Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Development and validation of Restaurant Game for measuring and training working memory

Master thesis

2024
Bc. Michaela Dlugošová

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Development and validation of Restaurant Game for measuring and training working memory

Master thesis

Study Programme:	Cognitive science
Field of Study:	Computer science
Department:	Department of Applied Informatics
Supervisor:	Mgr. Martin Marko, PhD.
Consultant:	Prof. Anja Podlesek

Bratislava, 2024
Bc. Michaela Dlugošová

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Bc. Michaela Dlugošová
Study programme:	Cognitive Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Development and validation of Restaurant Game for measuring and training working memory

Annotation: Executive functions naturally decrease during our life, lowering the quality of life. However, the decline can be observed and predicted by assessments and its effects can be mitigated using cognitive training. Both methods bear several disadvantages such as lack of flexibility, limited ecological validity, or cost. When developing a new method that addresses shortcomings of currently available measuring and training tasks, it is good to start with working memory as it was shown to highly influence executive functions and other higher-order cognitive processes. The first step when validating a new training method is comparing its results with measurements that are currently used to measure the construct in question, e.g. Digit span task. The reason for this step is that the training method has to reflect the targeted construct in the first place. For this reason, it is also crucial to take into account the shortcomings of current measuring tasks during the training's development.

Aim: Develop and validate a video game focused on working memory based on current knowledge of measuring and training methods, compare it with existing measuring methods, and evaluate its efficacy.

Supervisor:	Mgr. Martin Marko, PhD.
Consultant:	prof. Anja Podlesek
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	doc. RNDr. Tatiana Jajcayová, PhD.
Assigned:	08.03.2023
Approved:	08.03.2023

prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Michaela Dlugošová
Študijný program: kognitívna veda (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Development and validation of Restaurant Game for measuring and training working memory
Vývoj a validácia videohry na meranie a tréning exekutívnych funkcií

Anotácia: Exekutívne funkcie prirodzene degradujú počas našich životov, čím znižujú kvalitu života. Tento pokles môžeme sledovať a odhadovať pomocou meraní a mitigovať jeho dopad kognitívnym tréningom. Obe tieto metódy majú rôzne nevýhody ako napríklad nedostatok flexibility, obmedzenú ekologickú validitu, či cenu. Keď vytvárame novú metódu, ktorá adresuje nedostatky aktuálnych riešení na meranie a tréning, je dobré začať s pracovnou pamäťou, ktorá má veľký vplyv na výkon exekutívnych funkcií a ďalších kognitívnych funkcií vyššieho rádu. Prvý krok pri validácii novej metódy na tréning je porovnanie jej výsledkov s meraniami, ktoré sú aktuálne používané na meranie daného konceptu, napríklad tzv. Digit span task. Dôvod je, že trénovacia metóda musí v prvom rade reflektovať koncept, ktorý má trénovať. Z tohto dôvodu je dôležité počas vývoja nového tréningu zobrať do úvahy aj nedostatky metód používaných na meranie exekutívnych funkcií.

Cieľ: Vyvinúť a overiť videohru zameranú na pracovnú pamäť na základe aktuálnych teoretických poznatkov o meraniach a tréningoch, porovnať ju s existujúcimi meracími metódami a vyhodnotiť jej účinnosť.

Vedúci: Mgr. Martin Marko, PhD.
Konzultant: prof. Anja Podlesek
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 08.03.2023

Dátum schválenia: 08.03.2023

prof. Ing. Igor Farkaš, Dr.
garant študijného programu

.....
študent

.....
vedúci práce

Acknowledgments: I would like to thank my supervisors for their time and ideas. I am grateful they shared their expertise with me so that I could learn and proceed much faster. I also want to thank the professors at the Cognitive Science Department for accepting me into the program. Your program is amazing, practical, and so different from other degrees in Slovakia - in a good way. I loved my time there and always speak of it as the best decision I made. Last but not least, I am thankful for the patience of my family, to whom I cried about this thesis on several occasions. I will try to be a better companion from now on!

Abstract

Executive functions are crucial for our voluntary behavior and highly influence the quality of our lives. Their poor performance is connected with different cognitive impairments, mental disorders, or natural decline of functions due to aging. The current options assessing the decline bear several disadvantages, such as boredom, lack of research flexibility, lack of ecological validity, or high cost. We developed a new task, Restaurant Game, that limits the disadvantages aimes to reflect one of the executive functions, working memory. The development decisions were driven by research needs and user feedback gathered in UX testing. After preparing the game for our research, we aimed to do a pilot validation study for convergent and divergent validity, test-retest reliability, and user experience. In the first part of the study, participants ($N = 62$) completed computerized and computer tasks. Two tasks were more similar in nature to the game (Digit span task (backward) and Corsi task (backward)), one task was different (Shape-filling task), and the last was the Restaurant Game itself. The participants were divided into two groups that differed in the placement of the game after ($N = 28$) or before ($N = 34$) the other tasks. They also reported on user experience metrics for each task, specifically engagement and motivation. After approximately a month, the participants were asked to play the Restaurant Game again, and approximately half of them did ($N = 33$). Both parts of the study were explained in detail in step-by-step instructions that were sent to the participants who completed the experiment in their free time. There were no significant differences in the two groups' scores, so the data were further analyzed regardless of the initial division. The analyses showed a significant moderate correlation between the results of the Digit span task (backward) and the Restaurant Game, supporting the convergent validity of the latter. The correlation between Restaurant Game and Corsi task (backward) did not reach significance. A negligible correlation between the results of the Restaurant Game and the Shape-filling task indicated the divergent validity of the game. Participants felt significantly more engaged when playing the Restaurant Game than when doing the Corsi task (backward). Additionally, they felt significantly more motivated in the Restaurant Game than in any of the other tasks. Lastly, the game results between the two parts of the study were not statistically significantly correlated, indicating a poor retest reliability of the game. Possible interpretations of the results and several directions for further game development and research are discussed.

Keywords: working memory, cognitive assessment, serious games, user experience, validation

Abstrakt

Exekutívne funkcie sú kľúčové pre naše kontrolované správanie a výrazne ovplyvňujú kvalitu našich životov. Ich nízky výkon súvisí s rôznymi kognitívnymi poruchami a duševnými ochoreniami, ale taktiež prirodzene klesá v dôsledku starnutia. Súčasné metódy zaoberajúce sa týmto poklesom majú niekoľko nevýhod, ako napríklad že sú nudné, neposkytujú dostatočnú flexibilitu pre výskum, nemajú dostatočnú ekologickú validitu alebo majú vysokú cenu. Vyvinuli sme novú metódu Restaurant Game, ktorá limituje tieto nevýhody a sústredí sa jednu z exekutívnych funkcií, konkrétne pracovnú pamäť. Rozhodnutia v procese vývoja tejto metódy sme robili na základe výskumných potrieb a spätnej väzby od používateľov, ktorú sme získali počas UX testovania. Naším cieľom bolo uskutočniť pilotnú štúdiu konvergentnej a divergentnej validity, spoľahlivosti a používateľskej skúsenosti. V prvej časti štúdie účastníci ($N = 62$) absolvovali testy a hru na počítači. Robili dva testy, ktoré sa svojím charakterom viac podobajú našej hre (Digit span (spätná vezia) a Corsi (spätná verzia)), jeden test, ktorý je rozdielny od hry (Shape-filling), a samotnú hru Restaurant Game. Účastníci boli rozdelení do dvoch skupín, ktoré sa líšili umiestnením hry po ($N = 28$) alebo pred ($N = 34$) ostatnými testami. Pri každom teste hodnotili aj svoju skúsenosť prostredníctvom ukazovateľov používateľskej skúsenosti, konkrétne angažovanosti a motivácie. Zhruba po mesiaci boli účastníci požiadaní, aby si hru Restaurant Game zahrali znova, a približne polovica z nich tak urobila ($N = 33$). Priebeh oboch častí štúdie bol podrobne vysvetlený v pokynoch, ktoré sme zaslali účastníkom. Účastníci experiment absolvovali vo svojom voľnom čase. Vo výsledkoch skupín neboli signifikantné rozdiely a údaje boli ďalej analyzované bez ohľadu na ich počiatočné rozdelenie. Analýza ukázala signifikantnú miernu koreláciu medzi výsledkami testu Digit span (spätná verzia) a hry Restaurant Game, čo podporuje konvergentnú validitu tejto hry. Korelácia medzi Restaurant Game a Corsi testom (spätná verzia) nebola štatisticky signifikantná. Zanedbateľná korelácia medzi výsledkami Restaurant Game a testom Shape-filling naznačuje divergentnú validitu hry. Účastníci sa cítili signifikantne viac angažovaní v rámci hry Restaurant Game ako v teste Corsi (spätná verzia). Okrem toho sa cítili signifikantne viac motivovaní v Restaurant Game než v ktoromkoľvek inom teste. V neposlednom rade, výsledky hry medzi jednotlivými časťami štúdie neboli štatisticky signifikantne korelované, čo naznačuje slabú spoľahlivosť hry. V závere práce diskutujeme o možných interpretáciách výsledkov a niekoľkých smeroch ďalšieho vývoja a výskumu hry.

Kľúčové slová: pracovná pamäť, kognitívne hodnotenie, seriózne hry, používateľská skúsenosť, validácia

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Executive functions highly influence the quality of our lives, and their poor performance is associated with various cognitive impairments and mental disorders. Training these functions can delay the onset of the decline or mitigate its negative effects.

We explore current training options, focusing on gamified formats and games, as these often show very promising results. Game-like elements increase motivation and engagement and allow for more ecologically valid tasks. However, existing cognitive training games have several disadvantages, such as cost and lack of research support.

Our ultimate goal is to develop a new training game that aligns with the latest research and meets the needs of researchers and users. This training will have several parts, each focusing on different executive functions. Prior to this thesis, we attempted to develop the part focusing on working memory, which we named the Restaurant Game due to its restaurant setting. The game, however, did not reflect working memory.

The first part of this thesis focuses on preparing Restaurant Game for its pilot validation study. We make necessary adjustments based on the user experience testing and current research knowledge on working memory. After developing a more suitable version, we proceed with the validation study. We test convergent and divergent validity using tasks that are similar and different in nature to the Restaurant Game. Additionally, we gather participant feedback on their engagement and motivation within each task. We repeat the measurement with the game to test its test-retest reliability.

1.1 Executive functions in cognitive psychology

Executive functions are an encapsulating term referring to several top-down cognitive processes essential for goal-directed and non-routine behavior and thoughts. They are crucial for dealing with novelty and adapting to changing situations, solving problems, planning and effectively carrying out those plans, and sorting information based on its relevance to the goal. Moreover, they help us regulate our emotions and impulsive re-

actions, which is important for our social behavior. In summary, they are fundamental to our lives. (Anderson, 2002; Banich, 2009; Diamond, 2013; Jurado & Rosselli, 2007)

Despite being extensively studied and discussed, no formal definition of executive functions exists. They encapsulate a wide range of processes, and there is no consensus on which mental functions are executive and which are not. Studies trying to clarify the topic often yield contradictory results, and currently, there is no definite way to decide which results are more reliable. (Barkley, 2012; Jurado & Rosselli, 2007)

Even though we lack consensus regarding the precise nature of executive functions, we find three separated but related components commonly mentioned across literature: updating working memory, shifting, and inhibition (Best & Miller, 2010; Davidson et al., 2006; Diamond, 2013; Miyake & Friedman, 2012; Miyake et al., 2000). These three are sometimes referred to as basic or core ones. Best and Miller (2010) even named them as foundational components of executive functions, which speaks of their importance. Through their combinations, we get higher-order executive functions such as planning, problem-solving, or decision-making. (Diamond, 2013; Lunt et al., 2012)

The first of the basic executive functions is updating working memory. It enables us to monitor and manipulate the working memory content (Miyake & Friedman, 2012). When naming executive functions, updating working memory is often referred to simply as working memory. In that case, the authors also include the storage aspect, i.e., the ability to store a limited amount of information for a limited time. We discuss this discrepancy more at the end of Section 1.2. Regardless of the term used, this executive function is crucial for our ability to reason and make decisions. (Diamond, 2013)

The second executive function is shifting. It refers to our ability to shift our attentional focus to a different task, multitask, adapt to new environment requirements, take a different approach from the previous one, and be creative. Sometimes, it is called cognitive flexibility. (Diamond, 2013)

The third one is inhibition. It enables us to suppress irrelevant stimuli, override our automatic reactions and reactions that do not lead to our goal, and resist temptations. Eventually, we can choose and control ourselves thanks to inhibition. (Diamond, 2013)

Despite the complex nature of the phenomenon and the lack of a precise definition, research agrees that executive functions are crucial for good quality of life (Brown & Landgraf, 2010; Davis et al., 2010; Diamond, 2013). Executive functions are mainly associated with the functionality of frontal lobes, in particular the prefrontal cortex (Barkley, 2012). The prefrontal cortex is a brain region known to keep developing until adulthood, and the same is true for executive functions, which undergo different developmental stages with varying maturity levels. (Anderson, 2002; Diamond, 2013; Fuster, 2002; Reynolds & MacNeill Horton Jr, 2008; Stuss, 1992)

On the other hand, when we are adults, aging negatively affects the whole brain, including the frontal lobes, and therefore the executive functions (Barkley, 2012). This

decline is attributed to anatomical changes and slower information processing in the brain that accompany the aging process and changes in the interconnectivity of brain regions (Anderson, 2002; Fisk & Warr, 1996; Jurado & Rosselli, 2007; Raz et al., 2005).

