we have sufficient experience in working them. The effect might be even reversed as the naming would point to, with students currently studying and working with such elements getting benefits from such descriptions.

The guidance-facing effect considers the need for guidance at the beginning stage of learning, which becomes somewhat redundant and possibly detrimental to cognitive load later in the learning process (Sweller et al., 2019, p. 17; van Merriënboer & Kirschner, 2018). For instance, when a player begins playing a totally new game, tutorials are necessary to aid in the learning process and aid player cognitive load by breaking down mechanics and gameplay into more digestible chunks, or by limiting certain actions until the player has sufficient experience using earlier knowledge. If a player encounters similarly presented aid later on, they may be especially burdened by an abrupt stop to the natural learning flow one would have developed throughout the gameplay process. The completion strategy presented in Sweller et al. (2019, p. 17) is thus somewhat similar to the approach games take, with games starting off by heavily guiding the new players and transitioning to less and less handholding.

The transient information effect occurs when information present is in a non-transient format, requiring active learner retention during processing (Sweller et al., 2019, p. 17).

Strategies presented in Sweller et al. (2019, p. 17-18) include segmenting the content or using multi-modal medium for short-form transient content (Leahy & Sweller, 2016). We can observe both effects with most transient video game tutorials; they segment the learning from basic to more complex and are inherently a multimodal medium. The self-management effect considers that the learner can apply cognitive load management tools to learning material to aid their cognitive load, despite lack of design for better cognitive load (Sweller et al., 2019, p. 17).

Sweller et al. (2019, p. 18-20) present five non-compound cognitive load effects found in current literature; self-explanation effect, imagination effect, isolated elements effect, collective working memory effect, and human movement effect. This thesis will choose to skip over the collective working memory effect, as it does not fit into the context of the thesis. The self-explanation effect occurs when the learner is trying to process information through an internal monologue of sorts, when the learning is within reasonable cognitive load boundaries (Sweller et al., 2019, p. 19; Renkl et al., 1998). The imagination effect considers the effects of imagining the matter being learnt, creating a system of mental understanding between different pieces the material may consist of (Sweller et al., 2019, p. 20). The imagination effect seems to occur mainly with learners who are already familiarized with the matter being learnt, as a new learner may need to first understand the concepts it consists of before imaging and practicing the full image (Cooper et al., 2001; Sweller et al., 2019, p. 20).

The isolated elements effect occurs when elements of learning material are isolated into different blocks of learning, after which the learner can integrate the interactions between them into existing schemas (Sweller et al., 2019, p. 20; Pollock, Chandler, Sweller, 2002). Complex video games employ tactics that utilize this effect quite often; strategy games often begin by explaining basic mechanics like stats and gameplay like combat as distinct pieces, then either continue by teaching the interplay between them or letting the player learn them. Lastly, the human movement effect considers that learning is more optimal when visualizations are represented as moving and accurate

animations or the likes rather than stiff images (Sweller et al., 2019, p. 20; Höffler, Leutner, 2007). This finding was present in Höffler and Leutner's (2007) meta-analysis in motor-skill learning contexts, suggesting that having life-like and imitable presentations may improve germane load.

Cognitive load can also be affected by the format of instructional content, creating either extraneous or germane loads. For instance, Chandler and Sweller (1991, p. 5) examine studies around worked examples and more conventional instructional content, where worked examples were found to improve learning due to heightened focus on specific learning areas and removing extraneous loads that would go into solving problems. Alexiou and Schippers (2018, p. 4-5) note some effects of content complexity and structure can have on cognitive load. Structureless and complex content has been shown to lead to both higher cognitive load and feelings of frustration (Alexiou & Schippers, 2018, p. 4-5). Sevcenko et al. (2023) also found in their study that eye tracking data was able to predict intrinsic load of the players based on performance. Their data showed a negative correlation between saccades and subjective ratings, and that the number of fixations was lower on challenging levels, except for the successful participants (Sevcenko et al., 2023).

A related concept to cognitive load is flow, a state of cognition where concentration is high and performance relatively effortless (Csikszentmihalyi, 1975). Montani et al. (2020) found in their study that workers experienced the highest levels of engagement and innovation when workload was medium. This would suggest that achieving a flow state can be built off of a reasonable cognitive load that encourages learning. Additionally, Morin et al. (2016, p. 5) found that tutorials could aid specifically in casual player contexts to build a flow state, although more experienced players showed no effect with or without a tutorial.

Kiili (2005, p. 9) considers the effects of games as multimedia content on cognitive load, and how having different forms of media in the same learning context can create

extraneous load by limiting working memory functions. However, visualization via more life-like animations seems to have an increased germane load as opposed to just text (Höffler & Leutner, 2007; Sweller et al., 2019). Kiili (2005, p. 9) also suggests using haptic feedback to ease extraneous loads via higher levels of immersion.

2.2 Learnability

Human-Computer Interaction (HCI) is a field that studies the relationship between humans and computers and how we can improve it. User experience (UX) is one common area of study in HCI research, which can be seen as encompassing usability as well (Sanchez et al., 2012, p. 1). In video game research, however, the concept of playability is often used instead of UX and usability, as games by their nature are interactive and subjective multimedia products (Sanchez et al., 2012, p. 3).

Learnability as a concept exists both in HCI and playability research. In HCI it encompasses teaching a new user how to use a system and reach a desired level of knowledge (Pretorius et al., 2010, p. 2). In playability research, learnability has similarly been defined as the players' "capacity to understand and master the game's system and mechanics" (Sanchez et al., 2012, p. 7). Although learnability is usually considered from a new user's standpoint, learnability research in games has also studied effects on more experienced players (Poretski & Tang, 2022).

When it comes to learnability research from an HCI standpoint regarding games, Andersen et al. (2012) examined four different tutorial characteristics, presence of tutorials, context-sensitivity, freedom and availability of help for different games to find their effects on the learnability of games. The study found that the presence of tutorials as well as context-sensitivity increased engagement only in the most complex game (Andersen et al., 2012, p. 8). This could be due to the complex game warranting a tutorial to fully grasp it.

Lee et al. (2024) compared some ways that tutorial presentation could affect the learnability of games. They found that for learning game mechanics there was a large gap between having a tutorial and having none, as well as that context-sensitive tutorials were more favourable than on-screen tutorials (Lee et al., 2024, p. 10). Similar effects were identified with motivation and learning effectiveness specifically, although differences between on-screen tutorials and context sensitive tutorials were lesser, albeit remaining statistically significant in the whole (Lee et al., 2024, p. 10). These findings would suggest that differences in learnability and general user experience in tutorial types may not be as large, but still present.

Having no tutorials has been shown to be not favoured by users due to lacking understanding in game functions or reduced immersion (Lee et al., 2024). Andersen et al. (2012, p. 9) also showed that freedom within the tutorial had no effect on player behaviour, and that on-demand help had varying effects on player engagement and retention. Andersen et al. (2012, p. 9) posits that as tutorials had significant positive effects on only the complex games, effort may be needed to be placed in integrating tutorial into game design in a more general way, such as by allowing the players to experiment and discover functions via gameplay.

Blaskovic et al. (2023) presented a framework via which the effects of several concepts such as player engagement and learnability could be evaluated in a gameplay setting. Gameplay mechanics was identified via a statistically validated analysis as a moderately relevant factor in affecting learnability, as well as that user interface sensibility had a weak relationship with learnability (Blaskovic et al., 2023, p. 18). Gameplay mechanics were also identified as moderately related to player enjoyment (Blaskovic et al., p. 17). These findings would suggest that core gameplay features and mechanics can influence game learnability to an extent, and by affecting player enjoyment also further affect learnability. Additional factors that can affect player enjoyment include engagement, challenge, success, self-expression and more (Korhonen et al., 2009).

Eye tracking technologies have been used to measure learnability among other usability factors in HCI research. Eye tracking uses an optical device paired with software to record data around eye movements and patterns that occur from them. The advantage of eye tracking with learnability research is that it does not place an unnecessary burden on the participants' cognitive functions, allowing for more natural data (Pretorius et al., 2010, p. 3).

Several game design choices can affect learnability to a significant degree. Blaskovic et al. (2023, p. 15) found that user interface (UI) and gameplay mechanics affected learnability, with UI having minor effects on learnability and mechanics a moderate effect. It is to be noted that their study examined these effects in a specific platformer-game context. Meanwhile Poretski and Tang (2022) focus on approachability and accessibility of a game as factors that increase learnability. The definitions of the two concepts vary, but in their paper, they focus on the initial gameplay experience of both new and more experienced players (Poretski & Tang, 2022, p. 2-3).

Poretski and Tang (2022, p. 5-6) note some design choices used regarding learnability in their study's sample. Some of the strategies employed include tutorials before the start of the game, contextual help, and controlled practice in a sandbox environment (Poretski & Tang, 2022, p. 6). Meanwhile Gee (2005, p. 2-11) divides design choices that facilitate learning into three main categories: understanding, problem solving, and empowering players as learners. For example, empowering players as learners built upon making them active parts of the process, allowing them to adapt their experiences to their own capabilities, creating a sense of engagement with the learning material, and by having them take actions that have distinct and observable feedback (Gee, 2005, p. 2-5).

3 Cognitive load theory and learnability in video game tutorials

Although somewhat scarce, research exists on the links between cognitive load and game learnability. Lee and Heeter (2017) examined the effects of cognitive capabilities and gaming knowledge levels on the cognitive functions of attention and comprehension. They found that working memory capacity could not predict comprehension of the media for the expert knowledge group, but could for the non-experts (Lee & Heeter, 2017, p. 8). The findings were then studied further in a second experiment, where Lee and Heeter (2017, p. 10) found that the effect was caused by a misallocation of their cognitive functions in the expert knowledge group. They thus applied their existing schemas to the tasks, which caused blindness to factors outside the schema. The findings would suggest that games hold a very strong set of schemas which experienced players grab onto, which may cause a break in comprehension and engagement with the game if it deviates too much.

Related to schemas, Pretorius et al. (2010) studied the differences between adults and children learning to play a game. Adults can be considered more knowledgeable users that have not only encountered more learning situations, but that also may have varied experiences playing games or engaging in similar activities. Pretorious et al. (2010, p. 5-7) found that adults focused more on the instructions and other guiding text when playing alone or being guided by a more experienced player. This would seem to suggest that children are generally more likely to engage directly with the systems and learn from active participation.

Chang et al. (2017) looked at game-based learning and its effects on cognitive loads. They compared a control group using traditional learning material with a group that used game-based learning material and found that the game-based learning material affected extraneous loads positively albeit to a lesser extent, and germane loads positively to a

greater extent (Chang et al., 2017, p. 16). The findings showed no difference with intrinsic loads, which should not be affected by structure or form of the material.

One major aspect of game-based cognitive load is extraneous load created by unnecessary or complex details and presentation. Reducing these factors by limiting them can aid in reducing the extraneous load (Seyderhelm & Blackmore, 2023, p. 23).

When discussing the learnability of games, other aspects that can affect player cognitive load and overall cognitive functions rise. For example, engagement and immersion can affect the willingness to engage and keep up with the learning process (Andersen et al., 2012). Alexiou and Schippers (2018, p. 12) discuss how appropriate difficulty and reasonable and timely feedback can affect not only player engagement through immersion into the task but also their motivation to continue playing and learning the game's systems. This is done by giving the player agency to make mistakes and learn from them, whilst not letting them feel helpless or unskilled (Alexiou & Schippers, 2018, p. 12).

Another factor that can affect game learnability through engagement and immersion is tutorial design. Cao and Liu (2022) studied the use of implicit and explicit tutorials with game learnability. They found that implicit tutorials were considered more enjoyable and engaging, although less helpful as well, and that effects on learnability are weaker in expert players, but enjoyment factors carry through (Cao & Liu, 2022, p. 5-6). The findings suggest that implicit tutorials do little in improving learnability but can affect emotional aspects of learning.

Tutorial design in games can not only vary by the content within the tutorial and the way it is presented to the players, but also temporal factors. Chen et al. (2024) studied the effectiveness of tutorials of varying temporalities within a virtual reality -game context. Temporalities in tutorials can include real-time learning, bullet time, paused learning moments and such. They found that a bullet time tutorial was the most effective when

it came to teaching the players controls and mechanics, as well as having the lowest cognitive load ratings amongst other temporalities (Chen et al., 2024, p. 14-16). A bullet time tutorial usually consists of slowing down time to a significant degree, allowing for the player to process the scene for a longer period of time. There were notable effects on player engagement and increased feelings of autonomy alongside context sensitive tutorials, as well as showing slightly better ratings in the more complex game (Chen et al., 2024, p. 14-15). It is to be noted that VR games may inherently favour bullet time tutorials, as the level of immersion is already higher due to the setting, which may cause a more moderate workload on the player. This in turn can support real-time learning outcomes due to a more optimal cognitive load, especially when paired with a timelier but engaging and autonomy-enhancing nature of bullet times.

Several design choices aimed at improving game learnability have also been shown in research by Poretski and Tang (2022). Poretski and Tang (2022, p. 6) identify the following design strategies: recaps, seeding in the cutscene, assessing prior knowledge, tutorials, the invisible hand, practice in a sandbox, the sixth sense, just-in-time reminder, personal advisor, debriefing, and documentation. For example, practice in sandbox occurs when a game presents an area to the player specifically for experimentation and practice of the game mechanics. This classification is divided into before, during, and after gameplay. Poretski and Tang (2022, p. 14-18) further classify these design strategies into accompanying user interface (UI) integration via which they are used in games: purpose, format, presentation, trigger, constraints, and repetitiveness. For example, constraint can occur when capabilities are limited to correspond with current learning, such as by not allowing for jumping before the mechanics around jumping are explained.

Although the challenge-level of a task can affect the cognitive load of players, the link between it is challenged. Seyderhelm and Blackmore (2023, p. 22) found in their study that cognitive load and relative difficulty of a level was inconsistent, with one level showing high cognitive load and low performance, whilst another level showed high levels of both. Another aspect privy to improving players' cognitive load is germane load.

Seyderhelm and Blackmore (2023, p. 22) found in their study that adjusting difficulty level of a task could induce a flow state, which can affect both extraneous load and germane load.

4 Explanation of the experimental design

The experiment included a NASA-TLX survey, an STRT test, a heuristic evaluation of the tutorials, an eye tracking section, an evaluation of player performance, and a statistical analysis to validate the results. The experiment involved having the participants play through a tutorial of one of the games for 15-30 minutes. The participants were asked if they had prior knowledge about any of the games or genres the games belong to, and they were assigned a game that they had the least previous experience with. If there were several options available for a participant, they were assigned a game that had the least number of participants assigned to. This was done to ensure a somewhat equal number of participants for each game.

Before beginning to play the game, the participants were first briefed with the nature of the experiment. They were then given information about the NASA-TLX survey. During the actual experiment, the participants would play the game assigned to them in the eye tracking area. The area was the University of Vaasa's HCI-laboratory VME Interaction Design Environment, where eye tracking and other biometric studies can be carried out. The device used in the study was the Tobii Pro Nano -eye tracker. The peripherals used were standard keyboard and mice that one can find in an office, including the 27 inch monitor the game was played on. The participants could adjust table height, monitor height and tilt, increase the backlighting of the study area, as well as peripheral positions. The analysis tool used for the eye-tracking data was the iMotions-application (2024, ver. 10.1). The gaze data taken was based on Sevcenko et al.'s (2023) study, tracking fixation frequency and saccades.

Participant performance during gameplay was also assessed. For Baba Is You, performance was measured in terms of time taken to successfully complete each puzzle. Failures were not measured, as although they could occur due to the player soft-locking themselves, the occurrence of these soft-lock possibilities between stages was not consistent enough to allow for statistical analysis. For Terra Nil, performance was measured in the time taken to complete each stage of the tutorial. Failure- or success

rates were not able to be tracked, as the tutorial would not allow for mistakes. After completing the game session, the participants took the NASA-TLX survey.

The STRT test as used by Lee and Heeter (2017, p. 5) measures the time it takes for a participant to react to a secondary task whilst undergoing a primary task. In this experiment, during the gameplay section the participant would receive an audio cue, after which they would press the spacebar similarly as in Lee and Heeters' (2017) study, as the spacebar is not used for any of the games. There would be three audio cues per participant, from which an average response time could be estimated.

The heuristic evaluation used in the thesis is the tutorial impact assessment framework by Benvenuti et al. (2023). The framework combines Andersen et al. (2012) framework with items by Benvenuti et al. (2023, p. 5) to fit more modern gaming contexts. By running the games through this framework, we can evaluate the quantity and quality of tutorial features each game has. The framework is presented in Table 3.

Table 3. Tutorial impact framework.

Item	Description
1. Tutorial presence	True if the game under analysis provides some in-game tutorial
2. Context sensitivity	True if the instructions for a certain action or game mechanic are shown to the players only when they really need to use them during gameplay
3. Freedom	True if the players are provided with some freedom during the tutorial, e.g., they can make choices in the game
4. Availability of help	True if, during the tutorial, the game understands when the player is in need of help and reacts to that through textual, visual, or graphical cues
5. Printed	True if the game provides printed documentation with the game instructions
6. On-screen text	True if the game delivers instructions to the players through text
7. Voice	True if the game delivers instructions through voice
8. Video	True if the game explains mechanics or commands employing dedicated videos, including cut scenes
9. Helping avatar	True if an in-game assistant supports the player. In-game assistants range from a non-playable character (NPC) to fictitious characters speaking to the player through a phone call, etc. We notice that a helping avatar is considered part of the game world; i.e., it is not simply a voice or a text providing specific instructions
10. Controller diagram	True if the game shows the players (on request) an image describing how commands are mapped to the input device
11. Command scheme customization	True if the game lets the player change the command scheme mapping
12. Skippable	True if the tutorial can be completely skipped
13. Story integration	True if the tutorial is integrated into the game's story
14. Practice tool presence	True if the players are provided with a safe place to practice and get used to commands and game mechanics

Benvenuti et al's (2023, p. 5) framework includes several different aspects of tutorials that can aid in learnability of the game. What one might commonly experience in game are items such as context sensitive tutorials that only appear when needed, i.e. press x to jump when a player first encounters an obstacle that needs jumping, or press x twice to double jump when encountering a higher obstacle. Some games also include printed

documentation to help a player grasp basic mechanics before starting the game, or for them to go back to whenever needed, although in the thesis' games this is not applicable, as the games are digital. Instead, printed is interpreted as the game offering separate written instructions, such as an instruction manual inside a game case.

Table 4. Tutorial impact framework applied.

Trait	**Baba Is You**	**Terra Nil**
Tutorial presence	X (minimal guidance)	X
Context sensitivity		X
Freedom	X	X (limited to given choices, i.e. buildings unlocked at current stage)
Availability of help		
Printed		X (limited information through a glossary, no explicit instructions)
On-screen text	X	X
Voice		
Video		
Helping avatar		
Controller diagram	X	X
Command scheme customization	X	X
Skippable	X (implicit tutorial)	
Story integration	X (implicit tutorial)	X
Practice tool presence		X

The framework of Benvenuti et al. (2023, p. 5) was applied to the games chosen for this experiment. Based on this, an evaluation on the games' tutorials was created, found in Table 4. Each item that holds true to the game is marked by an X. The games share a lot of similarities in their tutorial design, although Terra Nil is more explicit. Terra Nil employs context sensitive tutorials that pop up when the player unlocks a new feature, showing what they need to do with it. To some extent we can argue that the stages in Baba Is You

are context sensitive tutorials themselves, as they usually present a new novel way to manipulate the sentences for a solution. However, as a lot of the stages have multiple solutions, we do not call this context sensitive in the context of this thesis. As Baba Is You's tutorial design is baked into the levels themselves, they can also be classified as skippable, meaning that a skilled player is able to use their pre-existing schema and quickly solve the stages. Terra Nil also incorporates a small glossary that a player can access whenever they want, meaning that printed instructions exist, albeit in a lesser extent.

5 Analysis and results

5.1 Integrity of the dataset and participant performance evaluations

During the first analysis phase of the full data set, STRT-values of each participant were written down for statistical analysis. However, only eight out of 27 participants had audio recording present in their research files, possibly due to a technical error either before or after the data collection. However, crass subjective evaluations of performance on the STRT-test were made and recorded, which allowed for the participants to be grouped into low and high capacity working memory groups. This allocation allowed participants to be evaluated based on a more subjective working memory evaluation scale, meaning that their data could still be statistically analysed to observe possible differences. It is to be noted, however, that this leaves some room for researcher bias. The existing STRT-responses are given in Table 5 in milliseconds. The groupings created are listed in Table 6. Members of Group 1 in Table 5 are marked with an asterisk (*).

Table 5. Successful STRT-test results.

Participant	Attempt 1	Attempt 2	Attempt 3	Average
P1*	518	543	362	474
P2*	196	348	778	441
P3*	666	228	345	413
P4	284	401	288	324
P5*	964	101	403	489
P6*	219	398	347	321
P8*	649	674	993	772
P9*	417	405	507	443

Participant performance ratings were calculated based on time taken in seconds to successfully complete a stage. The exact timing was measured via frame-by-frame analysis of the eye tracking video. Start time for Baba Is You was based on the frame in

which the full stage had loaded onto the screen. The end time was chosen as when the full “Congratulations” -text prompt had loaded onto the screen. For Terra Nil, the start time was similar to Baba Is You taken to be the frame during which all of the user-interface (UI) elements had fully loaded in. End time was based on the frame in which these elements had disappeared. During 2 stages in Terra Nil, this UI-element disappearance happens only after the guidebook has spawned onto the screen, so for those stages the first frame that the guidebook appears in was chosen instead.

In Table 5 we can observe that the results show some variance. Some participants had consistent results across all three attempts, whilst others experienced a raising or lowering score with each subsequent attempt. The average response times show less variation outside of participants P4, P6 and P7.

Table 6. Analysis groups based on working memory evaluation.

Baba Is You Group 1	Baba Is You Group 2	Terra Nil Group 1	Terra Nil Group 2
P1	P7	P3	P12
P2	P16	P5	P14
P10	P17	P6	P15
P11	P20	P8	P19
P13	P21	P9	P27
P18	P22	-	-
P25	P26	-	-

Group 1 in both games includes the participants with successful STRT-results and with high working-memory capacities as shown in (Lee and Heeter, 2017, p. 5). In Group 2 are placed the participants with low or moderate working memory capacities, as well as ones with insufficient notes to properly allocate them.

Performance data is listed for each stage, listed in the first row, in Table 7 for Baba Is You and Table 8 for Terra Nil, both games in seconds. For both games Group 1 is denoted with an asterisk (*). For stages that a participant did not complete, their performance is noted as not applicable (N/A). No participant managed to complete stage four on Terra Nil, although participants P6, P8, P9 and P19 got to the stage. Average time taken to complete a stage is present in the second column from right.

Table 7. Baba Is You performance data.

	0	1	2	3	4	5	6	7	Avg.	Total
P1*	19,45	94,10	57,48	64,82	61,61	221,54	35,37	278,36	104,09	832,73
P2*	9,30	52,17	40,84	100,36	135,84	N/A	142,40	287,40	109,76	768,31
P7	52,22	496,48	N/A	N/A	N/A	N/A	N/A	N/A	274,35	548,70
P10*	5,40	35,38	17,74	25,36	30,31	213,27	32,97	402,13	95,32	762,56
P11*	3,48	15,10	48,46	15,36	21,23	268,06	35,51	99,38	63,32	506,58
P13*	5,90	191,60	56,53	44,87	159,15	N/A	98,51	104,67	94,46	661,23
P16	11,27	124,94	550,35	339,55	103,97	156,32	N/A	N/A	214,40	1286,40
P17	36,98	244,89	255,73	390,86	262,55	242,47	N/A	N/A	238,91	1433,48
P18*	9,84	18,45	13,72	34,73	64,29	62,19	25,13	111,78	42,52	340,13
P20	17,61	91,58	27,14	N/A	215,96	319,11	N/A	N/A	134,28	671,40
P21	24,20	246,52	61,51	222,96	111,49	N/A	N/A	N/A	133,34	666,68
P22	12,26	209,50	140,90	242,18	91,36	N/A	163,54	354,18	173,42	1213,92
P25*	18,85	259,77	25,99	88,31	58,39	94,25	41,17	N/A	83,82	586,73
P26	79,88	527,79	N/A	N/A	N/A	N/A	N/A	N/A	303,84	607,67

Table 8. Terra Nil performance data.