Consequently, there is a decline in performance of executive functions among elderly, which is associated with mental disorders such as depression, dementia, Alzheimer's disease, and others (DeBattista, 2005; Diamond, 2005; Guarino et al., 2019; Rock et al., 2014; Willcutt et al., 2005). These impairments impact one's independence, emphasizing the importance of understanding and appropriately addressing the cognitive changes, especially in our aging population (Jurado & Rosselli, 2007; West, 1996).

Clinical neuropsychology needs valid assessments of executive functions to monitor their age-related decline and their development in various cognitive impairments, brain diseases, and injuries. These assessments are essential for accurate diagnosis, establishing effective treatment strategies mitigating the negative impact of the decline on individuals' lives, and may even help to prevent or postpone the onset of the decline (Karbach & Verhaeghen, 2014; Pennington & Ozonoff, 1996; Shah et al., 2017). From an academic perspective, these measuring and training methods are important for understanding the executive functions and cognition itself.

We proceed to discuss the practical aspects of measuring and training executive functions in Sections 1.1.1 and 1.1.2. Then, we establish the theoretical background and outline several models of executive functions in Section 1.2. We narrow our focus to the chosen executive function, working memory, in Section 1.2.1 and introduce the current assessment methods associated with it in Section 1.2.2. Finally, in Section 1.3, we introduce our task for measuring and training working memory, Restaurant Game, whose validity is the topic of this thesis.

1.1.1 Measuring executive functions

Executive functions are not a single mechanism, therefore, we need different measuring and training methods implied by each function. Every training method has to reflect the concept it is supposed to improve, and if it is not used repeatedly, we could say that this method can be used as measurement. Additionally, when validating a new training method, the first step is to check if its results align with already used assessments.

The tasks measuring executive functions were previously used with patients with frontal lobe dysfunction, as the phenomena date back to neuropsychological studies on frontal impairments and injuries. In these tasks, people without any frontal lobe issues and patients with frontal lobe dysfunction have statistically significant differences in performance. These consistent differences prove that the frontal lobes, therefore executive functions, are involved in cognitive processes needed to perform a given task. (Burgess & Shallice, 1996a, 1996b; Jurado & Rosselli, 2007)

Potentially due to this historical reason, commonly, the only criterion to validate a new task is to check if it is a marker for frontal lobe dysfunction (Jurado & Rosselli, 2007; Miyake et al., 2000). This approach is problematic as these patients are not commonly approachable for such studies, and in the case of older people, whose executive functions naturally decline with age, the results may be skewed.

The next issue with the measuring tasks is the task-impurity problem. Currently, it is not possible to measure any of the functions by themselves as participants also need to use context-specific function-irrelevant processes (e.g., color or language-processing, sustained attention) (Miyake & Friedman, 2012; Snyder et al., 2015). A latent variable analysis was used by (Miyake et al., 2000) to reduce the effect of the task-impurity problem and get better results. However, this analysis is not used on individual tasks, where the task-impurity problem persists and should be considered. On the other hand, in real life, the executive functions are always used within a present context and not as self-standing processes. Therefore, it is up for discussion if we should even aim to tackle the task-impurity problem at the expense of the ecological validity of the tasks.

From the participant's point of view, standardized tasks are generally considered boring, not interesting to do, and may even result in negative emotions (Lumsden et al., 2016). Moreover, executive functions require novelty as they guide our actions when we can not rely on automatic processes (Anderson, 2002; Chan et al., 2008). However, novelty is individual, can not be assumed for all the participants, and usually wears off after we repeat the task (Alexander & Stuss, 2000). This are the reasons for common low test-retest reliabilities in tasks measuring executive functions (Rabbitt, 2004).

We lack ecologically valid assessments, which means we can not draw conclusions from the findings and apply them to real-world situations (Burgess et al., 1998; Eslinger & Damasio, 1985; Manchester et al., 2004; Sbordone, 1996; Shallice & Burgess, 1991). The use of executive functions in standardized tasks and real-world situations differ, and the task results do not tell us a lot about all the deficits the patients are to experience in their lives (Burgess et al., 2006; Chan et al., 2008; Diamond, 2013). That being said, the lack of ecologically valid measurements implies that we have to keep using the available tasks, consider all their problems, and try to develop new, more satisfactory measures. (Doebel, 2020; Jurado & Rosselli, 2007)

Despite numerous problems with standardized tasks for measuring executive functions, their importance is undeniable. Additionally, they lay the groundwork for developing cognitive training methods focused on improving the functions' performance. Such interventions can help mitigate the effects of aging on the functions and maintain a good life quality which will gain importance due to the increasing age average of the human population. To conclude, taking into account the aforementioned shortcomings when designing new tasks for measuring and training executive functions holds much potential not only for research purposes but also for improving people's lives.

1.1.2 Cognitive training

Our ultimate goal is to develop a new training method for executive functions. To do that, we need to understand the advantages, disadvantages, and shortcomings of the current training option but also of current measuring options that lay the groundwork for cognitive training. Based on that, we can develop a new and better method.

Cognitive training shows a big potential in improving the performance of executive functions (Belleville, 2008; Butler et al., 2018; Jaeggi et al., 2011; Levine et al., 2011; Verghese et al., 2006). Metaphorically, Diamond (2013) compares cognitive training to physical exercising. In the same way as we build our muscles by regularly challenging them, we can improve our executive functions when regularly training them.

There is an ongoing discussion on the transferability of the training to executive functions that were not part of the training. Transferability is crucial for cognitive training, as it suggests the possibility of improving several cognitive processes at once. Otherwise, we would need to train our functions on many more tasks to achieve the same results. On one hand, there are studies showing non-significant changes in other tasks, which limits the potential of cognitive training (Butler et al., 2018; Owen et al., 2010; Sala et al., 2019; Sala & Gobet, 2020). On the other hand, there are studies supporting the transferability (Brehmer et al., 2012; Buschkuehl et al., 2008; Holmes et al., 2009; Karbach & Kray, 2009; Karbach & Verhaeghen, 2014; Klingberg, 2010).

A follow-up question is whether a near or far transfer is possible. In the former, the performance improves only for tasks closely related to the one used during training. In the latter, the performance improves also for untrained functions. The current evidence supports the near-transfer, which still suggests that training chosen executive function with one task should improve its performance in other tasks on that function (Diamond & Ling, 2016; Karbach & Kray, 2016; Melby-Lervåg et al., 2016).

With technological progress, the domain of cognitive training is changing, and technology can play a crucial role in reinventing methodologies for assessing and training executive functions. The advantages of technology are computational power, automatization, precision in measurements, flexibility, collection of larger amounts of data, and requirements for fewer human resources (Kueider et al., 2012). On top of that, many studies show promising results in cognitive training done on computer (e.g., Bond et al., 2001; Klingberg et al., 2005; Nouchi et al., 2012; Nouchi et al., 2013; Shahmoradi et al., 2022; Toril et al., 2014; Wei et al., 2022).

It is important to point out that we see a difference between computerized and computer tasks. Computerized tasks are the same as "on paper" ones, just done on a computer. They have many of the original disadvantages, such as a lack of ecological validity. Computer tasks are designed with technology as an inseparable part. We focus on the computer tasks that were already designed specifically for and with computers.

A big step towards designing computer tasks suitable for cognitive training is gamification. Gamification means using game-like principles (e.g., goals and points) to avoid boredom, increase engagement and motivation, and eventually improve the results (Deterding et al., 2011; Prins et al., 2011). According to Hawkins et al. (2013) and more recently Scharinger et al. (2023), gamification does not change the data results but it does make the experience better. Despite the studies having small samples, they are among the few studying gamification's influence on psychologically relevant results.

Even though the research on gamification is scarce, Lumsden et al. (2016) provides a great overview of gamified cognitive assessment and mini-games used in research between January 2007 and October 2015. Its results show that gamification was not used much in research, as only 31 gamified tasks were used in those eight years. Moreover, when looking at a more recent review by Koivisto and Malik (2021) focused on gamification for older adults, there are only 12 relevant studies. These results suggest that the domain of gamification within research, especially in interventions for older adults, is not sufficiently covered. A little attention has been paid to this topic by the research community, which is why it is a bit more difficult for us to rely on previous findings. Even though the usage of gamification grows every year, it is more commonly used in business than in academia (Nacke & Deterding, 2017).

Instead of using gamified tasks, we can use games themselves. The disadvantage is that developing a game is much more difficult. However, the advantages are that games make it possible to design real-world environments and create tasks more natural for using executive functions. This is a possibility to improve the ecological validity of the tasks regardless of their purpose. We decided to develop our new method in game format, as it brings all the advantages of technology, gamification, and even some more.

When games are designed primarily with another objective than entertainment, they are called serious games. In the case relevant to us, the objective would be to improve the performance of executive functions and fulfill the function of cognitive training. The requirements on the tech savviness of the players are commonly questioned, but it has been shown that no special skills are necessary for serious games (Kueider et al., 2012). That does not mean we should not consider the digital literacy of our population. Kueider et al. (2012) just pointed out that technical skills are not of the utmost importance. Additionally, even though the evidence on serious games is not very robust yet, current research shows promising results in the effectiveness of serious games for all ages (Nouchi et al., 2012; Nouchi et al., 2013)

The technological revolution underlined the importance of the user experience. Incorporating it in the game design process can make the game more attractive, enjoyable, and easy to use which also means minimizing the requirements on technical skills of the players (Toril et al., 2014). The result of increasing the players' engagement can be an increase in their inner motivation and thus performance but also an increase in the

probability of repeating the training (Lumsden et al., 2016; Milyavskaya et al., 2015; Ninaus et al., 2015; Procci et al., 2012; Vermeir et al., 2020). These advantages lack in traditional training, eventually decreasing their efficacy compared to game formats (Kueider et al., 2012). We discuss user experience in more detail in Chapter 2.

Currently, several organizations design serious games, e.g., LumosLabs[1], BrainHQ[2], HappyNeuron[3], CogniFit[4], CogMed[5], and Elevate[6], that is not even focused on executive functions. However, they all have several disadvantages. We start with the cost which in CogMed may reach 1500-2000$/program, and 14$/month on average in the other organizations as they are based on monthly subscriptions. Most organizations are commercial with strict legal rights (LumosLabs, Elevate, BrainHQ, CogniFit). We found one non-commercial option, which is a self-standing custom-designed multitasking video game called NeuroRacer by Anguera et al. (2013)). The game shows promising results and is available for download, however, we could not access its code.

The issue also lies in the way in which such cognitive training is commonly developed. Usually, it is a commercial product with no proper research background or with researchers developing a product based on theoretical models and not taking into account user validation. An example of the first can be HappyNeuron, which lacks validation by the scientific community. An example of the second can be CogMed, which is considered to be the most common working memory training program but without end-user validation. (Marcelle et al., 2018; Shah et al., 2017)

What all of the mentioned have in common is that their games are closed systems and lack the flexibility needed for the variance of directions in research. The parameters of the games are set in advance, which means if the games were to be used in research, their conditions could not be changed to test the hypothesis in question. Moreover, the commercial solutions output only the final score, not partial results or other psychologically interesting measures such as error rate and response time.

To conclude, cognitive training is more and more leaning toward technological implementations. It brings some significant benefits, which are closely linked to using gamified elements or games themselves. Unfortunately, currently available options for serious games are inappropriate from the research point of view and have several other disadvantages. Due to that, we decided to design our own computer training. We continue to discuss the theoretical models of executive functions, choose one to guide our development and choose one of the functions to start with.

[1]https://www.lumoslabs.com/
[2]https://www.brainhq.com/
[3]https://www.happy-neuron.com/
[4]https://www.cognifit.com/
[5]https://www.cogmed.com/
[6]https://elevateapp.com/

1.2 Models of executive functions

A multitude of models try to explain and conceptualize executive functions and identify the functions' underlying mechanisms, core components, and the connections among the components. However, due to the complexity of brain functioning and the limited knowledge in neuroscience and psychology, it is not yet possible to determine which model is more appropriate and should be generally accepted. The purpose of this section is not to provide an exhaustive review but to complement the topic of executive functions from a more structured and theoretical perspective.

As pointed out by Anderson (2002), the reason for having an appropriate conceptual model lies in establishing adequate assessments, correctly interpreting results, and developing effective cognitive training. The model provides a needed theoretical framework to guide the design, development, and analysis of new methods.

Chronologically, Luria (1973) was the first to try to conceptualize the functioning of the frontal lobes and executive functions (even though the term was not used yet). Their model takes the physical structure of the brain and divides it based on distinctive functions into three interconnected and concurrently functioning brain units.