Stage	1	2	3	Average	Total
P3*	541,82	1031,83	N/A	786,83	1573,65
P5*	153,18	N/A	N/A	153,18	153,18
P6*	306,69	459,82	318,92	361,81	1085,43
P8*	196,41	540,33	654,72	463,82	1391,46
P9*	320,54	524,66	467,84	437,68	1313,04
P12	711,97	N/A	N/A	711,97	711,97
P14	450,29	666,79	N/A	558,54	1117,08
P15	473,00	600,25	N/A	536,63	1073,25
P19	350,32	675,91	312,73	446,32	1338,96
P27	623,55	N/A	N/A	623,55	623,55

For the statistical analysis of performance, a subjective difficulty rating for each stage was calculated from the times given in Tables 7 and 8. This was divided by the total time taken for the whole gameplay session without including failed stages where the participants ran out of time, as this could affect the validity of the data. Total time is shown on Tables 7 and 8 on the right-most column. As Terra Nil had only five people per group with varying completion rates, regression analysis could not be completed for that dataset. After getting a subjective difficulty rating per stage per participant, these values were put into a regression analysis on Baba Is You's dataset. Alongside the difficulty rating for the stage, the participants number of fixations and saccades per minute during that stage were inputted into the regression analysis. Analysis of the eye tracking data is explained in detail in chapter 6.3.

Fixation and saccades for the most part hold a similar linear relationship. Exceptions to this occur in stages 0 and 1, where outlier saccades create an inverse linear relationship

with fixations, shown in Figures 5 and 6. This contrasts with Sevcenko et al.'s (2023, p. 9-11) findings where the associations between difficulty and physiological measures were similar in direction and scale. This may be due to Baba Is You requiring a higher level of non-visual thinking compared than Sevcenko et al.'s serious game, as Baba Is You utilizes sentences as a core gameplay feature (Sevcenko et al., 2023, p. 4).

There is a weak inverse linear relationship between fixations and saccades in stages 5 and 6, but one that is statistically insignificant. It is to be noted that these stages rank in the three most difficult ones. All other figures are available in Appendices 7-12. By observing the whole data set instead of by stages and treating each stage of a participant as an independent variable, we can observe some interesting results. Although the linear relationship is maintained, we can observe a higher degree of freedom in the clusters of the saccades. Justification for this attempt at normalization is weak at best, but left here as a curiosity, presented in Figure 7.

Table 9. ANOVA values on fixations and saccades per stage.

Stage	p-value (Fixations)	p-value (Saccades)
0	0,514	0,442
1	0,146	0,439
2	0,242	0,258
3	**0,022***	0,485
4	0,665	0,450
5	0,624	0,673
6	0,831	0,245
7	0,380	0,397

When comparing the p-values gained through ANOVA, present in Table 9 of regression analysis on fixations and saccades per stage, we see a similar result of lack of statistical significance. However, the number of fixations on stage three shows a statistically significant result of p=0,022. This may be due to the stage increasing the cognitive load

of participants, shown by a reduction in fixation count and rise in saccade rate (Chen et al., 2011, p. 3; Sevcenko et al., 2023, p. 3-4). This would suggest that the learnability in the context of the stage was compromised, leading to players having to leverage more of their cognitive capacity on solving the puzzle. The effect could have occurred due to the stage introducing a mechanic to function differently than in previous stages, forcing the players to adapt their mental schemas.

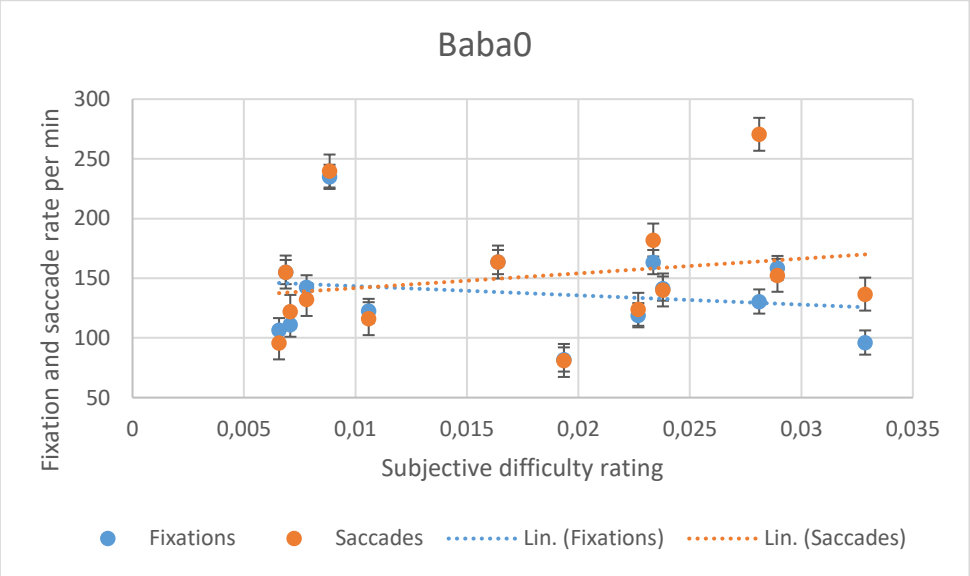


Figure 5. Stage zero linear regression.

Figure 5 shows the subjective difficulty rating of a stage plotted against the fixation and saccade rate per minute, with fixations shown in blue and saccades in orange. The dotted values are plotted into a linear regression line that shows us the linear relationship between the individual dots. In the case of saccades, the value is clearly rising, showing an increased saccadic rate with an increase in subjective difficulty. However, fixations show a decreasing line, showing the opposite effect.

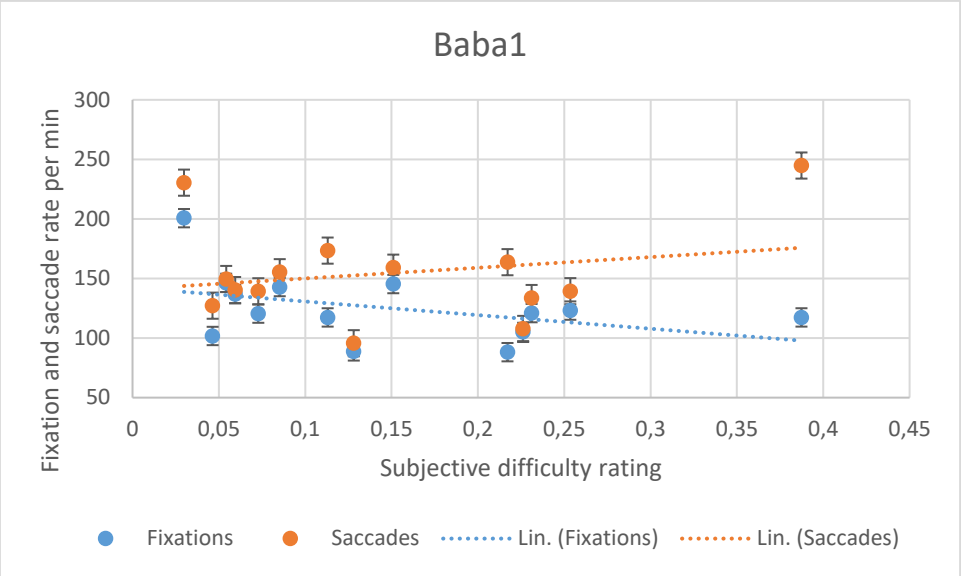


Figure 6. Stage one linear regression.

Figure 6 shows a similar trend as to Figure 5, albeit with a much stronger trend. The saccadic rate present is of a similar gravity, moving from a rate around 140 to 180, starting slightly higher than in Figure 5. The fixations show a much larger downwards trend, dropping as far as under 100 in the most difficult ratings. The effect can be explained to an extent by the fact that stage zero was quite easy to complete, with some

participants seemingly doing so accidentally. Stage one served as a litmus test, showing whether the participants had understood the core game mechanics.

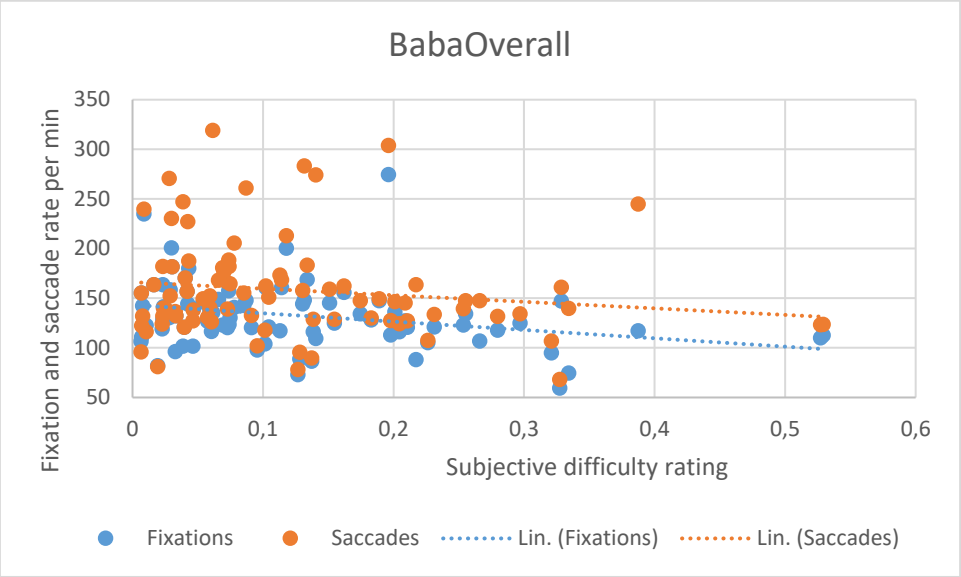


Figure 7. All stages linear regression.

In Figure 7 we can observe that the saccades occupy a wide range of values within the graph, whilst the fixations appear more clumped together and consistent. The reasons for this at this point are unknown, although the general linear relationship between the two is maintained.

5.2 NASA-TLX -findings

The NASA-TLX ratings on each item were plotted down per participant, alongside with item weight gained from the pairwise comparison. The sum of the pairwise comparison was double-checked to equal 15, as advised in the manual (NASA, 2022). The weighted ratings were calculated by multiplying the rating with the weight. Overall weighted ratings for the whole gameplay session were calculated by adding the weighted ratings of each of the 6 items together and dividing the sum by 15. The results of the weighted

ratings are presented in Table 10, with the overall weighted rating found in the right-most column. Group 1 is noted by an asterisk (*).

Table 10. NASA-TLX Weighted Ratings**.**

	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration	**Weighted rating**
P1*	28	2	0	30	28	6	**6,3**
P2*	14	1	0	24	40	40	**7,9**
P3*	48	0	8	39	20	3	**7,9**
P4	12	2	2	32	30	4	**5,5**
P5*	40	3	0	24	21	2	**6,0**
P6*	48	5	0	25	24	12	**7,6**
P7	52	0	26	25	27	4	**8,9**
P8*	60	0	6	39	6	70	**12,1**
P9*	52	0	18	36	5	80	**12,7**
P10*	40	0	7	42	20	20	**8,6**
P11*	36	0	5	28	24	36	**8,6**
P12	48	0	9	24	48	24	**10,2**
P13*	33	0	21	14	60	22	**10,0**
P14	2	1	3	25	15	1	**3,1**
P15	85	2	2	7	26	64	**12,4**
P16	22	0	30	40	36	16	**9,6**
P17	36	9	24	11	39	56	**11,7**
P18*	12	3	33	10	42	0	**6,7**
P19	30	2	4	15	12	6	**4,6**
P20	10	0	18	44	24	85	**12,1**
P21	95	0	30	17	51	76	**17,9**
P22	55	0	13	26	33	68	**13,0**
P23	24	0	6	12	32	60	**8,9**
P24	54	0	85	12	16	56	**14,9**
P25*	75	0	2	20	42	18	**10,5**
P26	27	0	54	16	36	80	**14,2**
P27	39	0	40	7	24	50	**10,7**
Overall avg.	40	1	17	24	29	36	**9,7**

The results would indicate that the participants rated the overall workload of the gameplay session as lower end of moderate with a value of 9,7 out of the 40 that the overall score can be. Out of the individual items, mental demand and frustration were the highest at 40 and 36 respectively, with the score capping at 100. The values are thus quite moderate. Mental demand has a varied rating amongst the participants, with ratings ranging from as low as 2 to as high as 95. Frustration has quite many of these lower values. These results are consistent with the types of games chosen for the experiment, as they are both strategy games that can require a lot of learning and cause frustration, especially in Baba Is You. Although frustration can negatively affect learnability by demotivating the learner, this could not be observed in the experimental design. Frustration could also have the effect of causing the learner to not pay proper attention to instructions or to absorb them properly.

The lowest value from the survey average ended up being physical demand with an average of 1 out of 100, which is caused by the large majority of answers to this rating being 0 on the pairwise comparison. In the actual ratings, most evaluated physical demand still as either the lowest at 1 or somewhat low as 2 or 3 out of 20. This can be explained by the experimental design specifically attempting to minimize possible physical demands of the games to ensure that findings would limit to the participants' cognitive load.

With the groups presented in Table 6, the overall weighted ratings were analysed per game via a t-test. For Baba Is You we get a p-value of 0,005, meaning that the difference between the two groups' overall workload is statistically significant. Thus, hypothesis H2B is acceptable. This can be explained by the more expert group of Group 1 having a pre-formed mental schema that helped them in allocating their working memory resources more effectively (Lee, Heeter, 2017, p. 8-10). It could also be that Group 1 had entered some kind of a flow state as theorised by Seyderhelm and Blackmore (2023, p. 22), or that Group 2 could not gain the benefits of an implicit tutorial as outlined by Cao and Liu (2022). We cannot discount the possibility that the tutorial did not keep the

engagement levels of Group 2 at an appropriate level for learning and managing cognitive load (Alexiou & Schippers, 2018, p. 12).

For Terra Nil a p-value of 0,327 was observed, which is not statistically significant. This can be explained by Terra Nil having participants that had uncertain STRT-results being placed into Group 2, possibly causing a lack of variance between the two groups. This effect is enforced by the smaller size of the two groups as well. Finally, both Groups 1 from Baba Is You and Terra Nil were put together to compare against Groups 2. This resulted in a p-value of 0,076, meaning that the difference between the groups across games was not statistically significant.

5.3 Findings from eye tracking

As Pretorius et al. (2010, p. 4) notes, analysing interactive elements that change with each participant is a complex process. As such, the analysis of the eye tracking was done by creating areas of interest (AOI) within the frames outlined in chapter 6.1. Although this does not allow us to create heatmaps due to interactive elements affecting the reliability of analysis, such as moving player character, camera, or other elements, we can still utilize fixation and saccade counts within the AOI in a similar fashion as Sevcenko et al. (2023). The fixation and saccade counts for Baba Is You are listed in Tables 11 and 12 respectively, whilst the counts for Terra Nil are listed in Tables 13 and 14. In all of the tables members of Group 1 are marked with an asterisk (*). It is to be noted that iMotions can give half values for fixation counts, for which the reason is unknown. These values were few but were rounded up for clarity's sake.

Table 11. Fixation counts for Baba Is You.

	0	1	2	3	4	5	6	7	Total	**Per min**
P1*	53	184	118	152	127	395	85	347	1461	**105**
P2*	19	119	95	269	283	1669	370	285	3109	**123**
P7	123	870	447						1440	**111**
P10*	10	60	38	56	61	419	99	738	1481	**117**
P11*	9	51	79	47	56	504	100	455	1299	**154**
P13*	14	393	123	108	320	1172	237	203	2570	**119**
P16	20	251	871	639	202	314	611		2908	**109**
P17	51	363	719	758	380	296	524		3088	**115**
P18*	26	45	39	93	158	133	66	274	834	**147**
P20	48	218	63	977	490	665	692	264	3417	**137**
P21	48	497	130	473	225	522	600	683	3178	**115**
P22	48	508	244	542	227	812	546	793	3719	**143**
P25*	41	508	44	218	144	172	93	433	1653	**118**
P26	128	776	163						1067	**87**
Avg.										**121**

Table 11 shows the fixation counts for both Group 1 and 2 in Baba Is You. Some participants were not able to complete all of the stages, which are shown as blank squares. There is a large variation in number of total fixations between the participants, which is explained by the time taken to complete the stages. There are some participants that show greater variance in per minute values from the average, such as P26 with 87 or P11 with 154, although the overall variance is low.

Table 12. Saccade counts for Baba Is You.

	0	1	2	3	4	5	6	7	Total	**Per min**
P1*	59	272	173	222	187	545	134	648	2240	**161**
P2*	18	122	94	282	292	1798	385	326	3317	**131**
P7	122	891	460						1473	**114**
P10*	11	75	39	56	61	468	103	827	1640	**129**
P11*	9	58	83	47	56	553	105	504	1413	**167**
P13*	13	445	155	114	337	1274	259	225	2822	**131**
P16	18	290	981	722	219	346	688		3264	**123**
P17	50	391	781	811	393	316	562		3302	**123**
P18*	25	46	39	94	160	135	79	300	878	**155**
P20	48	237	63	1047	528	714	724	278	3639	**146**
P21	50	549	134	540	281	609	645	780	3588	**130**
P22	49	556	277	596	256	939	581	871	4123	**159**
P25*	85	1060	107	417	254	431	219	1173	3746	**267**
P26	182	1440	256						1878	**153**
Avg.										**149**

Table 12 shows a similar trend as Table 11, with generally low variance in the per minute values outside of P25. Total saccades show a large variance again, with values ranging from 878 to 4123. P25's larger saccades per minute could be explained by them processing the game as more textual content rather than visual, causing a difference in physiological markers of cognitive load.

Table 13. Fixation counts for Terra Nil.

	1	2	3	4	Total	**Per min**
P3*	1190	2148	112		3450	**127**
P5*	416	806			1222	**152**
P6*	749	1042	770	1219	3780	**146**
P8*	647	1433	2063	402	4545	**180**
P9*	695	1076	1117	1257	4145	**135**
P12	1798	1555			3353	**143**
P14	1129	1492	1304		3925	**140**
P15	1267	1520	1243		4029	**165**
P19	1073	1629	922	573	4197	**165**
P27	1604	509			2113	**151**
Avg.						**150**

Terra Nil shows a more consistent number of total fixations than Baba Is You, with most participants hovering around 3500-4000. Exceptions occur in P5 and P27 showing lower values, and P8 having a clearly higher value. Fixations per minute show a similar trend of consistency, with values hovering around 140-160. P3 and P8 show up as distinct values. P8 having a higher fixation count can be interpreted as the player spending more time trying to understand the instructions and how to play the game (Pretorius et al., 2010, p. 7). P8 did spend a lot of time on stage three compared to other participants, which can also partially explain this result. Majority of this time was spent on exploring the newly opened map with new mechanics and visuals.

Table 14. Saccade counts for Terra Nil.

	1	2	3	4	Total	**Per min**
P3*	1395	2536	131		4062	**150**
P5*	420	826			1246	**155**
P6*	909	1311	951	1489	4660	**180**
P8*	682	1512	2179	429	4802	**190**
P9*	748	1175	1181	1390	4494	**147**
P12	1921	1704			3625	**154**
P14	1202	1588	1455		4245	**151**
P15	1329	1596	1356		4281	**175**
P19	1123	1753	961	603	4440	**174**
P27	1742	550			2292	**163**
Avg.						**164**

Saccade counts follow a similar trend as fixations in Table 13. P5 and P27 show a clear deviation from total amount of saccades when contrasted with the group's average values around 4000. In per minute values P8 stand out again, alongside P6. P8 experiencing both lower fixation and saccade counts could indicate processing the game as more textual than visual content, which to some extent is consistent with strategy games as a genre. P6 could have experienced a heightened cognitive load, as evident from a raised saccade count compared to fixation count.

The total number of fixations and saccades was then counted as a per minute value based on the time it took for participants to play the stages. This was done to include the failed attempts for evaluating cognitive load during those stages, as well as to create a more equal way to assess participants with their varying stage completion rates. This count is present in Tables 11-14 on the right-most column, with the total amount of fixations or saccades on the second column to the right.

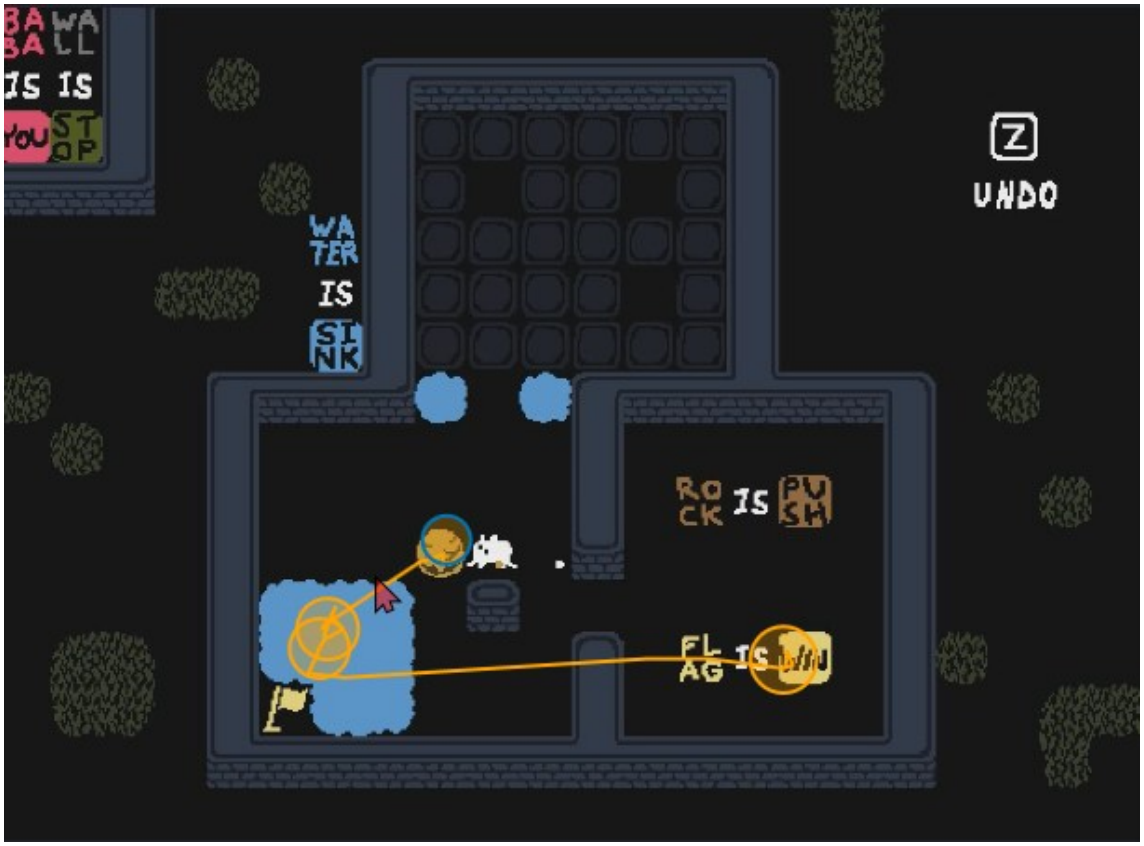


Figure 8. Example of eye tracking features in Baba Is You.

In Baba Is You the players focused usually on the player character, or the sentence structures and words present on the stages, as shown in Figure 8. The player has just begun to solve stage three and is observing the moveable objects and a new element included in the stage, water. After this, they observe the win conditions; the flag and the sentence flag is win. The player is seemingly trying to observe the relationship between the objects given to them and how to achieve the win condition.