The first unit is located in the brain stem and regulates the arousal of the cortex making it responsible for states of alertness and vigilance. The second one is for receiving, processing, and storing sensory information from the outside world and consists of sensory regions across the temporal, parietal, and occipital lobes. The last unit is for programming, regulating, and evaluating behavior and mental activity and is located in frontal lobes with the prefrontal cortex as a crucial regulatory part. (Luria, 1973)

Next, Norman and Shallice (1986) proposed a model referred to as the "SAS" model that provides more detail for the third brain unit proposed by Luria (1973). As noted by Norman and Shallice (1986), their model focuses on the attention needed for the automatic and controlled processes.

The model consists of two components: contention scheduling and supervisory attentional system (SAS). The first component is responsible for the activation and functioning of processes that do not require our attention as they are habitual and part of our routine. The second is responsible for the conscious processes in novel situations requiring our directed attention and also for controlling the contention scheduling component. (Norman & Shallice, 1986)

Another model proposed by Baddeley and Hitch (1974) conceptualizes working memory, and even though it is not directly focused on executive functions, it does show their relationship with memory. While the first version was published before the SAS model by Norman and Shallice (1986), these models were later combined.

According to Baddeley and Hitch (1974), the working memory consists of three components: phonological loop, visuo-spatial sketchpad, and central executive. The

phonological loop and visuo-spatial sketchpad enable the storage maintenance of audio (verbal) and visuo-spatial information, respectively. The central executive represents an attentional control of actions, and as there was little known about attention under given circumstances, the SAS model was incorporated into this component.

Baddeley (1996) outlines four functions of the central executive: attending two tasks simultaneously (multitasking), switching between tasks, focusing attention while inhibiting irrelevant stimuli, and holding and manipulating information in long-term memory. These characteristics are aligned with those of executive functions, and therefore, the central executive component is believed to represent executive functions.

Subsequently, a fourth component was added to chronologically integrate multidimensional information. This new component, an episodic buffer, linked the previous three components with long-term memory and semantics (Baddeley, 2000).

In a later work, the author mentioned the model of working memory (WM) was still not complete: "My overall view of WM therefore comprised, and still comprises, a relatively loose theoretical framework rather than a precise model that allows specific predictions" (Baddeley, 2012, p.7). Despite that, the model could integrate and explain many previous and new findings, making it one of the most common and influential models in the domain of executive functions.

A different approach to executive functions is via a unity/diversity framework studying correlations between executive functions. This approach can provide insight into the functional structure of the functions and answer questions like whether executive functions are a construct with a single mechanism, if they can be divided into multiple components, if the components are related, and if yes, then how. (Miyake et al., 2000)

Originally, Miyake et al. (2000) tried to address a task-impurity problem. The core issue is that the specific executive functions are not assessed as self-standing constructs because they are always put into context. The results of such tasks include aspects specific to the studied function but also commonalities across executive functions, other processes, and error, which makes them in some sense impure (Snyder et al., 2015). Consequently, when we compare two tasks that measure the same executive function, they might have different results and be weakly correlated (Friedman & Miyake, 2017).

To tackle the task-impurity problem and get more accurate measures, we can opt for latent variable analysis (Friedman & Miyake, 2017). It is used when we are aware of a hidden variable that can not be directly observed, but we have some assumptions about it. Based on these assumptions, we can infer the latent variable. Specifically, an example of such a variable may be the common aspect among the tasks for a chosen executive function that better reflects the function itself.

Miyake et al. (2000) chose shifting, inhibition, and updating of working memory for their study on normal individual differences. They found out these functions are correlated and can be decomposed into a part common for all three functions (unity)

and parts that are function-specific for each function (diversity). Interestingly, they also found that inhibition lacked the function-specific part and was almost perfectly correlated with the common part.

Several studies came to the same conclusions when studying correlations between the same but also different executive functions (e.g., Brydges et al., 2014; Fisk & Sharp, 2004; Fournier-Vicente et al., 2008; Hull et al., 2008; Rose et al., 2012; Saylik et al., 2022; Vaughan & Giovanello, 2010; Willcutt et al., 2001; Zabelina et al., 2019). These results support the view that the central executive component is fractioned into smaller separated but connected parts corresponding to executive functions.

Even though we lack a definition of executive function, their models and frameworks implicitly (as a supporting component) or explicitly (as a self-standing component) assume the incorporation of working memory. Working memory plays a vital role in the performance of executive functions (Engle, 2002; Pennington & Ozonoff, 1996), and its capacity reflects the functions' performance (Engle et al., 1999; Hester & Garavan, 2005). Additionally, the working memory model by Baddeley and Hitch (1974) provides a great theoretical framework to hold onto. Therefore, we decided to focus on the working memory and elaborate more on the concept in Section 1.2.1.

Before we continue, we need to address potential terminology misunderstandings. The term "working memory" is used in different ways in the literature. The confusion arises from the fact that Baddeley and Hitch (1974) define working memory as a memory unit including a central executive component with the same characteristics as executive functions, and at the same time, one of the main executive functions is commonly referred to as working memory (Best & Miller, 2010; Davidson et al., 2006; Diamond, 2013). This creates a loop in explanations. As the aim of this thesis is not to settle this discrepancy, we state how we use the terms to avoid further confusion.

When Baddeley and Hitch (1974) discusses working memory, they refer to the storage components along with the central executive component. The former is responsible for maintaining the information in an accessible state after the stimulus is no longer present, and the latter does the work with and on the stored information.

The central executive component is considered to be the equivalent of the executive function. One of the basic executive functions is updating, and we continue to use the term similarly to Miyake et al. (2000) when talking about the function responsible for adding, manipulating, and pushing out the information of working memory storage. In this case, updating is not concerned with the storage capacity it works upon.

According to Diamond (2013), working memory is one of the executive functions that maintains and manipulates the stored information. This view adds the storage-related processes as part of the executive function while simplifying the description of working memory by Baddeley and Hitch (1974), as it excludes functions other than updating. It positions updating as an inseparable part of working memory.

Further on, we use the term updating to talk about the information manipulation processes, the term working memory to refer to the storage and its updating, while using the term working memory model to specifically refer to the model proposed by Baddeley and Hitch (1974) which also includes other executive functions.

1.2.1 Working memory model

We can take a look at the working memory from two perspectives relevant to our thesis. The first is the point of view of executive functions, which brings some issues with definitions and is addressed in the previous Section 1.2. The second is the memory point of view and we elaborate more on it in this section.

The first prominent memory model was proposed by Atkinson and Shiffrin (1968) and is referred to as the multi-store or modal model. It distinguishes three separate types of memory: sensory, short-term, and long-term memory. Additionally to the memory units, the model includes the processes describing the flow of information.

First, the information arrives at our senses and gets to the sensory memory. Next, when the information is attenuated, it proceeds to short-term memory. There, the information decays after tens of seconds if not rehearsed. From short-term memory, the information is transferred to long-term memory more or less automatically as long as the information is present in short-term memory. Therefore, the strength of memory depends on the time during which the information was attended to. The information is remembered and stored within long-term memory from where it can be later retrieved back into short-term memory. (Atkinson & Shiffrin, 1968)

For each of the memory components, we can talk about capacity and durability. Even though the exact numbers are still not agreed upon, we can generally say the following. Sensory memory can hold a large amount of information, but only for a very short time (usually less than a second or two). For short-term memory, the duration is generally limited to thirty seconds, but the discussion on the capacity is more interesting. For long-term memory, both of the aspects are considered to be unlimited. (Atkinson & Shiffrin, 1968; Sperling, 1960)

Miller (1956) refers to the short-term memory capacity as a memory span and defines it as the number of items in the longest sequence that can be correctly repeated in at least 50% of trials. According to his research, the capacity for an individual's short-term memory is 7 ± 2 chunks. A chunk is a meaningful group of information independent of other groups. What is and what is not meaningful depends on each individual and thus is too subjective to define. As we can not define how big a chunk can be, we can not generally estimate the exact capacity of short-term memory.

A more recent research by Cowan (2001) argues the capacity is only 4 ± 1 chunks. However, his estimate is made for more strict conditions within a task and better

reflects the capacity of the focus of attention. On the other hand, the estimate by Miller (1956) is more accurate in tasks more similar to the real world when there is no clear distinction of what a chunk is, making it more suitable for our thesis.

The short-term memory within the modal model was depicted as one coherent unit responsible for storing information for a short time period. This view was redefined by Baddeley and Hitch (1974), who defined the unit's structure and explicitly positioned the information manipulation processes related to it. The name of the memory unit was changed from short-term to working memory to reflect its active nature, thus the name working memory model. processes.

Even though the theoretical difference between working memory and short-term memory is defined, they are often used interchangeably (Aben et al., 2012). It may be due to their historical connection as working memory developed from short-term memory or maybe from the fact that there is no consensus on their separation and overlap in empirical studies (Aben et al., 2012; Baddeley, 2012). Their relationship changed over the years and differs among different memory models (Atkinson & Shiffrin, 1968; Baddeley, 1996; Baddeley & Hitch, 1974; Shiffrin, 1976). Further on, we use term short-term memory to talk about temporary passive storage and term working memory to talk about temporary active storage with updating function as an inseparable part.

Baddeley and Hitch (1974) in their model divided the working memory unit into two domain-specific storage components that can be considered short-term memory units as their primary function is to maintain the information. The third component is the central executive whose description matches the description of executive functions (Baddeley, 1996; Baddeley & Hitch, 1974).

The phonological loop is one of the buffers that serve as a short-term storage. As the component's name suggests, the information maintained within the loop is speech-based, such as the list of groceries we need to buy or someone's phone number. The phonological loop can be further divided into two subcomponents with the first being a phonological store that can keep the information for up to two seconds, and the second being an articulatory control process with a rehearsal function prolonging the durability of the information. (Baddeley, 2020)

The next component is the visuo-spatial sketchpad, which has a function similar to that of a phonological loop but is concerned with visual and spatial information. For instance, how something looks like, where it is, and its relative position to our body. As we have two different visual pathways to process visual information, the dorsal and ventral pathways, the sketchpad is divided similarly. It consists of spatial memory, which refers to the where component of visual information, and object memory, which refers to other visual features of the object. (Baddeley, 2020)

Lastly, the central executive is a processing component even though originally it also had storage capability. The central executive is an attentional control system that

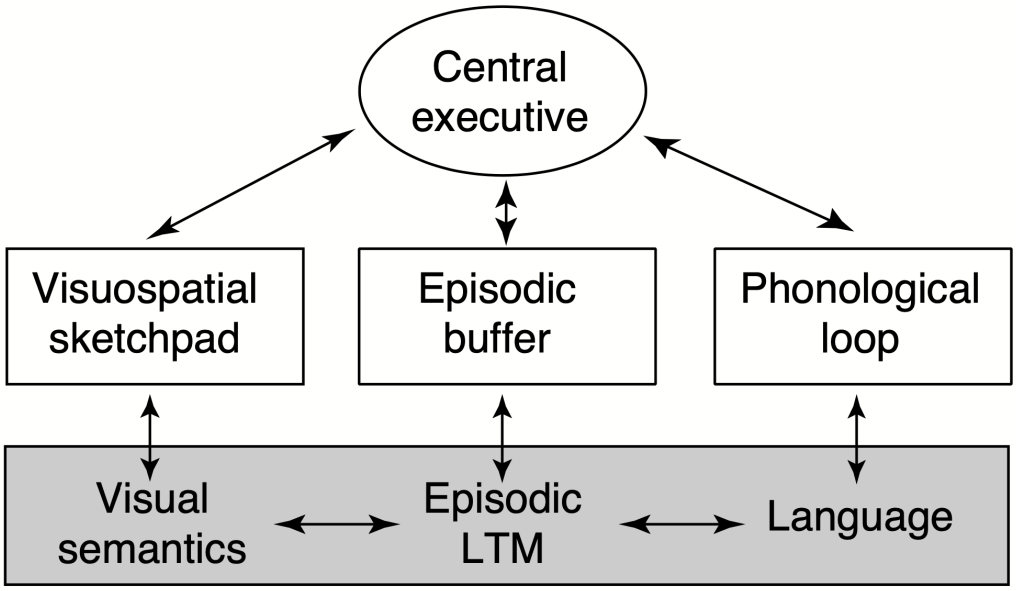

Figure 1.1: The revisited working memory model by Baddeley (2000) supplemented with episodic buffer and therefore consisting of four separate components.

allocates attention which is limited. This component coordinates the activity of the storage components as well and is linked to long-term memory. Even though it is vital, there is not much known about it. (Baddeley, 2020)

Later, Baddeley (2000) added a storage component to separate the central executive from its storage capabilities. The new component was named an episodic buffer and reflected new empirical findings (e.g., for patients with short-term memory deficits). Its function was to bind information from different sources and create coherent episodes as multidimensional representations. Figure 1.1 shows the complete revisited working memory model. (Baddeley, 2000)

One may argue that there are more models depicting working memory. Even though that is true, not all models are easily comparable and easy to use (Adams et al., 2018; Miyake, Shah, et al., 1999). The model proposed by Baddeley and Hitch (1974) is the most common and established one which is why we chose it to serve as a reference further on. It is a well-studied framework for the memory unit with the addition of an executive component, which suits our needs the best.

To briefly summarize, we introduced the concept of working memory also from the memory unit perspective. We differentiated between short-term and working memory and stated the most prominent research regarding its capacities. Further on, we use the working memory model by Baddeley and Hitch (1974) as our theoretical framework reference as it is well-researched and includes both the executive functions and storage units. We follow by discussing current measuring options for each of its components.

1.2.2 Working memory model - assessments

The most common measure of working memory is its capacity, the so-called working memory span. The difference from short-term memory capacity is that we require intentional manipulation of the stored information. The measurements are also called simple and complex span tasks for short-term and working memory, respectively. Complex span tasks have been shown to correlate with other cognitive processes, including other executive functions, reasoning, and fluid intelligence (Aben et al., 2012; Engle et al., 1999; Hester & Garavan, 2005; Kyllonen & Christal, 1990).

We start with the phonological loop. This component of the working memory model by Baddeley and Hitch (1974) stores speech-based information. A real-life example is when we need to remember a list of groceries to buy, but on our way to the store, we change the order of items to match the order of aisles in a store.

The Digit span task is used to assess the span of the phonological loop. It is commonly used within clinical neuropsychology and is part of the Wechsler Adult Intelligence Scale, which is a standard test to measure cognitive abilities. The Digit span task is used to measure short-term memory span, as the participant remembers the order of digits presented to them. The length of the sequence increases after the correct answer until the participant is no longer able to remember the digits correctly. The number of successfully remembered digits is the span. To measure working memory, the backward version of this task is better suited as the participant has to repeat the list of the digits in reversed order. (Caplan et al., 2010)

The visuo-spatial sketchpad is a storage of visual and spatial information. A common measure to assess its span is a Corsi block-tapping task (Corsi, 1972). This task is similar to the Digit span task, but instead of remembering digits, the participant remembers a sequence of tapped blocks. To have a more suitable version reflecting working memory capacities, we again need to use its backward version.

To assess the central executive, different measurements are used for different executive functions. The tasks differ in complexity and difficulty depending on the specific function(s) they are supposed to reflect. Here, we only focus on the three basic functions, namely shifting, updating, and inhibition.

Shifting enables us to change the task we are currently doing, in other words, to shift our focus of attention to something else. We usually talk about doing several tasks "at once," which is called multitasking. The tasks commonly present a stimulus to which the participant has to respond based on one of the rules.

An example is the number-letter task by Rogers and Monsell (1995), which was also used by Miyake et al. (2000) as part of their latent variable analysis. This task shows stimuli consisting of a pair of a letter and a number which is shown in an upper or lower part of the screen. The location determines the rule by which the stimuli

are evaluated. The first rule is concerned with the letter, whether it is a vowel or a consonant. The second rule is concerned with the number, whether it is even or odd.

What is typically measured is an error rate or switch cost. The latter can be calculated differently, for instance, as a difference or ratio in response times between pure and mixed blocks or task-switch and task-repeat trials. The pure block consists of presenting the stimuli only on one part of the screen, which means that only one rule determines the correct answer. In the mixed block, the stimulus can occur in either part of the screen. The task-switch trial has a rule different from the rule used in the preceding trial, and the task-repeat trial has the same rule.

Updating is about monitoring and changing the information that is currently stored for new ones. However, it is not directly linked to measuring the memory span.

An example is a letter memory task by Morris and Jones (1990) adapted by Miyake et al. (2000). The task shows letters one by one, and the participant has to repeat the last n of them. Another example is the n-back task first introduced by Kirchner (1958). The difference from the previous task is that the participant for each letter decides whether that letter was present exactly n letters back. What is measured is precision, which is calculated based on the number of correct trials.

Inhibition enables us to suppress irrelevant information for current tasks, while this irrelevant information usually comes into our attention more naturally.

A widely known example is the Stroop task first proposed by Stroop (1935) and also used in the analysis by Miyake et al. (2000). Within the task, the participant is presented with several names of the basic colors, and each word is written with colored ink. The inhibition happens when we need to name the color of the ink which is different from the color name written by it. For example, saying red when we are presented with the word green written in red ink would be more difficult than saying green in the same case, as our standard behavior is usually reading the word itself. What is measured is the time it takes us to complete the task and the error rate.

In this section, we presented several tasks used for measuring different components of the working memory model, except for the episodic buffer, as there is insufficient understanding of underlying processes necessary to establish proper assessment (Nobre et al., 2013). These theoretical findings will guide the further development of our new method to make it reflect the working memory.

1.3 Restaurant Game

Being able to measure and train executive functions is very important for both clinical and research purposes. Unfortunately, standard methods have several shortcomings. Therefore, we found ourselves in need of a new task that addresses them.

Having a non-commercial, accessible measurement that provides an opportunity to change its parameters gives flexibility to researchers in testing a variety of hypotheses and measuring different parameters (e.g., response time, number of errors). The research interest is commonly forgotten as it is not the main purpose of serious games. A possible explanation for this is domain expertise. Researchers are not often familiar and skilled with game development, and game developers do not see the scientific and experimental perspectives (Procci et al., 2012).

We aimed at designing a task within an environment, that allows designing different scenarios. These scenarios have to be naturally set within the environment and tackle different executive functions. Moreover, as we did not only aim to create a measuring task but a task with the potential to become cognitive training, we opted for the game format. This format shows promising results in engagement and motivation, which we perceive as important aspects of long-term training.

We wanted to create a game that brings a new way of measuring into the research field, ideally a more ecologic one. We searched for a modifiable environment in which we could naturally place various tasks targeting different executive functions. The restaurant setting fulfilled that, as there was potential for working memory tasks (remembering and assembling orders), switching tasks (while water for tea is boiling, we start preparing a hot dog), inhibition tasks (adding distractions from customers), and planning tasks (creating a path to deliver orders throughout a town in the shortest time). On top of that, these tasks could be made more difficult, e.g., by adding more customers. Moreover, the tasks could be combined, resulting in increased difficulty.

We started by focusing on working memory, as it was shown to play a very important role among executive functions. We developed a Restaurant Game, which, before this thesis, was more likely to reflect short-term memory instead of working memory.

The game is placed in the environment of a restaurant. There is one customer to serve displayed as an animated woman in the middle of the screen. On the left part of the screen is a panel with food and drink options illustrated by black and white icons from Flaticon[7]. The game starts when the application file is opened, and the first order is placed within a few milliseconds. The game environment is shown in Figure 1.2.

The customer places an order, which is shown to her right as text in a rectangle (we refer to it as a speech rectangle). The number of items is randomly generated from one to five, and the specific items are also randomly generated. The order consisting of item names is shown for n seconds, where n is the order size. After that, the speech rectangle disappears. While the order is shown, the panel with the icons is not visible.

The player serves the customer by placing the ordered items on a tray in the bottom center. The items do not have to be put on a tray in any specific order. After the player

[7]https://www.flaticon.com/

Figure 1.2: The Restaurant Game in its original version, with black-and-white icons, one customer with a tray in the center, a button to evaluate assembled orders, a countdown, and a scoreboard.

drags and drops the items into the tray, they press the "DONE" button to evaluate the order. They receive feedback in the speech rectangle saying whether their answer was correct or not. If the order is correct, they also get 2^*n points where n equals the order size, and next order is placed. If the order is incorrect, the player has one more try to assemble it again. In the second attempt, the correct answer results in n points, and the incorrect one in 0. Either way, after the second attempt a new order is placed.

Overall, the whole game lasts two minutes in which the player is trying to score as many points as possible. At the bottom right, the board shows the remaining time and the current number of points. After two minutes, the final score is announced.

The game was developed in Unity, which is the most common game engine. It is a game development framework with a free version for students and projects without high revenue. Unity is used to create the graphical part of the game. For the behavior of the objects and the game itself, we used the object-oriented scripting language C#.

The original source code is available for download on the link: https://davinci.fmph.uniba.sk/~dlugosova24/DiplomaThesis_RestaurantGame/UXTesting/%5bCODE%5dRG_ShortTermMemory_randomOrder.zip.

We believed this version of the game was not appropriate for measuring working memory. As player was not required to manipulate information in any way, the game was more likely to reflect short-term memory. However, there was potential to change the game to reflect working memory, and these adjustments are part of this thesis.

1.4 The aim of the present research

Overall, we aim to develop a cognitive task addressing the shortcomings of the available options that would enable us to add different modules to it. We started by focusing on working memory, and here, we aim to develop and validate a task measuring it.

The previously developed Restaurant Game in the format described in Section 1.3 is more likely to reflect short-term memory. We aimed to make appropriate changes based on the theoretical background to make the game reflect working memory. Additionally, as current options do not provide proof of making users part of the design process, we intended to do so as their experience is important in the long term. In conclusion, this thesis aims to make the game better in terms of alignment with current research findings and user feedback. After that, we conduct a pilot validation study on the psychometric characteristics (validity, reliability) and experience of the game.

1.5 Research questions and hypotheses

In this thesis, we are interested in the psychometric characteristics, specifically in validity and reliability, of the developed and adjusted Restaurant Game as a method for measuring working memory. Additionally, we are interested in its user experience.

We expect the game results to be positively correlated with the results of a task measuring working memory and to have a negligible correlation with a task measuring a different executive function. Regarding the test-retest reliability, we expect the results of two measurements to be correlated across participants. Lastly, we assume the experience will be better with the game compared to the classic assessments.

Chapter 2

Restaurant Game redesign

In this thesis, we aim to conduct a pilot study to test the validity and reliability of a newly developed game called Restaurant Game. The original version of the Restaurant Game can be considered a minimum viable product or more of a prototype. It serves as a good starting point, a skeleton of the final product. In this chapter, we describe all the changes that were necessary to prepare the game for the pilot study of this thesis.

The changes can be divided into three categories: user experience, research, and general code polishing. Before we get to the specifics, it is important to note that several changes were made based on requirements from more than just one category.

The first category is about the user experience which aims to improve the overall experience of the users. When designing psychological tasks, this category is often overlooked and considered unimportant. However, we argue for the opposite. In the case of games, the user feedback can guide the development to make the game less frustrating, more intuitive, and ultimately more enjoyable. We discuss the user experience in more detail in Section 2.1 and the changes based on it in Section 2.2.

The researchers administering the game are also users of the game. If we were to be completely precise, they would be included in this category. However, at this stage of development, researchers mainly determine the requirements such as what data to log, and their experience in administering the game is not within the scope of our thesis. Therefore, we categorize the researchers' requirements as part of the research category and only focus on the participants when enhancing the user experience. Nevertheless, it is crucial to focus on the researcher's experience in the future.

The second category focuses on research, which is the most important at this game development stage. Despite its importance, we discuss it after the experience because the experience category serves as a helpful guide for deciding about specific implementation of the research needs. In Section 2.3, we discuss how the hypothesis determines game mechanics, how to create appropriate variables to adjust the course of the task, and how to record needed data about each trial.

The final category in Section 2.4 focuses on general code polishing. In software development, it is very important to consider the readability of the code, naming conventions, adding comments, and following good programming practices that ensure sustainable code. This category also covers general settings related to the game application. We need to focus on this category, especially as the game gains in its complexity. However, due to time restrictions and continually evolving requirements from the previous categories, our priority was to create a functional game rather than achieve perfection.

2.1 User experience

The user experience originates in the human-computer interaction and is concerned not only with the interaction itself but with entire user journey. The user experience is emotionally oriented and interested in designing products that provide a pleasant experience. Therefore, it is more often considered relevant for business, which is why it comes with many principles, standards, and good practices that can be easily implemented to improve product design rapidly. (Interaction Design Foundation, 2016a)

User experience design places users at the center of developing products, systems, or services. It involves creating user profiles and improving their experience when interacting with the product. This design draws upon principles from psychology to inform decisions. For instance, understanding Gestalt principles helps designers to understand how people visually perceive elements on websites. The principle of common region uses boundaries to distinguish groups of elements, leading users to unconsciously assume the elements in one group have the same functionality, which is different from the functionality of the elements in another group. (Harley, 2020; Norman, 2013)

User experience is concerned with subjective emotions. As emotions and cognition are closely related, achieving positive emotional experience has been shown to positively influence cognition as well (Lumsden et al., 2016; Ninaus et al., 2015; Norman, 2013; Procci et al., 2012; Vermeir et al., 2020). This is very important for cognitive training, as it may improve its results (Ninaus et al., 2015).

The domains of user experience and its design are business-oriented, which is why the term "user experience" lacks precise definition and proper academic research. It is an umbrella term often mistaken for user interface and usability. Even though they are related and important to consider while designing products, they have their differences.

User interface (UI) is focused on the interface that enables users to interact with the product and can be graphical, voice-controlled, or gesture-based (Interaction Design Foundation, 2016c). In the Restaurant Game, the interface is entirely graphical and includes elements such as buttons, colors, and different screens to navigate the game.