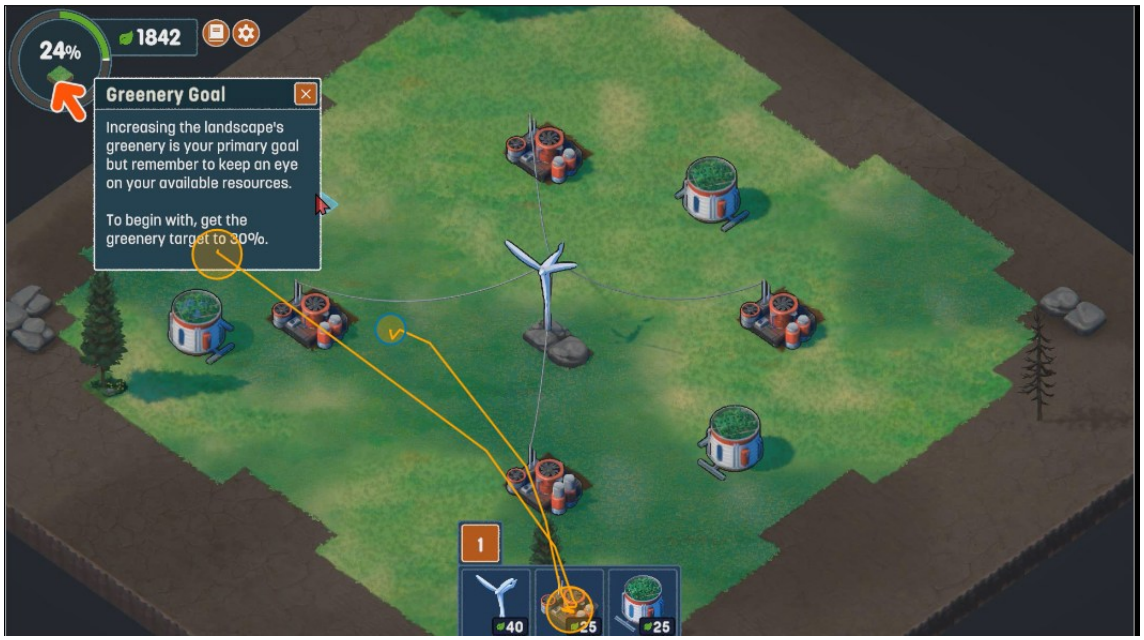


Figure 9. Example of eye tracking features in Terra Nil.

In Terra Nil the players focused on the instructions provided, UI-elements, as well as buildings that they placed on the map, some of which are shown in Figure 9. The player has just passed the first phase of stage one, where they are prompted to achieve a new goal before the map opens up. They focus first on the buildings they had just placed, before noticing the instructional prompt.

To evaluate whether there were significant differences between the groups outlined in Table 6, a t-test was run on fixations per minute and saccades per minute between the groups in their respective games. Fixations between Groups 1 and 2 in Baba Is You yielded a p-value of 0,182, meaning that the difference between the groups was not statistically significant. The same repeated for saccades in Baba Is You, with a p-value of 0,097, although being much closer to statistical significance. For Terra Nil's groups the t-test on fixations gave a p-value of 0,344, being not statistically significant. The same repeated again on saccades with a p-value of 0,474. This can be explained by a lack of proper grouping due to a failed STRT-test, causing the groups to mix in their actual working memory capacities, or by low sample size due to the participants being spread into two games. Due to this, we reject hypothesis H2A.

For the next analysis pass, the number of fixations and saccades per minute per game were calculated for each stage. This was done to create an evaluation on the difficulty of each stage, which we can then assess via a t-test between the two groups. The values for each stage are presented in Tables 15-18, with fixations and saccades for Baba Is You in Tables 15 and 16 respectively, whilst for Terra Nil in Tables 17 and 18. Members of Group 1 are noted with an asterisk (*). After the fixation and saccade rate for each stage was calculated, a difficulty rating based on the data was formulated. Terra Nil was left out of the calculation as the relatively low group size and low number of stages caused a shallow level of analysis. Graphs on Terra Nil are shown in Appendices 13 and 14.

For Baba Is You, the difficulty was calculated by assigning a numeric value to each stage based on how much time it took for successful participants to complete the stage, shown also in Table 7, where the shortest average completion time would rank as 1 whilst the longest as 8. Then the number of participants that successfully cleared the stage was ranked similarly, with the highest number of participants ranking as 1 whilst the lowest as 8. In both time taken and the amount of participants, any ties scored the lowest value to share whilst the next stage would be ranked as normal, i.e. if stages 0 and 1 tied for first they both ranked as 1, whilst the next stage would rank as 3. The two ranking values were added up to create a difficulty coefficient, which were combine with the average fixations and saccades per minute per stage to form an assessment of the stage's difficulty. This was plotted into a dotted graph with a fitted line to indicate possible relationships between the difficulty and the physiological markers. These are shown in Figures 10 and 11. The stage rated as difficulty 8, stage six shows up noticeably in the number of fixations and saccades, which may be due to one participant getting somewhat stuck on this level, although the average completion times seen in Table 7 do not show this. In addition to this, stage three with a difficulty rating of 10 followed closely. The fitted line fails to note any significant trends in the number of saccades but notes a similar effect with fixation counts as in Sevcenko et al's (2023, p. 9) study. H1 can be

rejected, as more difficult stages do not seem to correlate with cognitive load markers in fixation and saccade counts as a whole.

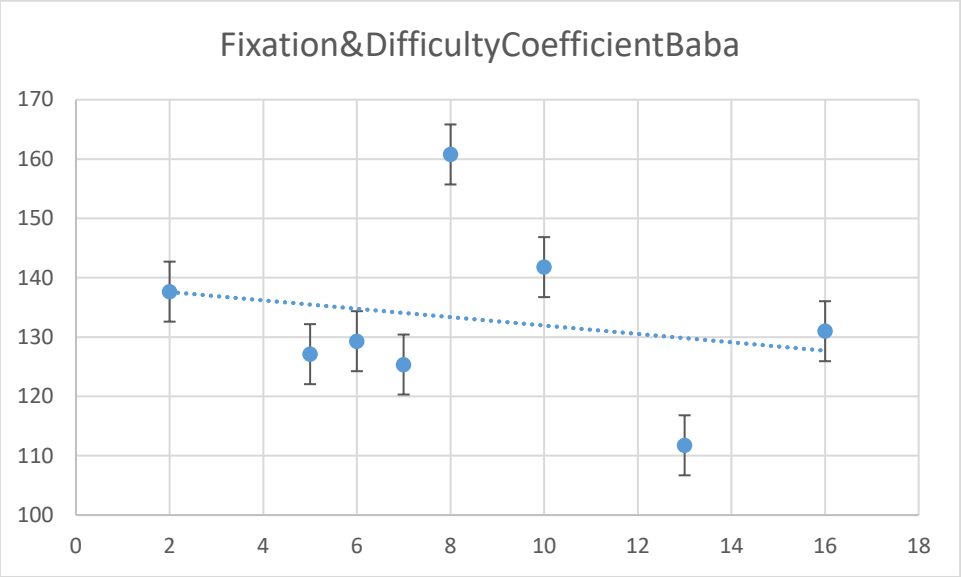


Figure 10. Relationship of stage difficulty coefficient and fixations per minute in Baba Is You.

Figure 10 shows the fixations per minute on the y-axis being contrasted with the stages difficulty coefficient on the x-axis. From this dotted graph a linear relationship line is formed, showing a rise in difficulty causing a lowered number of fixations for the stage. The relationship is not particularly steep, with around 10 fixations per minute of variance. Stage six shows up as an anomaly, having a difficulty coefficient of eight but a high fixation per minute count. This anomaly can be explained by the relatively low sample size making it easier for singular participants to skew the results if they got particularly stuck.

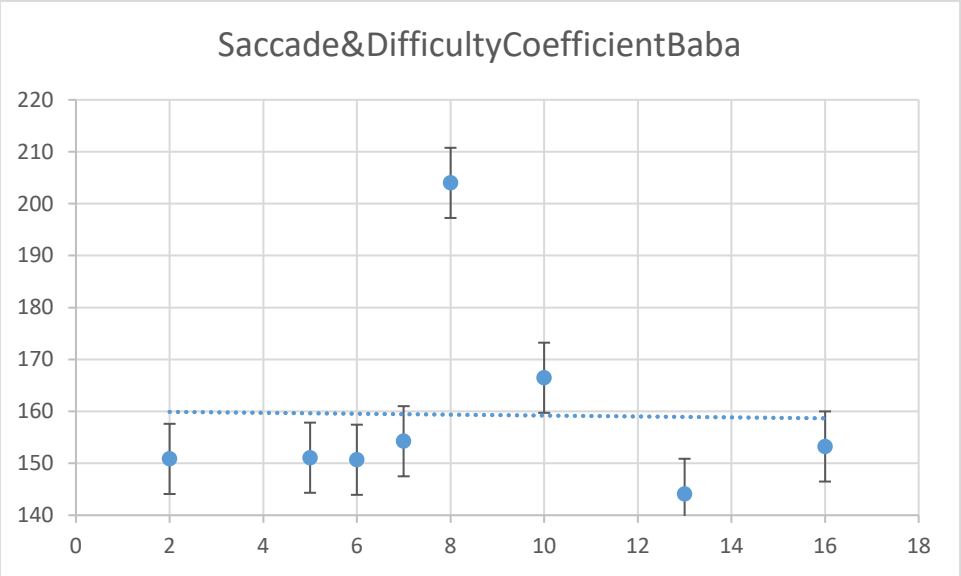


Figure 11. Relationship of stage difficulty coefficient and fixations per minute in Baba Is You.

Figure 11 shows a similar graph as Figure 9, but with saccades per minute in the y-axis. Although stage six appears yet again as an outlier, it does not affect the overall relationship line, which is almost straight. This could be due to Baba Is You having a unique combination of visual and textual problem solving, causing saccade counts to plateau.

Table 15. Fixations per minute per stage in Baba Is You.

Stage	0	1	2	3	4	5	6	7
P1*	163,50	117,32	123,17	140,70	123,68	106,98	144,19	74,80
P2*	122,58	136,86	139,57	160,82	125,00	N/A	155,90	59,50
P7	141,33	105,14	N/A	N/A	N/A	N/A	N/A	N/A
P10*	111,11	101,75	128,52	132,49	120,75	117,88	180,16	110,11
P11*	155,17	200,66	97,81	181,64	156,85	112,70	168,12	274,70
P13*	142,37	123,07	130,55	144,42	120,64	N/A	144,35	116,37
P16	106,48	120,54	94,96	112,91	116,57	120,52	N/A	N/A
P17	81,94	88,82	168,69	116,28	86,73	73,12	N/A	N/A
P18*	158,54	146,34	170,55	160,67	147,46	128,32	157,58	147,07
P20	163,54	142,83	139,28	N/A	136,14	125,04	N/A	N/A
P21	119,01	120,96	126,81	127,29	121,09	N/A	N/A	N/A
P22	234,91	145,35	103,90	134,28	148,75	N/A	200,32	134,34
P25*	130,50	117,33	101,58	148,11	147,97	109,50	135,54	N/A
P26	96,14	88,16	N/A	N/A	N/A	N/A	N/A	N/A
Avg.	137,65	125,37	127,12	141,78	129,30	111,76	160,77	130,98

Table 15 shows the fixations per minute across the stages in Baba Is You. All of the stages average across the participants to about 130, with stages five and six showing as outliers. Stage five was the stage that caused a lot of problems for the participants, being either the stage they got stuck on or the stage they skipped and returned to later on. Stage six showed a similar trend, which can be explained with the solution to the stage being the most creative of all the stages. The results can also be explained by a lack of participants at these stages, making the data more susceptible to singular anomalies affecting it.

Table 16. Saccades per minute per stage in Baba Is You.

Stage	0	1	2	3	4	5	6	7
P1*	182,01	173,43	180,58	205,49	182,11	147,60	227,31	139,68
P2*	116,13	140,31	138,10	168,59	128,98	N/A	162,22	68,06
P7	140,18	107,68	N/A	N/A	N/A	N/A	N/A	N/A
P10*	122,22	127,19	131,91	132,49	120,75	131,67	187,44	123,39
P11*	155,17	230,46	102,15	181,64	156,85	123,78	176,57	303,98
P13*	132,20	139,35	164,51	152,44	127,05	N/A	157,75	128,98
P16	95,83	139,27	106,95	127,58	126,38	132,80	N/A	N/A
P17	81,12	95,68	183,24	124,42	89,70	78,20	N/A	N/A
P18*	152,44	149,59	170,55	162,40	149,32	130,25	188,62	161,03
P20	163,54	155,27	139,28	N/A	146,69	134,25	N/A	N/A
P21	123,97	133,62	130,71	145,32	151,22	N/A	N/A	N/A
P22	239,80	159,09	117,96	147,53	168,13	N/A	212,98	147,55
P25*	270,56	244,83	247,02	283,32	261,00	274,38	319,16	N/A
P26	136,71	163,70	N/A	N/A	N/A	N/A	N/A	N/A
Avg.	150,85	154,25	151,08	166,48	150,68	144,11	204,01	153,24

Table 16 shows an average around 150 for saccades across all of the stages of Baba Is You. Stage six shows up as an outlier again with 204 saccades per minute on average. Interestingly, this stages solution focuses more on the visual side, having the player realize they can transform into the wall and move towards the goalpost. The outlier can be explained similarly as in Table 15 with a lower number of participants, although the same does not hold true for the saccades of stage five, indicating that another factor may be the cause.