The UI design provides visual clues to the users on how to interact with the elements, where to find the information they need, and what the current product's state is. The most known principles of interactive design were introduced by Norman (2013), and there are six of them. Visibility refers to making items more visible and thus easier to find so that they are more likely to be used by a user. It is important to think about what draws the user's attention. Feedback is important to inform the user about what happened and what the system's state is. Constraints simplify the interface by restricting possible interactions. Mapping states the relationships between controls and their effect. Consistency leverages visual similarity to infer functional similarity. Last, affordance hints at how the object should be used. (Norman, 2013)

Usability is about how easy something is to use and how easily users can achieve their goals (Soegaard, 2019). Nielsen (2012b) defined five components of usability. First is learnability, which focuses on how easy it is for new users to figure out how to accomplish their goals. In other words, how intuitive the design is for its first-time users. The second is efficiency, which is concerned with how quickly users who have already interacted with our product can achieve their goals. Third is memorability, which is about how easy it is for users who have not interacted with the product for some time to re-learn the necessary interactions to achieve their goals. The fourth is focused on errors, which involves questions about the number of errors users usually make, when exactly they make them, and how they deal with them. Last, satisfaction is concerned with how users feel when interacting with the product. (Nielsen, 2012b)

Usability is a measurable metric often studied through usability testing. In this process, individuals representing the target population are asked to complete basic tasks while the designers observe their interactions with the product. This method is currently considered to be the most valuable one for usability. (Nielsen, 2012a)

In usability testing, we can collect both quantitative and qualitative data. To gather the quantitative data, we can measure the time it takes users to find a specific button, count the number of errors, or assess how many users were able to complete a given task. For qualitative data, we can ask the participants to verbalize their thoughts while interacting with the product and follow up with open-ended questions.

Gathering qualitative data may seem subjective and requiring many users, but it is quite the opposite. According to calculations by Nielsen and Landauer (1993), just five users from a specific target population will reveal most of the important issues. While additional users would indeed provide more insights and point out new problems, Nielsen (2000) argues for repeating the testing after making changes instead of putting all resources into a single large testing round.

Usability testing helps the designers to see how the users interact with the product, what they expect, and what they intuitively understand. The testing is about observing and asking open-ended questions to gain insights into user interactions rather than

asking the users for specific preferences. Observation is essential as the users usually do not know what they need. It is up to the designer to identify core issues and decide what changes are necessary to increase the users' satisfaction. (Moran, 2019)

Additionally, usability testing can pinpoint the problematic aspects that can be addressed to eliminate users' confusion and minimize the mental processes unrelated to the task. These processes may occur as a result of non-intuitive or misleading design. To state a specific example related to the Restaurant Game, if the player can not intuitively and quickly distinguish which icon represents the ordered item, they will need additional mental processes to decide which icon to choose.

User experience (UX) is more challenging to understand than the previous two terms. It focuses on the overall experience of the user when interacting with the company, service, or product in any way. Designing a good user experience requires a deep understanding of the user, their needs, expectations, and even previous knowledge. (Interaction Design Foundation, 2016b)

Don Norman is considered to be the inventor of the term UX trying to shift the focus from technology to user. In an interview done by Merholz (2007) Norman explained: "I invented the term because I thought human interface and usability were too narrow. I wanted to cover all aspects of the person's experience with the system, including industrial design graphics, the interface, the physical interaction, and the manual. Since then, the term has spread widely, so much so that it is starting to lose its meaning."

As UX designers have to think about improving interactions with the interface and the usability of the design, some may say that UI and usability are sub-sets of UX. They would not be entirely wrong, as the goal of all three is to put the user in the center of the design. Despite that, UX, UI, and usability have different focus areas. UI is more focused on the interface and the individual elements, usability on ease of completing tasks, and UX on the overall emotional experience. (Norman, 2013)

All three domains will help us to make our Restaurant Game more intuitive, motivating, and engaging. To better address the needs of our users, we prepared a usability and user experience testing to collect their feedback, which will guide further changes in the game development before the validation study.

2.1.1 User experience in Restaurant Game

To understand the users' perspective, it is crucial to ask them directly. We wanted to learn more about the usability and overall UX of the original version of the Restaurant Game described in Section 1.3 and to improve it. That is why we did a round of usability and user experience testing, which we refer to simply as UX testing.

We designed the entire testing process and formulated specific questions based on the article by Moran (2019), the theory mentioned at the beginning of this chapter, and

knowledge gained by attending the UX course for beginners taught by the Slovak User Experience Association[1]. As one part of the course, we had an hour-long consultation with a UX professional that we used to learn about the specifics of game UX and game UX testing. In our testing, we targeted three out of five usability components defined by Nielsen (2012b). Those are learnability, errors, and satisfaction. We excluded the other two components, efficiency and memorability, as they are focused on repeat players.

It is important to stress that we gathered feedback from the population similar to the one chosen for our validation study and if our population was different, the feedback and the implied changes would differ as well. That is particularly true for Restaurant Game, which could be used for research and clinical purposes. Different populations and purposes require specific modifications depending on age, language, cognitive impairments, and other characteristics. Despite the tedious work, this approach improves the quality of the game and makes it more appropriate for given circumstances.

Before our study, another student chose the game to collect data for their student project. They modified the game to suit their hypothesis testing but kept its main characteristics. The student collected data from 72 Slovene-speaking students attending the University of Ljubljana and asked them to comment freely on the game after the experiment. Even though they used a modified game version, we requested the data analysis and feedback provided by the participants. We used the data as an additional source to the UX testing. We refer to it as a preceding collection.

2.1.1.1 Participants

All participants were 25 or 26 years old, with no cognitive impairments and Slovak as their mother tongue. Although we planned to do the validation study on Slovene-speaking participants, here we omitted the language requirement to afford greater attention to other game aspects. Additionally, we already had feedback from Slovene-speaking participants from the preceding collection.

Participants were recruited through the author's personal network. We aimed to include participants with diverse genders, work backgrounds, and gaming experience. Even though we initially recruited 6 participants, one of them canceled at the last minute. Therefore, the final sample consisted of 5 participants (3 males and 2 females), which is a sufficient number for UX testing according to Nielsen and Landauer (1993).

Despite the participants speaking Slovak, the game and data collection were in English. Even if we had used the participants' mother tongue, the inputs would have been irrelevant as our validation study was intended to be done in the Slovene language. Furthermore, the consultant of this thesis did not speak Slovak, which would make the results inaccessible to them.

[1] https://www.suxa.sk/ux-kurz

As English was used throughout the study, we selected participants considering their proficiency in English with a minimum requirement of B1-B2 level to express their thoughts with no significant difficulties. All participants met this criterion, with three reporting their level as B1-B2 and two as C1-C2.

We aimed to choose participants from diverse backgrounds while including someone from the game development field to give us professional insights. The reported fields of work consist of Creative and cultural industry, Analytics and statistics, Artificial intelligence, Event organization and staff hiring, and Mobile game development.

Regarding their gaming experience, three participants reported they used to be occasional or regular players in their childhood or teenage years. One participant stated to occasionally play logical games, while the last one was a professional game designer of mobile games as a work occupation and an occasional player in free time.

Testing the game with people who don't play games much is beneficial, as they have fewer assumptions about how the game should work. Even though it may sound counterintuitive, it helps to identify what really is easy to figure out and what is a common practice that users can not be assumed to know. For instance, regular players will know to use W-A-S-D keys to move their game avatar, while non-players might not be familiar with this commonality.

Lastly, the participants played the game on various devices using Windows and MacOS. This enabled us to observe platform specifics and test the process of making the game available on each platform. While making the game available for both operating systems, we encountered several technical issues described in Section 2.4.

2.1.1.2 Procedure

We did the UX testing online through an hour-long interview. Before the interview, the participants were asked to download a game application compatible with their operating system, and they were told not to open the game. The game application is available for download for both operating systems on the following link, together with a complete list of introductory instructions, predetermined interview questions, and notes from the interviews: https://davinci.fmph.uniba.sk/~dlugosova24/DiplomaThesis_RestaurantGame/UXTesting/.

At the beginning of the interview, each participant was asked to read, sign, and agree to an informed consent. The participants did not receive any compensation for taking part in the study. After signing the consent form, they were asked to complete a questionnaire about their demographics, background, and gaming experience.

After completing the questionnaire, all participants were informed the interview would be recorded and what was the purpose of the recording. All participants consented to it by acknowledging it after the recording started. We recorded the interviews

to ensure no important information was overlooked, as typically, more people monitor the testing. However, we only had one researcher, and the recording made it possible to review the interviews and capture all relevant information, including verbal and non-verbal cues, such as facial expressions and body language.

The participants were informed about the interview schedule and were given a chance to ask questions. After that, they were shown how to start the game and asked to play it as they would normally do while telling us their thoughts.

The game started immediately after opening the game application, and the first order was placed within milliseconds. Just to shortly summarize the game, the player has two minutes to score as many points as possible by correctly assembling the entire order. Each order consists of randomly generated items, with the number of items being randomly generated from the range of one to five.

After playing the game for the first time, the participants were asked questions about their initial thoughts and assumptions. It is important to point out that we tailored the questions to each participant according to what they said while playing the game, and we did not strictly follow the pre-prepared list.

If participants did not understand the game's objective, missed some elements, or were uncertain, we provided a further explanation to ensure that the next time they played the game, they knew what to do. After a short ten-minute dialogue, the participants were asked to play the game again. This time, they did not have to speak out loud their thoughts so that they could completely focus on the game.

The second round was followed by a more in-depth interview. The questions we asked slightly varied for each participant from those we prepared beforehand, as we adjusted them to each participant's answers. The adjustments could help us better understand the participant's experience.

One part of the questions was focused on the experience and asked about the motivation to play again and highlights and pitfalls of the game. Additionally, we asked about recall strategies, which could provide insight into what is easier to remember and how the participants approach similar tasks. The second part of the questions was concerned with the game and asked about the way the participant controlled the game, its visual aspect, and how the information was presented.

Most of the questions were qualitative, as this type of question provides more authentic insights into the user experience. However, we also included three quantitative questions to get a more explicit measure of motivation. The first question asked about the participant's motivation to score as many points as possible on a scale of 1 to 5 where 1 is not at all motivated and 5 is they were in a complete flow. The second question asked about the likelihood of playing the game each day for the next two weeks, knowing the game could improve their cognition. The third question asked about the same likelihood but under the condition of playing the game each day for a month.

The second and third questions used the same scale with 1 referring to they would not play the game at all and 5 they would play it every day.

The more in-depth interview lasted approximately half an hour. At the end, the participants were asked about additional comments and observations in case the questions did not cover them.

2.1.1.3 Results

We start by analyzing the quantitative results of our testing. The motivation to score as many points as possible had answers $[3, 4, 4.5, 4.5, 5]$ with 4.5 representing answers "4 or 5". It is worth noting that the participant who reported their motivation as 3 also failed to score any points in either round. In the first round, the participant did not understand the objective of the game. In the second round, they had difficulty memorizing orders as each one consisted of 4 or 5 items. We attribute their lower motivation to their final game score of 0. Nevertheless, considering the circumstances, the motivation value 3 is still relatively high.

The responses for the likelihood of playing the game repeatedly were $[1, 1, 2, 4.5, 5]$ with the same principle of interpreting answer "4 or 5" as 4.5. Additionally, one of the participants said that if there was evidence that the game improves cognitive abilities, their response would change from 1 to 3. When asked about the likelihood of playing the game for a month, the responses were $[1, 1, 1, 2.5, 5]$.

The participants were asked to provide reasoning behind their report, and the answers may explain the low values. First, the participants suggested making the game available to phones, increasing the convenience of playing it anywhere. Next, adding greater variability to the game would create more opportunities for improving their performance posing as a motivation factor to keep playing. Lastly, one participant reported a low score, believing they did not need to improve their cognitive abilities.

Despite the low likelihood reports, the high motivation scores indicate there is a good potential for further development in case the game is intended to be used daily. Another supporting evidence is that all participants reported they felt they could get better at playing the game and score more points.

We continue to analyze qualitative data that constitute most of the collected data. This section provides an overview of our findings that were used to guide further game development. The complete notes on participants' answers and the researcher's observations are available on the link: https://davinci.fmph.uniba.sk/~dlugosova24/DiplomaThesis_RestaurantGame/UXTesting/UX_testing_interviewNotes.pdf.

Four out of five participants understood the objective to some degree. Despite that, all of them explicitly mentioned that they lacked instructions. Three participants also struggled in the beginning as they did not focus on the initial order.

One confusing element was the time limit. While all participants noticed the two-minute countdown, three of them misunderstood its purpose. They believed the limit applied to assembling a single order rather than as many orders as possible.

On the other hand, all participants intuitively used the drag&drop to interact with the game and place icons of items on the tray. When commenting on the visual part of the game, the participants described it as simple and easy to understand. All main elements were understood and noticeable, showing the game interface is user-friendly, very clear, easily understandable, and easy to use. The only exception was the speech rectangle containing the initial order which was not noticed by three participants. However, as stated before, it was mainly because the participants did not expect it.

When asked about understanding the scoring system, none of the participants got it completely correct. Furthermore, four participants skipped or reported they would skip the second attempt to assemble the order if their first attempt was not successful. This indicates that there is no difference if the participants are given more attempts to complete the same order or not.

On the contrary, the scoring system itself was positively evaluated as a great feedback mechanism. What is more, as a consequence, the participants felt engaged, motivated, and competitive to score as many points as possible and praised getting more points for remembering larger orders.

Our questions about the food and drink items were mostly focused on their discriminability. As the order is placed as a list of words, the participants have to look for the correct items in an icons panel. Participants reported difficulties in differentiating between coffee and tea, as well as between apple and orange juice. On top of that, a very interesting comment was made about the shape of the items. This participant pointed out that all drinks were more similar in shape than foods, making it harder to quickly distinguish which icon represents which item.

Even though we did not ask about the difficulty of remembering different numbers of items, four participants found it challenging to remember five items. They explicitly commented on this difficulty or expressed it through their facial expressions. This was the case especially if the participant was given an order consisting of five items repeatedly or if it was the first order placed in that round.

Participants also provided several other suggestions for game changes. They included replacing the drag&drop mechanism for clicking as it is faster, placing the scoreboard into the eye field of the player and not in the corner of the screen, considering color influence, improving the visual side of the game, and using colorful icons.

The feedback received from Slovene-speaking participants in the preceding collection highlighted the importance of considering language and cultural differences in the development process. These participants provided valuable insight into the selection of icons. Specifically, we chose a donut as one of the food items. The Slovenian word

for a donut is "krof". However, this word also refers to traditional Slovenian food with no hole in the middle and jam as a filling (in Slovakia, they are called "šišky"). In Slovenia, the word "krof" is used more often to refer to the Slovenian traditional food, not the American-style donut. Therefore, the participants did not immediately realize they were looking for the icon of the classical American donut when they needed to find the "krof". We believe this required additional mental resources unrelated to the task and, based on this feedback, can be prevented in the future.

This example emphasizes the importance of considering the specifics of each population before proceeding to a validation study. Having said that, it is clear that if the game was supposed to be used for a different population (different language, age group, etc.), the user experience and usability testing would need to be done again.

2.2 Game adjustments - user

During our UX testing, we noticed two unwanted behaviors. The first occurred while the participants dragged items onto the tray. Sometimes, their mouse cursor moved further away from the icon. Although this may not seem like an issue, we discovered a bigger problem when reviewing the interview where the additional movement occurred.

When the icon of an item was above the tray at the moment of the drop, but the mouse cursor was not, the tray did not register that the item was put onto it. As a result, the item was not part of the assembled order, which resulted in an incorrect evaluation of the participant's answer.

The second unwanted behavior was that the speech rectangle with an order was shown even after the game had ended. Both issues were due to incorrect settings in the Unity part of the game, and we successfully fixed them.

The changes implied by the UX testing results that we considered easy to implement and relevant to a better user experience were immediately made. We moved the scoreboard below the button to be closer to the game's main focus field. Next, as scoring points was reported to be motivational, we added the amount of scored points as part of the feedback message the player receives after assembling a correct order.

At the beginning of the game, we added several screens with game instructions. The instructions are a common part of standardized psychological measurements and were requested by both the users in our testing and the participants in the preceding collection. As part of the instructions, we added screenshots showing the game's environment and progress. After that, we included a list of items with their names and respective icons. We believe the added instructions will prevent the initial confusion we observed during UX testing, ensure the initial order is not overlooked, help the players familiarize themselves with all game elements, and correctly identify each item.

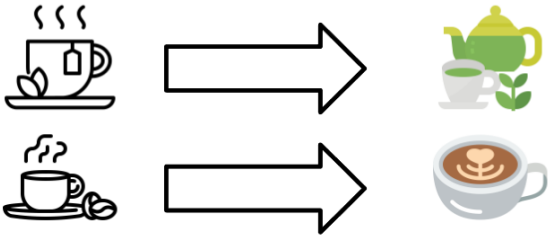

Figure 2.1: The left icons were used in the original version of Restaurant Game, and users reported them to be hardly distinguishable, which is why they were exchanged for the icons on the right with more unique features.

To improve the graphics, we followed the UI principles by Norman (2013) and changed the elements to attract the player's attention to the most important ones. First, we changed the background colors from the previously used red, which is commonly used to signal danger and naturally draw attention. We opted for a plain, faded background instead. Second, we changed the icons from black-and-white to a colored version. The items are a very important part of the game, which is why they should stand out. Additionally, we suspected the non-colored version of items was part of the reason why the participants had difficulty differentiating between specific items.

We replaced all of the icons with different colored versions from Flaticon[2]. We obtained a paid license for each graphical element used in the game using a Flaticon visual. Some of the icons within the game were additionally redesigned to have more distinguishable features. For instance, if we look at Figure 2.1, we see that the tea icon was originally a cup with a tea bag hanging from the edge with leaves on the side. It was difficult to distinguish it from the coffee icon with coffee beans. This was also true when we used the colored versions, as the main part of both icons was a cup. Therefore, we completely changed the icons and chose ones that differed as much as possible at the expense of complicating the visual of the tea icon.

The feedback from the preceding collection included the problematic nature of the donut, even in its colorful version. That is why we replaced it with an egg. The egg was different in color and shape from the other food items, making it easy to recognize. For similar reasons, we exchanged orange juice for lemonade, and to create a higher diversity in shape among drinks, we exchanged a beer bottle for a beer glass.

We also changed the arrangement of the icons to make it less likely for the players to make accidental mistakes. Icons with similar shapes or those that can be easily exchanged (such as tea and coffee) were put close to each other. One of the participants mentioned this adjusted arrangement as potentially beneficial during an interview. Despite the rearrangement, we kept the items separated by the drink and food categories.

[2]https://www.flaticon.com/

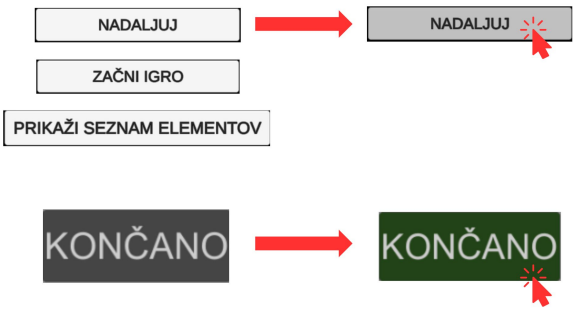

Figure 2.2: The buttons in the upper part are used within the instructions, and all look the same while looking differently from the button on the bottom that is used to evaluate the assembled order. All buttons change color when clicked to provide feedback to the user that they were clicked.

While implementing all of the mentioned changes, we kept in mind the theoretical background of UX, UI, and usability. We demonstrate this in the following example (see Image 2.2). We designed the instruction screens while considering the consistency principle by Norman (2013). That is why all buttons within instructions screens that move between screens look the same but differ from the button DONE *(KONČANO)*, which evaluates the order the user assembled. Another of Norman's principles is feedback, which we implemented as the color change that happens when the button is clicked, giving the players information that some action took place.

When analyzing the interview results, we came up with several large-scale ideas to implement for the validation study. First, the participants expressed the difficulty of assembling orders consisting of five items. That is why we decided to explore the option of changing the game mechanics so the higher number of items would not take the participants by surprise. Second, we decided to omit the second attempt to assemble the same order. The main reason is that the participants reported to ignore the second attempt, making it pointless. However, we kept for discussion the option of including more than one attempt in a different way. As both changes required extensive modifications and interfered with the game mechanics, we needed to include the research perspective. Therefore, we address these changes in Section 2.3.

We put aside some participants' suggestions, such as creating a phone version and adding more levels, as they were considered not beneficial or significant at the current development stage.

Up to this point, we discussed what needs to be changed, but it is important to mention the positive things pointed out by the participants so we know what to keep. Participants found the drag&drop to be an intuitive way of interacting with the game and the texts to be easy to read. They also praised the immediate feedback in the form of announcements in the speech rectangle and were curious about their current score.

We want to highlight two comments from interviews. In the first one, the participant described the restaurant setting: "The story behind it is nice - creates the feeling I am doing something meaningful and not just memorizing the items, which can make me more engaged." The second was from the researcher from the preceding collection: "I liked it was placed in the restaurant. I like it is more ecologically valid even if you are not a waitress, remembering food and drinks is more natural than remembering squares." These comments emphasize benefits of real-world settings and game elements.

In summary, talking to participants gave us many insights about how people perceive the Restaurant Game. We could identify the good aspects of the game that should be kept. The UX testing also helped to identify and implement necessary adjustments to make the game more intuitive. On top of that, we came across several discussion points that require input from the research perspective to which we now proceed.

2.3 Game adjustments - research

The next category of changes focuses on research requirements. These are the most important ones, as they imply adjustments without which we could not test our hypothesis. They helped us develop a version of Restaurant Game that reflects working memory and addresses shortcomings of the current measuring and training methods.

First, the game must fulfill the research needs on functioning and the task content. In our case, the participants are shown a list of items and assemble the order according to the given instructions and researchers can adjust game parameters to fit their study.

Second, we need to be able to record the data necessary to reject or support the research hypothesis. These data include but are not limited to the research codes, the number of remembered items, final score, error rate, response time, and time stamps.

Below, we describe the adjustments that we implemented based on the research findings presented in Chapter 1, the results of the preceding collection, UX testing, and several consultations with the consultant of this thesis.

The biggest research issue was that the original version of Restaurant Game most likely reflected short-term memory instead of working memory. To create a game that would be aligned with research on working memory, we primarily needed to ensure intentional manipulation of the remembered information.

When deciding on new game mechanics, we included the UX testing results and knowledge on the functioning of current measurements from Section 1.2.2. The first topic left to discuss was adjusting the overall difficulty, and the second was considering whether the player should have a single try or multiple tries to proceed to the next difficulty. Describing each change individually does not make sense, as we had to completely change how the game functions. Instead, we introduce whole new mechanics.

The new version has one customer who places an order. This order is randomly generated and displayed in a speech rectangle. The order is shown for $1.5 * n$ seconds, where n is the number of items within the order. We extended the showing time from one second to ensure the players have enough time to comprehend each word. The initial order size is set to two and gradually increases. We do not start with order size one, as there is no possibility for the necessary manipulation to reflect the working memory. However, the number two is not hardwired into the script but is implemented as a variable corresponding to a minimum order size, which the researcher can modify from Unity. This is the first significant change in the game mechanics.

The player has two attempts at each order size, but the orders differ. In other words, the first order of a specific order size is shown, and regardless of the correctness of the player's answer, it is followed by a second order of the same size. If at least one of the attempts is successful and the order is correctly assembled, the number of items in the order increases by one. This is a transformed method of limits with two stimuli presented at each level of difficulty. In other words, the player must succeed on at least 50% of trials, which is aligned with the definition of measuring memory span by Miller (1956). The game ends when the player does not succeed at remembering any of the orders for a current number of items or reaches the maximum order size. This is the second significant change to the game mechanics.

With the change in providing the second attempt, we also modified the scoring system. The UX testing showed that the original system was not easily understandable, yet the points posed a motivating factor. Therefore, we wanted to include them but in a more straightforward way, which is why the correct order earns the participant n points, with n being the order size. An incorrect answer does not earn any points.

The third significant change is that we added a rule determining the sequence in which the player must place the items on the tray. This rule ensures manipulation of the information stored in memory. The rule is to repeat all ordered drinks first and food second, which is natural for the restaurant environment. However, the fact is that only adding this rule in the evaluation process of the order does not ensure the manipulation as long as we keep the order generation completely random. For instance, if the order is generated without any drinks or in the correct order, the manipulation is not necessary, and that could skew the results to an unknown extent.

For the reason mentioned above, we changed the order generation to not be entirely random. After several unsuccessful attempts, our implementation of the generator ensures the needed requirements: the order includes one drink and one food, and there exists a drink item presented after some food item. The implementation is described in in Appendix B.

We completely excluded the time element even though it was a motivating factor for two participants of UX testing who described themselves as competitive and tried

to beat their previous scores. In this new game version, the possibility of scoring more points comes with remembering a higher number of items. This shift in the scoring system is aligned with the evidence that children are more likely to focus on speed, but with age, we put more importance on accuracy (Jurado & Rosselli, 2007). This can be considered a fourth change, even though it naturally resulted from the previous ones.

Another thing we changed was the game difficulty. We started discussing the overall number of items, which was 14 in the original version. According to Miller (1956), the individual's memory capacity is 7 ± 2, in which case 14 is enough. However, we did not know whether the game reflected the working memory and matched current research findings. Therefore, we also took into account the results from the preceding collection, as the sample was from the population intended for our validation study.

The analysis of the gathered data showed that the average memory span was 6, which was also a median. The highest span was 12, and only one participant had it. The span of eleven had two participants, and the span of ten had again only two participants. As some participants had high memory spans, we decided to increase the number of items from fourteen to sixteen.