Table 17. Fixations per minute per stage in Terra Nil.

Stage	1	2	3
P3*	131,78	124,90	N/A
P5*	162,95	N/A	N/A
P6*	146,53	135,97	144,86
P8*	197,65	159,12	189,06
P9*	130,09	123,05	143,25
P12	151,52	N/A	N/A
P14	150,44	134,26	N/A
P15	160,66	151,89	N/A
P19	183,77	144,61	176,89
P27	154,34	N/A	N/A
Avg.	156,97	139,11	163,52

Table 17 shows the fixations per minute per stage in Terra Nil. Although the data is scarce, some trends can be observed from it. Stage one shows a fixation count consistent with Baba Is You's results. The stage also serves as an introduction that asks the player to grasp some core gameplay mechanics. Stage two shows a lower value by 17, which can be explained by it having less textual instructions present than stages one and three. Stage three then again adds layers of complexity, which cause the rate to jump to 163.

Table 18. Saccades per minute per stage in Terra Nil.

Stage	1	2	3
P3*	154,48	147,47	N/A
P5*	164,51	N/A	N/A
P6*	177,83	171,07	178,92
P8*	208,34	167,90	199,69
P9*	140,01	134,37	151,46
P12	161,89	N/A	N/A
P14	160,16	142,89	N/A
P15	168,58	159,53	N/A
P19	192,34	155,61	184,38
P27	167,62	N/A	N/A
Avg.	169,58	154,12	178,61

Table 18 shows a similar trend to Table 17, with stage two showing a clear drop in saccade count compared to the other two stages.

Lastly, a t-test was performed using the groups in Table 6 for each game and for each stage for both fixations and saccades per minute. Findings are presented in Tables 19 and 20 for Baba Is You and Terra Nil. Tests were run for stages where a group had at least two or more participants, as per t-tests requirements. Only two findings ended up with a statistically significant result, both occurring in stage three of Baba Is You. This stage proved to be a somewhat challenging point for especially Group 2, even if they completed the stage. This finding would suggest that there can be clear differences between the two groups, more of which could be visible with a larger sample size. Due to being somewhat of a defining stage in terms of difficulty, we can observe these effects here still. In this stage, the more successful group held lower fixation counts and higher saccade counts than the less successful group, as observed in Sevcenko et al. (2023, p. 9-11).

Table 19. P-values on Baba Is You's fixations and saccades per minute between Group 1 and 2.

Stage	0	1	2	3	4	5
p-value (fixations)	0,395	0,120	0,484	0,002*	0,163	0,325
p-value(saccade)	0,230	0,055	0,126	0,022*	0,155	0,110

Table 19 shows the results of the t-test performed on fixation and saccadic rates of Groups 1 and 2 in Baba Is You. Stage three shows p=0,002 and p=0,022, which are both statistically significant results. Group 1 showed less fixations than Group 2 in stage three, as well as a higher saccade count. This would indicate that Group 1 had a higher cognitive load than Group 2. The lower number of fixations could also indicate that Group 1 needed less time to process the actual textual content and was able to focus more on the visual aspects or their pre-existing schemas. The higher saccade count would also support the theory that the processing was happening more on a visual level. Although not statistically significant, stage one gets close to statistically significant values in saccadic rate, with p=0,055.

Table 20. P-values on Terra Nil's fixations and saccades per minute between Group 1 and 2.

Stage	1	2
p-value (fixations)	0, 332	0,230
p-value(saccade)	0,468	0,406

Overall, the results showcase that games that combine textual and visual content share similar trends of cognitive load markers in fixations and saccades but can differ in saccadic rate and count depending on the amount of textual information being processed. Participants with higher working memory capacity do not seem to perform differently overall from the lower working memory capacity participants, although

exceptions occur in certain stages. Although participants gave testimonies of frustration and confusion during and after the play session, they overall rated the workload as a moderate-to-low, indicating that tutorials can help in managing subjective workloads. The higher working memory capacity participants showed a statistically significantly lower workload assessment in Baba Is You, indicating that the tutorial helped them optimize their cognitive load or did not hinder their pre-existing schemas.

6 Discussion and limitations

This thesis aimed to answer the following research questions: how do video game tutorials affect a player's cognitive load, and how does cognitive load affect the learnability of the game? To answer these questions, the players' cognitive loads were assessed during gameplay. In the study the working memory capacity was attempted to measure and thus considered in the interpretation of the results. Answering the NASA-TLX workload survey, the participants evaluated the workload as low to moderate, with mental demand and frustration being the leading factors. A statistically significant difference was observed in the two group's evaluations in Baba Is You, which can indicate higher cognitive load for Group 2. This could be due to Group 1 experiencing a flow state, having used their pre-existing schemas to streamline to gameplay experience, or due to Group 2 not getting any of the cognitive load benefits of the implicit tutorial style (Seyderhelm & Blackmore, 2023, p. 22; Lee & Heeter, 2017; Cao & Liu, 2022).

Players performed across stages and games similarly to each other within each group with little to no statistically significant differences, meaning that most participants in Group 2 managed to optimize or control their cognitive load during the gameplay session. Group 1 most likely utilized pre-exiting schemas to streamline their gameplay experience (Seyderhelm & Blackmore, 2023, p. 22). Group 1 also had a slightly younger skew with more gaming expertise in self-reporting, which can indicate familiarity with the type of task, leading to lower cognitive load due to higher levels of engagement or immersion (Andersen et al., 2012).

The difference between the games is not statistically significant for the most part, although in the analysis Baba Is You's data leans towards statistical significance, or at the very least close to that. This can be attributed to an insufficient STRT-test creating mixed working memory groups in Terra Nil, as well as the game having a lower number of participants. It is to be noted that more participants complained aloud about not understanding what to do in Terra Nil than in Baba Is You, which can be explained by Baba Is You being a puzzle game, priming the participants towards accepting their

confusion as a planned experience. Some minor effects in cognitive load markers were observed in Baba Is You's stages 0 and 1 that contrast Sevcenko et al's (2023, p. 9-11) findings due to the inverse linear relationship. As mentioned before, this can be explained due to Baba Is You involving more textual processing than Sevcenko et al's serious game. Additionally, Baba Is You stage three showed a statistically significant difference in fixations and saccades per minute between Groups 1 and 2, indicating that Group 2 had a higher or less optimized cognitive load than Group 1, hindering their learnability of the new game mechanic introduced. The fixation and saccade counts are consistent with Sevcenko et al's (2023, p. 9-11) findings, with the more successful group having lower fixation numbers and higher saccade counts.

With the current findings, it is not possible to evaluate effects on learnability due to a lack of statistical significance. However, by performing a light-weight qualitative analysis, some small patterns can be observed. Baba Is You's stage zero had remarkably low completion times, but stage one had a lot more variance in completion times. Stage zero is supposed to act as the most basic tutorial, explaining how to move and how the sentences work. However, a fair number of participants managed to brute force through the stage without learning from it, leading to the higher completion time in stage one. This is one of the risks of implicit tutorials, where players can skip important knowledge or misunderstand it.

This thesis is not without its limitations. Firstly, the control runs suggested by NASA-TLX were omitted due to possible time constraints. Similarly, the participants in the original experimental design were supposed to play one game as a practice run before playing their real allocated game, which would have improved ecological validity. Although small talk was incorporated to reduce the stress of the environment, a practice run could have had similar effects with additional benefits of familiarizing the peripherals. Small talk could also add additional social pressure or take away from the participants working memory capacity. The small talk was controlled by engaging in it only every 5 minutes if not initiated by the participant. By removing these factors, the study was able to

incorporate more participants but lowered its ecological validity. Second, by running two games and dividing them into several groups, the statistical analysis of data became extremely difficult. A more focused experimental design could have shown more significant findings if only one game was used. Related to this, the STRT-test failing due to technical issues also caused parts of the data to be analysed differently from the original plans and removed some possible applications of it. This could have been minimized by using a similar setup as Lee and Heeter (2017), where a visual stimulus would appear on the screen during gameplay. Some participants were also quite uncomfortable with the peripherals, taking until the 3rd attempt at the STRT-test to properly locate spacebar to answer stimulus. This could have also been alleviated by performing a control run.

Cognitive load can also be measured via more physiological measures other than fixation and saccade count, such as blink rate, although the measure can be unreliable at times. The physiological markers were kept at a minimum to ensure reliability of research setup, but a researcher whose more familiar with the devices and software should run a more robust dataset. The difficulty ratings employed for the two different statistical tests also varied greatly, furthering the unreliability of the findings. These decisions were made due to a lack of individual difficulty ratings, which could have been administered in all the statistical tests. Further research in a similar context should utilize these subjective difficulty ratings. A majority of the statistical analyses was done via excel, whilst a large chunk was done on hand. Despite double-and triple-checks, errors within the results could still exist. Utilizing R or similar solutions for statistical analysis would be recommended in future research.

Future research should focus on one game at a time to create more statistically robust data. Difficulty ratings could be done both subjectively by the participants rating the stages after playing as well as via performance indicators. More qualitative data could be used, such as interviews or thinking-out-loud to gain understanding on why certain stages might have felt more difficult, or what design choices players felt affected the

games' learnability. This should be contrasted with quantitative data to gain a more complete understanding of factors affecting learnability, as users often report vastly different issues than what hinders the experience.

7 Conclusion

This thesis set out to study the effects of tutorials in video games on players cognitive load, and thus how the cognitive load may affect the learnability of the game. The thesis set out to do so by combining qualitative and quantitative measures of cognitive load to gain a more comprehensive understanding of factors affecting it. In studying anything related to cognition, we often deal with subjective opinions or self-reporting, as measures of cognition are often either unreliable or difficult to administer. Eye tracking offers a light-weight method in estimating cognitive load, as subconscious eye movements tend to correlate with changes in cognition. When combined with self-reporting in the form of the NASA-TLX survey, this thesis was able to combine the subjective markers with more objective physiological signs. This thesis showed that the combination of the aforementioned methods can create robust data, even if issues did occur during the study. This thesis was also able to show that participants with higher estimates of working memory capacity had higher self-reported cognitive load levels, indicating that the tutorials were either not doing enough to optimize their cognitive load, or that they were otherwise encumbered during the study.

Despite this, the performance of the groups was comparable, indicating that the tutorials did still serve their purpose in teaching the games mechanics, thus improving learnability. The study also showed that the trend of successful participants having lower fixation count and higher saccade count applied in this study as well, supporting previous literature. This thesis showed that the area of video game learnability still remains a fruitful place for research, which can be applied to serious games or other multimedia content. Video games, such as many other hobbies, have paved the forefront for technological advancement such as phones, processors and more, so it is not unimaginable that exploring the ways games try to teach their players could aid in other areas of technology. Games of all forms have existed since the dawn of humanity, giving us a way to explore our creativity, challenge ourselves, as well as connect with each other.

References

Alexiou, A. & Schippers, M. (2018). Digital game elements, user experience and learning: A conceptual framework. *Education and Information Technologies, 6, 2545-2567.* https://doi.org/10.1007/s10639-018-9730-6

Andersen, E., O'Rourke, E., Liu, Y., Snider, R., Lowdermilk, J., Truong, D., Cooper, S. & Popovic, Z. (2012). The Impact of Tutorials on Games of Varying Complexity. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 59-68.* DOI:10.1145/2207676.2207687

Benvenuti, D., Ferro, L., Marrella, A. & Catarci, T. (2023). An Approach to Assess the Impact of Tutorials in Video Games. *Informatics, 1.* DOI 10.3390/informatics10010006

Blaskovic, L., Zuzic, A. & Orehovacki, T. (2023). Evaluating a Conceptual Model for Measuring Gaming Experience: A Case Study of Stranded Away Platformer Game. *Information, 6, 350-376.* DOI 10.3390/info14060350

Cao, S. & Liu, F. (2022). Learning to play: understanding in-game tutorials with a pilot study on implicit tutorials. *Heliyon, 11.* https://doi.org/10.1016/j.heliyon.2022.e11482

Chandler, P. & Sweller, J. (1991). Cognitive Load Theory and the Format of Instructions. *Cognition and Instruction, 4, 293-332.* DOI:10.1207/s1532690xci0804_2

Chang, C-C., Liang, C., Chou, P-N. & Lin, G-Y. (2017). Is game-based learning better in flow experience and various types of cognitive load than non-game-based learning? Perspective from multimedia and media richness. *Computers in Human Behavior, 71, 218-227.* DOI 10.1016/j.chb.2017.01.031

Chen, B., Yan, X., Hu, X., Kao, D. & Liang, H.N. (2024). Impact of Tutorial Modes with Different Time Flow Rates in Virtual Reality Games. *Proceedings of the ACM on Computer Graphics and Interactive Techniques, 1, 1-19.* https://doi.org/10.1145/3651296

Chen, O., Kalyuga, S. & Sweller, J. (2016). Relations between the worked example and generation effects on immediate and delayed tests. *Learning and Instruction, 45, 20-30.* DOI 10.1016/j.learninstruc.2016.06.007

Chen, O., Kalyuga, S. & Sweller, J. (2017). The Expertise Reversal Effect is a Variant of the More General Element Interactivity Effect. *Educational Psychology Review, 2, 393-405.* https://doi.org/10.1007/s10648-016-9359-1

Cooper, G., Tindall-Ford, S., Chandler, P. & Sweller, J. (2001). Learning by imagining. *Journal of Experimental Psychology: Applied, 1, 68-82.* DOI 10.1037/1076-898X.7.1.68

Csikszentmihalyi, M. (1975). *Beyond Boredom and Anxiety: Experiencing Flow in Work and Play.* Jossey-Bass. ISBN: 978-0-787-95140-5

Gee, J. P. (2005). Learning by Design: Good Video Games as Learning Machines. *E-Learning and Digital Media, 2, 5-16.* https://doi.org/10.2304/elea.2005.2.1.5

Höffler, T. & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction, 6, 722-738.* DOI 10.1016/j.learninstruc.2007.09.013

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *The Internet and Higher Education, 1, 13-24.* DOI 10.1016/j.iheduc.2004.12.001

Korhonen, H., Montola, M. & Arrasvuori, J. (2009). Understanding playful user experiences through digital games. *International Conference On Designing Pleasurable Products and Interfaces, DPPI09, 13-16.* https://www.researchgate.net/publication/242084991_Understanding_playful_user_experiences_through_digital_games Retrieved 13.12.2024.

Leahy, W. & Sweller, J. (2016). Cognitive load theory and the effects of transient information on the modality effect. *Instructional Science, 1, 107-123.* https://doi.org/10.1007/s11251-015-9362-9

Lee, Y. H. & Heeter, C. (2017). The effects of cognitive capacity and gaming expertise on attention and comprehension. *Journal of Computer Assisted Learning, 5, 473-485.* DOI:10.1111/jcal.12193

Lee, Y., Kim, G., Lee, K. H., Park, J. & Kim, H. K. (2024). Comparison of Tutorial Methods in Virtual Reality Games for a Better User Experience. *Applied sciences, 16.* https://doi.org/10.3390/app14167141

Li, C., Dai, C., Chan & W. K. (2024). Optimizing Tutorial Design for Video Card Games Based on Cognitive Load Theory: Measuring Game Complexity. *HCI in Games, 55-67.* https://doi.org/10.1007/978-3-031-60692-2_5

NASA. (2022). *NASA TLX Task Load Index – Paper/Pencil Version.* Retrieved 11.12.2024 https://humansystems.arc.nasa.gov/groups/TLX/tlxpaperpencil.php

van Merriënboer, J. & Kirschner, P. (2018). *Ten steps to complex learning: A systematic approach to four-component instructional design (3rd ed.).* Routledge. https://doi.org/10.4324/9781315113210

Montani, F., Vandenberghe, C., Khedhaouria, A. & Courcy, F. (2020). Examining the inverted U-shaped relationship between workload and innovative work behavior: The role of work engagement and mindfulness. *Human Relations, 1, 59-93.* http://doi.org/10.1177/0018726718819055

Morin, R., Léger, P-M., Senecal, S., Bastarache, M-C., Lefèbrve, M. & Fredette, M. (2016). The Effect of Game Tutorial: A Comparison Between Casual and Hardcore Gamers. *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts, 229-237.* https://dl.acm.org/doi/10.1145/2968120.2987730

Orru, G. & Longo, L. (2019). The Evolution of Cognitive Load Theory and the Measurement of Its Intrinsic, Extraneous and Germane Loads: A Review. *Communications in Computer and Information Science, 1012, 23-48*. https://doi.org/10.1007/978-3-030-14273-5_3

Pollock, E., Chandler, P. & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction, 1, 61-86.* DOI 10.1016/S0959-4752(01)00016-0

Poretski, L. & Tang, A. (2022). Press A to Jump: Design Strategies for Video Game Learnability. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1-26*. https://dl.acm.org/doi/10.1145/3491102.3517685

Pretorious, M., Gelderblom, H. & Chimbo, B. (2010). Using eye tracking to compare how adults and children learn to use an unfamiliar computer game. *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer*

Scientists and Information Technologists, 275-283. https://dl.acm.org/doi/10.1145/1899503.1899534

Renkl, A., Stark, R., Gruber, H. & Mandl, H. (1998). Learning from Worked-Out Examples: The Effects of Example Variability and Elicited Self-Explanations. *Contemporary Educational Psychology, 1, 90-108.* DOI 10.1006/ceps.1997.0959

Sanchez, J. L. G., Vela, F. L. G., Simarro, F. M., & Padilla-Zea, N. (2012). Playability: analysing user experience in video games. *Behaviour & Information Technology, 10, 1033-1054.* http://dx.doi.org/10.1080/0144929X.2012.710648

Sevcenko, N., Appel, T., Ninaus, M., Moeller, K. & Gerjets, P. (2022). Theory-based approach for assessing cognitive load during time-critical resource-managing human-computer interactions: an eye-tracking study. *Journal on Multimodal User Interface, 1, 1-19.* https://doi.org/10.1007/s12193-022-00398-y

Seyderhelm, A. & Blackmore, K. (2023). How Hard Is It Really? Assessing Game-Task Difficulty Through Real-Time Measures of Performance and Cognitive Load. *Simulation & Gaming, 54, 294-321.* https://doi.org/10.1177/10468781231169910

Shankhwar, K. & Smith, S. (2022). An interactive extended reality-based tutorial system for fundamental manual metal arc welding training. *Virtual reality 26, 3, 1173-1192.* DOI:10.1007/s10055-022-00626-6

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science, 12, 257-285*. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (1994). Cognitive load theory, Learning difficulty, and instructional design. *Learning and instruction, 4, 295-312.* DOI 10.1016/0959-4752(94)90003-5

Sweller, J., van Merriënboer, J. & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review, 2, 261-292.* https://doi.org/10.1007/s10648-019-09465-5

Appendices

Appendix 1. Consent form page 1

Study - An assessment on video game tutorial learnability from a cognitive load framework

We ask you to participate in this combined questionnaire and eye-tracking study. The study explores how video game tutorials affect a players' cognitive load and subsequently changes the learning experience.

After reading this form, you'll be asked to give your consent to participate in the study.

Aim of the study

The aim of this study is to explore how tutorials can aid in improving video game learnability for players from the framework of cognitive load theory by optimizing player cognitive load. The study looks to answer two research questions:

1. *How do video game tutorials affect a player's cognitive load?*
2. *How do the effects on a player's cognitive load affect the learnability of the game?*

Timetable and phases of the study

The study will be conducted between January and March of 2025. The study consists of a questionnaire and an eye-tracking experiment. Additionally, the study will measure task performance and functionality of the working memory. Task performance measures vary depending on the game, usually consisting of measures such as completion time or error rate. The questionnaire conducted will be the NASA Task Load Index (TLX), which consists of a self-evaluation of the tasks work load. The eye-tracking experiment will measure physical indicators of task load, such as saccades and fixations. During the gameplay session a STRT-working memory test will also be conducted, in which the participants response to an outside trigger is measured and evaluated.

Expanded information on methods of gathering the materials

Eye-tracking data and performance measures are taken automatically during the study by the researcher. Participants are only required to calibrate the eye-tracking device to themselves, play the game, and answer the STRT-test during the gameplay session. Afterwards, the participant will be asked to answer the NASA-TLX -questionnaire.

Benefits and possible risks of the study

The findings of the study can expand on the understanding on cognitive loads in different types of systems, not just games. Video games are form of rich multimedia, which places a unique load on the user at their first steps. The findings can also aid in expanding knowledge on how cognitive loads can be optimized, helping to create

Appendix 2. Consent form page 2

better understanding on how to formulate educational and informational content. There are no foreseeable risks associated with the study.

Confidentiality, data access and safekeeping

The data collected in this study will be handled confidentially according to GDPR and Finnish data privacy law.

Data will be accessed and used in a pseudonymized form, with no personal information gathered that could be connected to any natural persons. Information will not be given to individuals not associated with the study.

Persons right to privacy in publication

In reporting the results, all individuals are pseudonymized. Any possible information that could be connected to a natural person will not be gathered.

The research file and materials gathered during the study will be kept until the masters thesis has been completed, after which they will be deleted.

Voluntariness

Participating in this study is voluntary and you can cancel your consent at any stage of the study. Upon canceling, any information gathered before the cancellation can be used in the thesis going forwards.

Privacy in publishing and publicity of the thesis

The participants' names will not be mentioned in the thesis and the results will be handled with complete anonymity.

Complete results of the study will be sent to the participants upon their request.

Participant information for sending results

Researcher contact information

Mikael Nylund

University of Vaasa, Masters programme in Technical Communication

d119027@student.uwasa.fi 0458972405

Appendix 3. Data protection notice page 1

Data protection notice EU data protection regulation/GDPR (106/679) art 12-14

Date 16.01.2025

Name of registry
An assessment on video game tutorial learnability from a cognitive load framework

Data controller(s)
Mikael Nylund (d119027@student.uwasa.fi / 0458972405)
Personal information use cases and basis
Your personal information will be gathered for a thesis on video game tutorial learnability and cognitive loads.
The study consists of around a 30min. long gameplay session, during which eye-tracking data, performance measures (time spent, success rate), and working memory evaluation will be gathered. After the session, the participant will answer a questionnaire measuring the task load of the session. Materials gained from this will give a combined qualitative and quantitative estimate on participants cognitive load during the gameplay session. Before the study begins, the participants will be briefly introduced to the study environment.

Participating in this study is completely voluntary. The basis for accessing personal information is consent. Consent can be revoked at any point by informing the data controller. Revoking the consent will not affect the legality of access done before cancellation of the agreement.

Personal information retention period
After the completion of the thesis, the stored information will be destroyed.

Informational content and sources of the registry

- Demographic information
  - Gender, age, experience in video games
- Eye-tracking data (saccades, fixations), performance measures, STRT-working memory test, NASA-TLX -questionnaire

Information will be collected from the participants themselves during the study. Collected information will be changed so that they cannot be connected to the participant.
Rights of the registered
According to data privacy and protection law, you hold the right to access your data, correct them, delete them (right to be forgotten), control the access and use of personal information. If you wish to exercise your right, contact the data controller.