We did not have any other data to guide our decision, and we acknowledge that fourteen items may have been enough. However, we do not know if and how the difficulty changes when the participant reaches twelve items and only two items are left out compared to four items being left out. As we can not settle the debate with the available evidence, we leave it for future research.

Lastly, we changed the language of the game. We translated all texts and elements into Slovene, mother tongue of population chosen for our validation study. Otherwise, processes related to comprehending foreign language could influence the game results.

We continue to describe the changes necessary to enable effective data collection. Our study needed research codes, final scores, and the final order size. Even though these may be collected without any sophisticated solution, it would not provide flexibility and sustainability for future use of the game within research. That is why we implemented a logging system as a script named *GameLogSytem*.

The current information recorded within the log file includes the research code and version of the game (in our case, drinks first and food second). Next, the information about the game progress, such as how many items are in order, if it is the first or second attempt, and when something is clicked and put on the tray. During evaluation, we record the given and assembled order with the addition of an error record in case the orders do not match. We also record information on the final score and the order size.

To collect the research codes from the participants, we added a new instruction screen at the very beginning. We test the code to be null or white space, in which case the instructions do not continue and force the player to input the correct code. After that, the *GameLogSytem* creates a text file with the name corresponding to the code.

The log system has an internal clock that starts after the player goes through all instructions. Thanks to that, we can put a time stamp on each event that occurs, and researchers can easily calculate response times for whichever actions.

Currently, *GameLogSystem* is the only script with direct access to the log file. We implemented it in a way so that information in the file is not accidentally overwritten. Additionally, the log continues if the game is played repeatedly, and no data is lost.

To meet research requirements, we completely changed the game mechanics based on current research findings and implemented an extensive logging system to collect necessary and additional data from each trial. At this stage, the game looked and functioned as we wanted. However, we proceed to Section 2.4 to discuss general code changes needed to increase the quality and sustainability of our software.

2.4 Game adjustments - code

The last category of changes is focused on general code polishing and configuring game options. These changes reflect good programming practices, focus on fixing bugs, adding appropriate code comments, renaming the variables and functions to be self-explanatory, changing the code to make it easier to read, adding reset and exit buttons, and much more. Some of these changes were necessary to make in order to implement the changes from the previous two categories, and others improved the sustainability of the Restaurant Game. Even though we did all of the above, we do not mention every single change but choose a level of detail appropriate for this thesis.

We start with code bugs, which we discovered by more extensive testing. The first one was in displaying the speech rectangle. When the rectangle contained nine or more items, it extended the game window, and some of the ordered items were not visible. We changed several constants within the script, which were responsible for adjusting the size of the speech rectangle to correctly position the rectangle within the screen.

The second bug was less likely to affect anyone as it only occurred when a player remembered all sixteen items correctly. In this case, the game did not end even though it was supposed to. To fix this, we added a new condition for the end game based on the maximum number of items within the order. This provides researchers with another parameter to adjust to their needs within Unity. We decided to set our maximum number to fifteen as the last item is always deterministic and does not require more resources than remembering just fifteen items.

As the original scripts were intertwined and written with beginner's knowledge, there were many possibilities to increase the simplicity of the game implementation. We started with adding a script called *OrderChecker* that encapsulates the game mechanics. It checks whether the assembled order is correct and determines further progress

in the game. The benefit of encapsulation is that one functionality is put into a single class, which makes the code easier to read and build upon. Originally, the mechanics were split among multiple scripts.

Next, we added a property for each item, determining whether it was food or drink. This enabled us to generate orders ensuring manipulation as described in Section 2.3 and to check if the player assembled the order according to the game rules. The benefit of such implementation is that the property can be set from Unity when a new item is added without needing to access scripts.

Moreover, we significantly changed how other scripts refer to items. Originally, we worked with the items as with a *GameObject*, the most general and abstract class, and not by their attached script *Victuals* that is specific for the items. The biggest issue of that was that we could not easily access any parameters, features, or functions specific to the items. We changed the implementation, which simplified the usage of the item's script and enhanced the code's readability. We had to adjust all other scripts to match the new implementation subsequently.

Another benefit of this change is that it enables us to create different language versions more easily. It is sufficient to exchange the picture file with a visual representation of each item. The item name shown within an order will match the file name.

Improving code is a long process. For now, we wanted to make the code easier to read and understand and not include unnecessary interlinks. In that case, it is easier for researchers to see the relevant parameters that can be changed for their research, for new developers to quickly orient themselves in the code, and for game developers to make changes and further develop the game.

In the last part of this section, we provide an overview of the issues we dealt with when making the game available for different operating systems. We believe this can help future developers and researchers to know how to make it work but also what we tried, what worked, what did not, and where there may be potential to try again.

After a game is developed in Unity, it has to be built for a specific platform, such as Windows, MacOS, Android, or WebGL, to be put online. We were not successful in making the game online, so we had to make it available for different operating systems.

We developed, tested, built, and finally played the game on a Windows computer without issues. However, as we knew we could not rely on all our participants using Windows, we also needed to create a game available on MacOS. In this case, we encountered several issues. We looked at the discussion topics on an official Unity forum[3] to look for solutions or alternatives.

First, Unity allows one to switch building platforms and build the game for MacOS, even from a Windows computer. However, as we later learned, it is not a real option.

[3]https://forum.unity.com/

When trying to open the game application, an error occurs, announcing there is no application to open. Therefore, if the game is built on a Windows computer, it does not matter that it was built for MacOS, it will not be executable. From what we gathered, this is a long-lasting issue that has not been dealt with by Unity.

After many unsuccessful attempts to make the game application work, we downloaded Unity and the whole code to a MacOS computer and built the game there. We still encountered several error messages and game-loading issues. However, if the whole code is compressed on Windows, ideally with MacOS pre-settings, and then transferred to MacOS, further issues are minimal.

After we built the game, we zipped the file and uploaded it to Google Drive to make it easily accessible to all participants. However, the game was not playable after downloading and unzipping the folder. We expected there might be some issues with specific compression methods, which is why we tried different compressing methods and cloud services for uploading the file. The tar format worked, so we used it.

When trying to open the game, the computer did not allow the file to open as it was an application from an unknown source. This issue was the easiest to solve as it happens often and is not specific to Unity games. What needs to be done is to allow the game to open from the Security settings. After all this, we successfully transferred the game to another Mac computer and let somebody else play it.

The most promising solution to avoid any technical issues of such kind is to make the game available online. After building the appropriate version of the game for the website, we uploaded the game to a web server managed by Comenius University in Bratislava. The game is available at an address: https://davinci.fmph.uniba.sk/~dlugosova24/DiplomaThesis_RestaurantGame/web_version/.

We encountered two significant problems with this option. First, after all the instruction screens, no order is shown once the game starts. After pressing the evaluation button, the speech rectangle shows feedback. As the rectangle can appear, the problem is not with it. The second issue is that the game logging system creates a log file in the game folder. However, there is no such file for the online version.

We have limited knowledge of game development and did not find solutions to the presented issues. Still, we perceive the online platform as a very good option for the future and advise focusing more on solving them.

2.5 Final version of Restaurant Game

We end this chapter with a section summarizing the new version of the Restaurant Game. The game starts with an input field for the participant's research code. After entering the research code, a log file is created in the game folder, and the participant

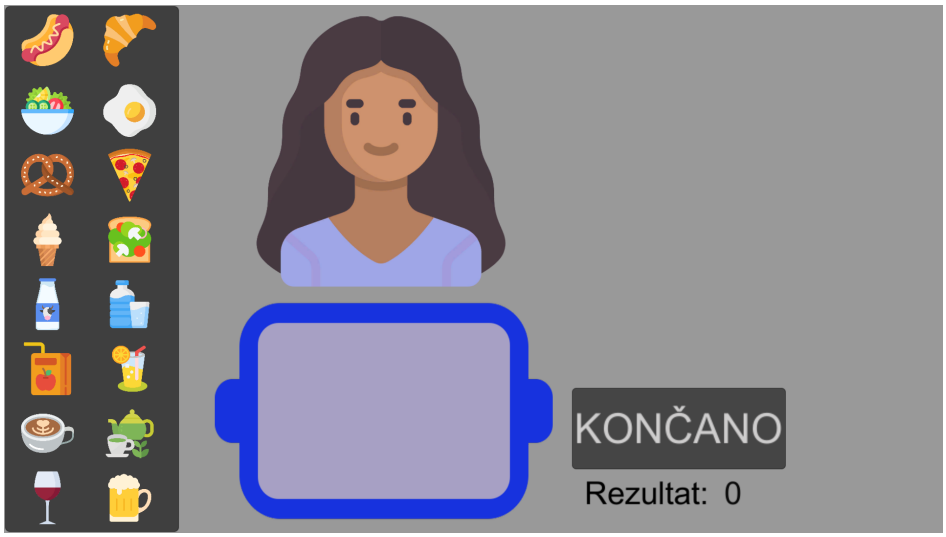

Figure 2.3: The new version of Restaurant Game, with colorful icons presented in a panel of the left, with one customer to serve and a tray on which the order is assembled. The button is pressed to evaluate the order, and the scored points are shown below.

continues through several screens with instructions about the game. These instructions also include exemplary screenshots of how the game proceeds and a complete list of items with their names. The last screen has a button to start the game.

After the start, the game is in its default state (see Figure 2.3). In 1.5 seconds, the speech rectangle shows the first randomly generated order consisting of two items. During this, the left panel containing the items is not visible (see Figure 2.4). After the order disappears, the items appear, and the participant can drag and drop items onto the tray. Each order is shown for $1.5^{*}n$ seconds, where n is the order size. The sequence of putting them on the tray is essential and given by a rule. In our case, the rule is drinks first and food second. When the order is complete, the participant presses the done button (in Slovene *Končano*), after which the assembled order is evaluated.

If the order is correct, the participant receives n points, where n is the number of items in the order. If the order is incorrect, they do not receive any points. The participant has two attempts for each order size, with each order being randomly generated. If at least one of the attempts is correct, the number of items in the order increases by one. When both attempts are incorrect, the $n - 1$ is the game result, which is the hypothesized working memory span equal to the size of the largest order successfully remembered. After the game ends, the participant is presented with their final score, which is the summary of scored points.

The code of the game and the game applications for Windows and MacOS can be retrieved from the following address: https://davinci.fmph.uniba.sk/~dlugosova24/DiplomaThesis_RestaurantGame/MainResearch/GameFiles/

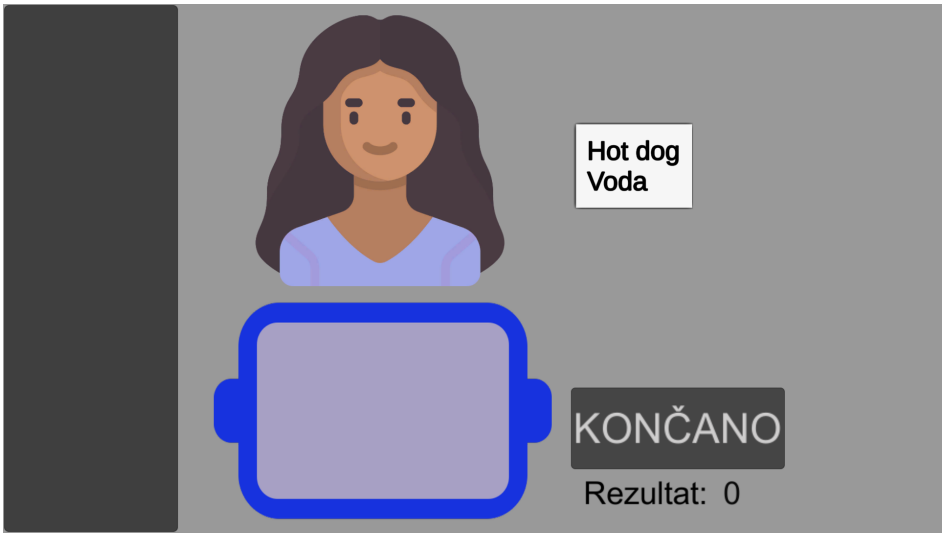

Figure 2.4: The order is shown to the customer's right as a speech rectangle containing the list of ordered items. Meanwhile, the icons are invisible.

Chapter 3

Method

The Restaurant Game is a prime subject of this thesis. It shows a customer who places an order, and the player's goal is to correctly assemble it according to a specific rule. The rule determines the sequence in which the items must be placed on the tray.

The Restaurant Game is a newly developed task and, for that reason, has not been tested as a measure until now. In this thesis, we focus on its convergent and divergent validity. Additionally, we are interested in the user experience in all of the tasks used. The participants will play the game twice, with the second trial being approximately one month after the first one. We use the results to examine the game's test-retest reliability, which reflects to what extent the obtained scores are stable over time. Our research was approved by the Ethical Committee of the University of Ljubljana.

3.1 Validation

In the validation part, the Restaurant Game is tested against selected psychological measures. To test convergent validity, we chose working memory tasks that are more similar to the game, specifically the Digit span task (backward) and the Corsi block-tapping task (backward). To test divergent validity, we chose Shape-filling task that is different. We used computerized tasks to eliminate possible differences among formats and translated the tasks into Slovene to match the language of our population.