Right to complain to an authority
You hold the right to make a complaint to an authority that governs the use of personal information if you suspect that your personal information is being used illegally: tietosuoja.fi, tel: 0295666700, email: tietosuoja@om.fi

Appendix 4. Data protection notice page 2

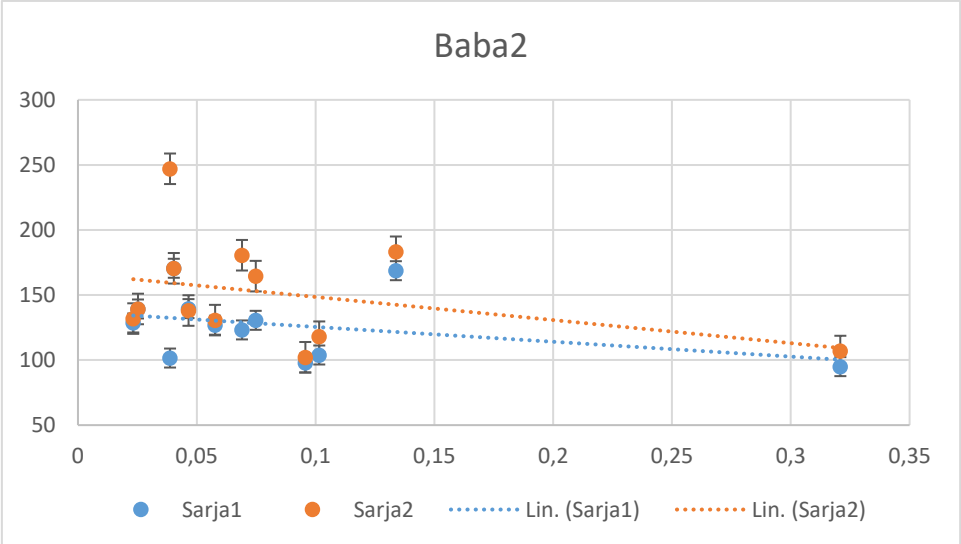
Recipients of personal information
Your personal information will not be given to outsiders.

Safekeeping principles of the registry
Manual materials will be kept in a locked environment. Digital materials will be protected by an username and a password or via two-factor authentication (MFA). Materials will have directly identifiable information deleted. The materials will be accessed only by University of Vaasa network (/tools) or on a local network.

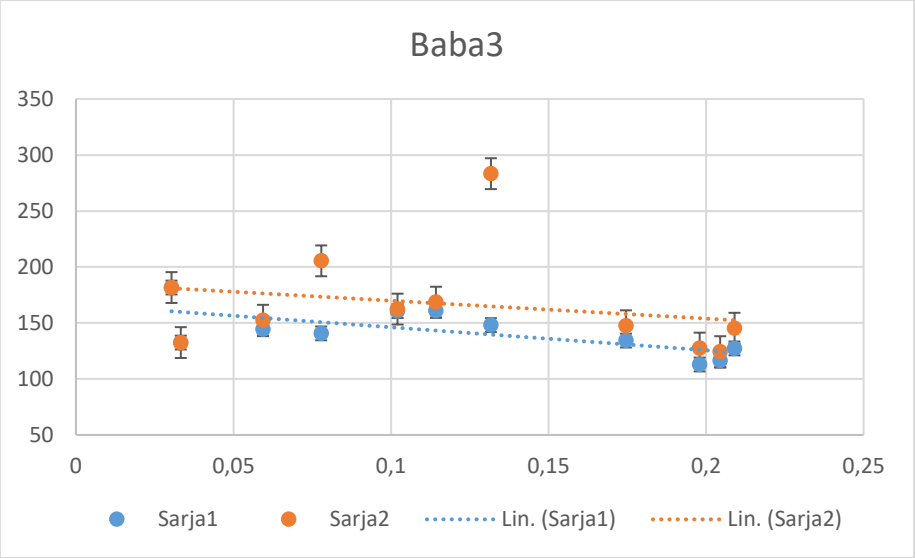
Appendix 5. Textbook instructions in the beginning of Terra Nil

Appendix 6. Instructional pop-up in Terra Nil

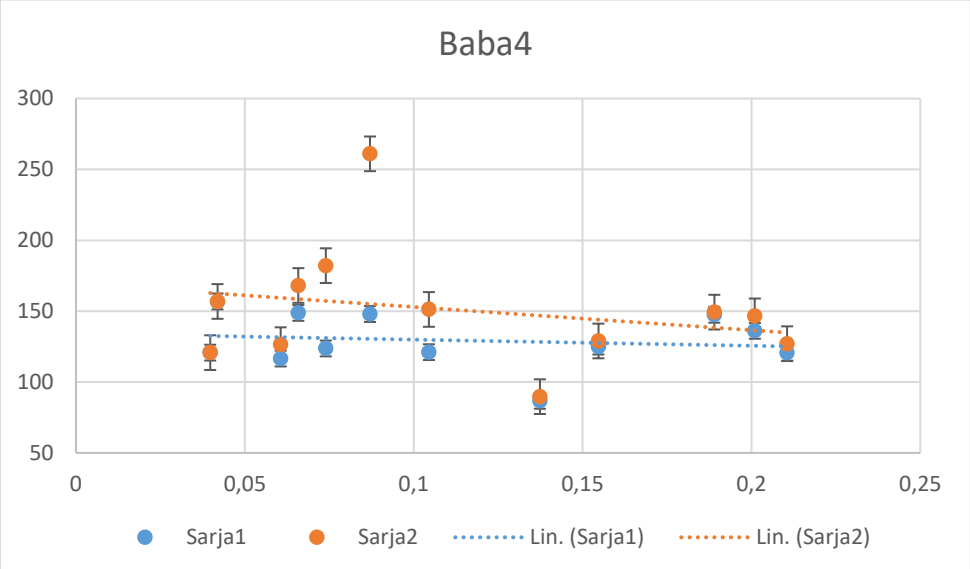
Appendix 7. Baba Is You's stage 2 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line



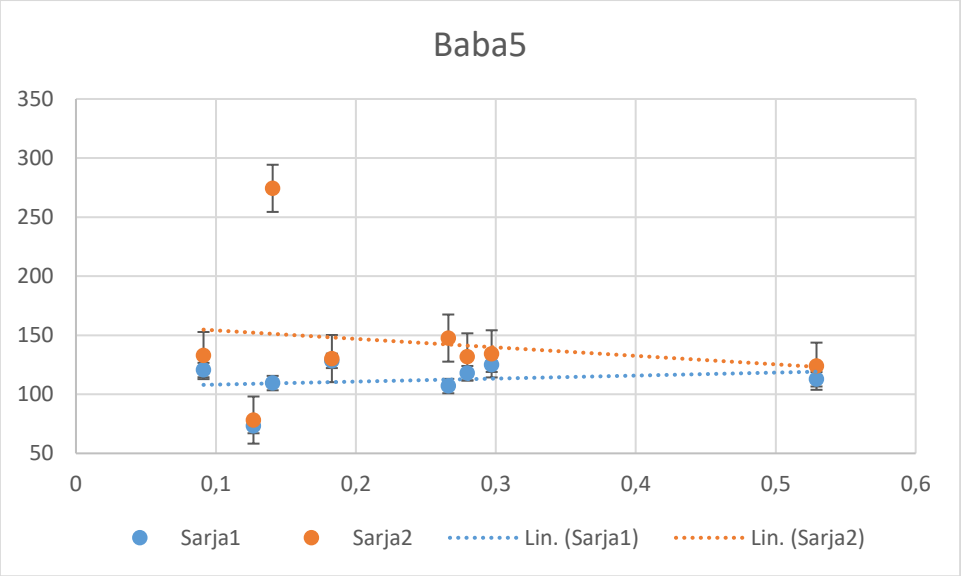
Appendix 8. Baba Is You's stage 3 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line



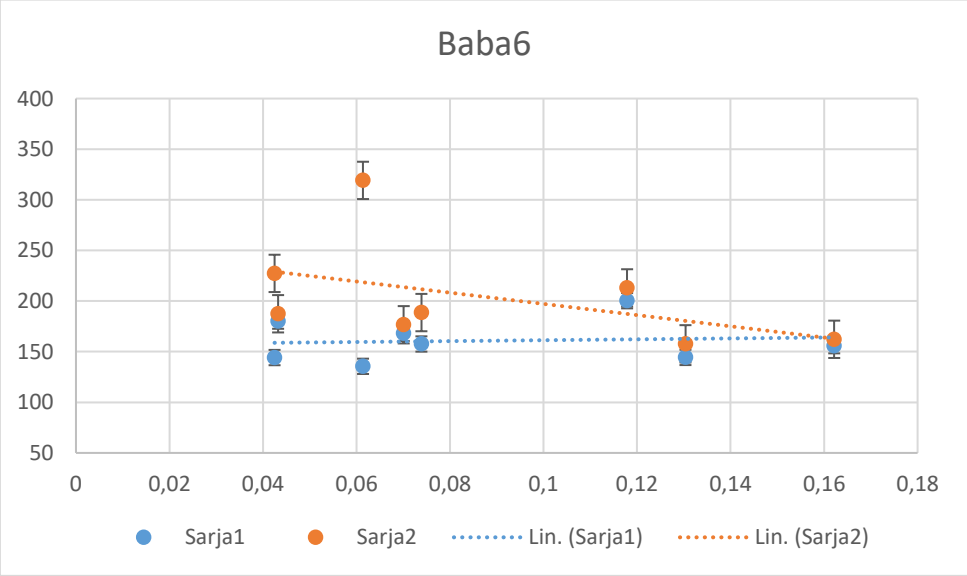
Appendix 9. Baba Is You's stage 4 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line



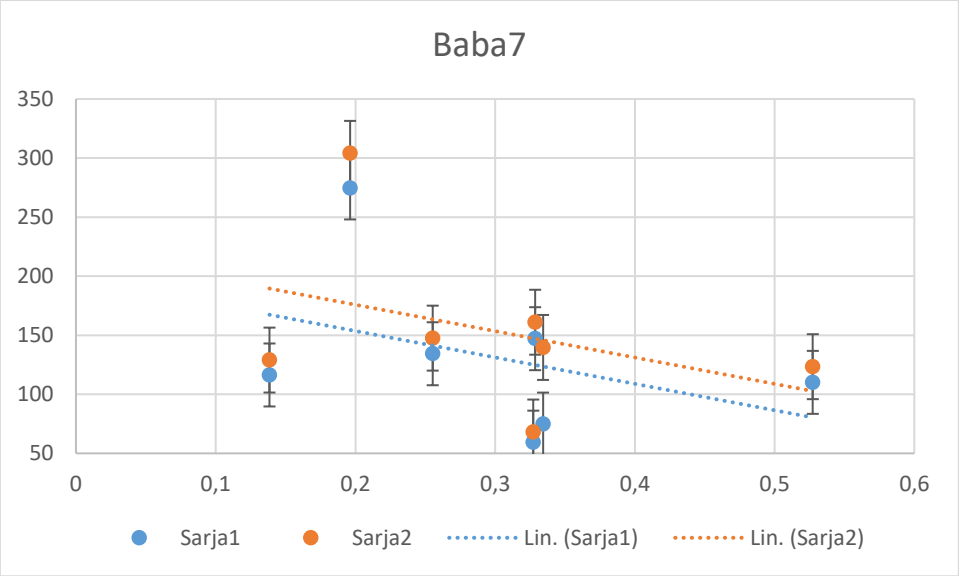
Appendix 10. Baba Is You's stage 5 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line



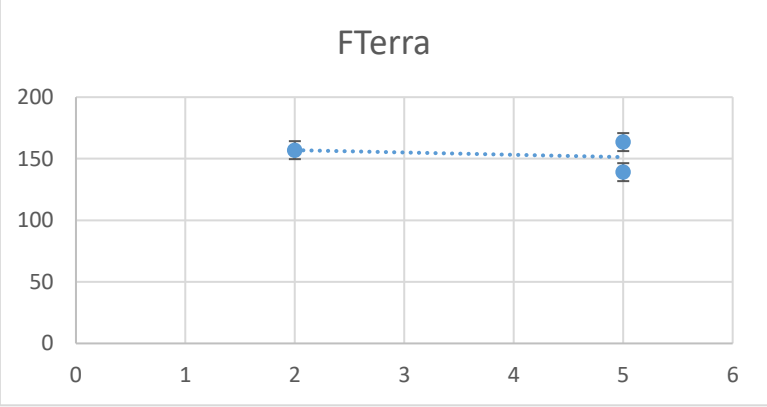
Appendix 11. Baba Is You's stage 6 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line



Appendix 12. Baba Is You's stage 7 fixations (blue) and saccades (orange) per minute plotted against the subjective difficulty rating with a linear relationship line



Appendix 13. Terra Nil's fixations per minute plotted against the difficulty coefficient with a linear relationship line



Appendix 14. Terra Nil's saccades per minute plotted against the difficulty coefficient with a linear relationship line