3.1.1 Preparing comparison tasks

Before describing the method of our study, we outline how we chose and modified the tasks for the purposes of our research. To save time and leverage existing resources, we used PsyToolkit (Stoet, 2010, 2017). This platform offers a wide variety of psychological tasks to choose from. Each task also includes a script that can be downloaded and adjusted to one's needs. The website also provides comprehensive documentation that helped us learn the syntax and implement necessary script changes.

The first task needed to be similar to the game and measure working memory. The possible options on PsyToolkit were the Digit span, N-back, and Corsi tasks. The second one was not an option as it reflects just the updating function and was shown to not reliably reflect working memory capacity (Jaeggi et al., 2010). The Corsi task is discussed below. However, it was not our first choice as it assesses visual memory, and the Restaurant Game relies on verbal stimuli. Therefore, we opted for the Digit span task. The PsyToolkit had only its forward version, which is used as a measure of short-term memory span (PsyToolkit, 2022a). Therefore, we needed to change the task to its backward version to assess working memory.

In the original script, the digits were saved in an array and shown one by one at the beginning of the trial. At the end of the trial, the answer was compared to the saved digit array. As there is no function to reverse an array, we created a new one called *numberSequenceReverse*, which would be used in the final comparison as the required answer instead. While the digits were shown one by one from the original array, they were also prepended to our *numberSequenceReverse*. That way the first digit would end up as last and the last digit as first, which made the needed backward order.

In the forward version, the participant had to remember two sequences correctly to level up. If one of the sequences was incorrect, an additional trial was done to decide. We changed this to match the game mechanics of the Restaurant Game, which means two attempts for each number of digits while at least one of them has to be correct.

The final version of the Digit span task (backward) starts with instructions followed by training on two digits. After that, the real data collection starts. Every sequence is shown digit by digit, each for 800 milliseconds. After the sequence ends, the participant is presented with a keypad to input the answer. There are two buttons. One enables the participant to delete the digits from their answer (*Izbriši*, Slovene for delete), and the other submits the answer (*Nadaljuj*, Slovene for continue). After submitting the answer, the digit sequence is evaluated, and the participant receives text feedback on whether the answer is correct or not. The participant has two attempts for each sequence size, and if at least one of them is successful, the sequence size increases by one. The digits in the sequence do not repeat. Therefore, the maximum sequence size is nine. To see the task environment, see Figure 3.1. (PsyToolkit, 2022a)

Even though we originally intended to include only one task measuring working memory, we added the Corsi block-tapping task (backward). Corsi is widely used to measure working memory, specifically its visuo-spatial component. The Digit span task (backward), which we already chose, is more focused on the phonological loop. We found it beneficial to include tasks measuring both types of working memory storage as we were not sure on which component the Restaurant Game relies more on.

The Corsi task (backward) was already available on PsyToolkit (PsyToolkit, 2021a), and we only needed to make few changes to the script. As in the Digit task (backward),

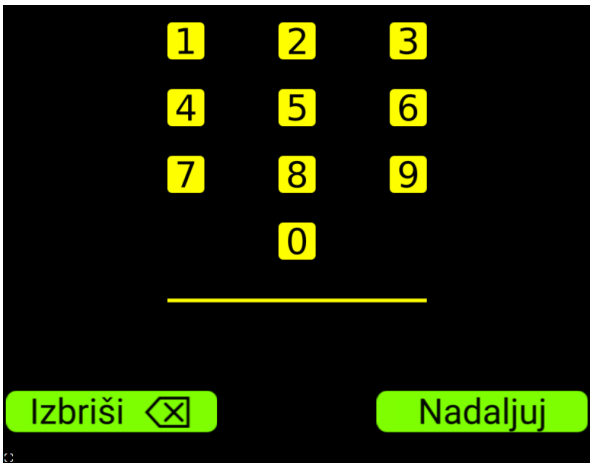

Figure 3.1: The Digit span task (backward) has a digit keypad in the center for inputting the answer. There are two green buttons. The left one, *Izbriši*, is for deleting digits from the answer, and the right one, *Nadaljuj*, is for submitting the answer.

we needed to change the task mechanics of increasing the difficulty to match the game. In the original version of Corsi, the difficulty increased if the first attempt was correct. Only if the first attempt was incorrect, it was followed by a second one. After two wrong answers, the task ended. We changed that to always provide the participant with two tries and check if at least one was correct to proceed to the next difficulty.

The final version starts with instructions, after which the participant is presented with nine purple squares. The position of each one is randomly generated for each trial. After 0.5 second, the squares start to light up to yellow one by one, with each lightening up for 0.3 seconds, followed by another square after 0.3 seconds. When the sequence ends, there is a signal to input the answer. The signal was originally auditory, but we changed it to visual to lower technical requirements. After clicking on the squares in the reversed order, the participant presses the finished button (in Slovene *Končano*). The button changes to a smiling or frowning face, depending on whether the repeated sequence is correct or not. The face is shown for 1.5 seconds, and then the task continues. The number of squares in a sequence increases until all nine squares are included. See the task environment in Figure 3.2.

Next, we needed a task different from the game but still measuring a related concept. We looked for a task on basic executive functions, as using one on complex cognitive functioning may complicate drawing conclusions. Updating tasks are too similar to the game, with updating itself as a crucial part of working memory. Miyake et al. (2000) shows inhibition is highly correlated with the common part of the basic functions. That is why we also excluded inhibition tasks. On PsyToolkit we looked for tasks on shifting, namely Number-letter (PsyToolkit, 2022b), Shape-color (PsyToolkit, 2021c), and Shape-filling (PsyToolkit, 2021b).

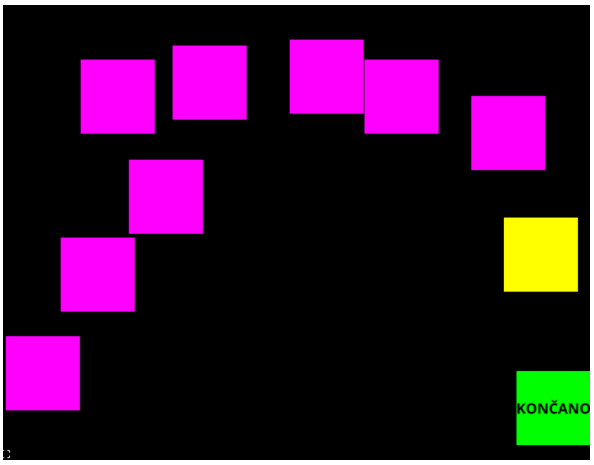

Figure 3.2: The Corsi task (backward) shows nine squares that are purple by default and light up to yellow one by one in a sequence. In the bottom right corner, there is a green button *Končano* that is pressed to submit the answer and continue the task.

The first one is implementing the number-letter task by Rogers and Monsell (1995). The response to stimuli is determined by one of two rules—first, whether the letter shown is a vowel or consonant, and second, whether the number shown is odd or even. This task requires more complex verbal and numerical processing and decision-making than deciding whether a shape is a square or a diamond. We wanted to avoid this additional mental load.

The second option has stimuli characterized by a shape and a color. The rules determining the correct answer depend on these characteristics, making the task dependent on color. If we wanted to use this task, color-blinded people would be excluded from our study. We believe this condition should not be relevant in a working memory study, which is why we did not choose this task.

We chose the third option, Shape-filling task, which is similar to the task used by Stoet et al. (2013). It has the same nature as previous two, but the response rules are simpler and do not depend on factors irrelevant to executive functions, such as color.

In this task, each stimulus is characterized by its shape (diamond or square) and filling (two or three dots). The stimulus is shown within a framework of two rectangles above each other (see Figure 3.3). The upper rectangle is marked with text *Shape* (in Slovene *Oblika*). When the stimulus appears here, the participant responds according to the shape of the stimulus, pressing b if it is a diamond and pressing n when it is a square. The lower rectangle is marked with text *Filling* (in Slovene *Polnilo*). When the stimulus appears here, the participant responds according to the filling of the stimulus, pressing b if there are two dots inside and pressing n when there are three dots inside.

The initial instructions are followed by training in shape-only, filling-only, and mixed blocks, with five trials for each block. The training is followed by a real data collection

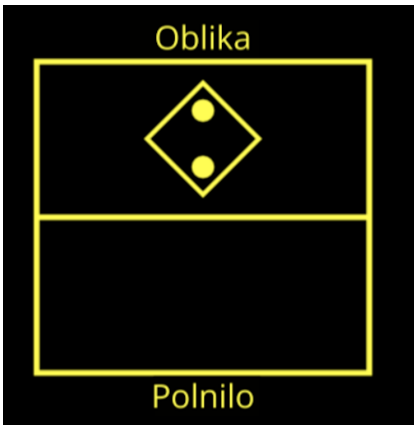

Figure 3.3: The Shape-filling task has two rectangles in the center within which the stimulus appears. The position of the stimulus determines whether the participant answers based on the stimulus's shape (*Oblika*) or filling (*Polnilo*).

with the blocks in the same order. The shape-only and filling-only have 48 trials, with each of the four stimulus combinations presented twelve times. The mixed condition has 96 trials as it combines the previous two blocks. The participant has 4 seconds to respond otherwise it is a timeout. In case of a timeout or incorrect key press, the participant is informed about it and, after a second, returns to the task. The next stimulus is shown 0.8 seconds after the participant's correct response or after they return to the task in case of an incorrect answer or timeout.

As the task was long and presented many stimuli, we discussed decreasing the number of trials. However, to maintain precision, we kept it the same and only decreased the number of training trials from the original ten to the current five per block. Additionally, a reminder of the rules followed each incorrect or timed-out answer. This summary was shown for five seconds, and the original version described it as a way to make participants concentrate (PsyToolkit, 2021b). With the initial training, we found this to be annoying to the participants and an unnecessary prolongation of the task, so we omitted it.

Ultimately, we created a survey in the PsyToolkit with informed consent, a Digit span task (backward), a Shape-filling task, and a Corsi task (backward). These tasks are available at https://www.psytoolkit.org/c/3.4.6/survey?s=3xkKx.

We chose the order of the tasks to separate the Digit and Corsi tasks (backward), which are both measures of working memory. Additionally, the Digit task (backward) uses a digit keypad. Our intuition was that more visually oriented participants might (un)consciously use a strategy to remember the digits as a visual pattern, thus tapping the visual component of working memory instead of the intended verbal one. If the Corsi task would follow immediately after the Digit task, the Digit task could serve as training, improve the results for the Corsi task, and skew our results.

3.1.2 First data collection

After preparing the experiment and starting to collect the data, we encountered an emergency that interrupted the whole experiment. While we know the data do not hold sufficient value, we include a description of this attempt. Based on this experience, we made several changes and, after that, had a successful second attempt.

3.1.2.1 Participants

All participants were psychology students in their second year at the University of Ljubljana. They were asked to participate during their lesson with Prof. Podlesek, a consultant on this thesis. Each one of our participants agreed to voluntarily participate in the study by signing the informed consent.

We sampled a homogeneous group of young people within the age range of $20 - 22$. Two participants did not fill out the demographic questionnaire, and thus, we can not report their age. None of the participants reported to have any cognitive impairments. Our sample consisted of 30 participants, from which one was excluded as their mother tongue was not Slovene. Another 16 would have to be excluded as they did not finish all parts of the study, leaving us with 13 participants. Therefore, the final sample was too small even if we wanted to use the data, which is why we did not analyze them.

3.1.2.2 Procedure

In our first attempt, participants were in one classroom and did the experiment simultaneously. They started by reading and signing the informed consent. After that, the experiment was completely computerized. Each participant received detailed step-by-step instructions (Group A or B) to follow to complete the (see Appendix A).

We started by gathering basic demographic information (age, gender), checked whether they fulfilled the study criteria (no cognitive impairments, Slovene is their mother language), and asked about their previous gaming and waitressing experience. The questions on previous experience are directly related to the nature of Restaurant Game and that is why they were of our interest.

The participants randomly chose instructions for one of two groups that differed in the task order, which was meant to eliminate the order effect. The instructions for Group A started with tasks in PsyToolkit, Digit span task (backward), Shape-filling task, and Corsi block-tapping task (backward). They continued with the Restaurant Game, and after playing it, the participants were asked to upload the game log file to a dedicated Dropbox. Group B had the order reversed, so they started by playing the Restaurant Game and continued with the tasks in PsyToolkit. The game was described in more detail in Section 2.5 and the PsyToolkit tasks in Section 3.1.1.

After finishing all the tasks, the participants were asked to complete the final questionnaire. In this questionnaire, we asked the participants how they would rate their engagement on a standard seven-point Likert scale, where 1 is not engaged at all, and 7 is completely engaged. We asked the same question about motivation. To collect qualitative data, we asked the participants to compare their experience among the tasks, to provide additional comments on the game, and to describe a strategy they used to achieve better results, if they used any. The qualitative data are not directly relevant to our study but can bring insights for further game development and future research.

3.1.2.3 Evaluation

Our data collection was done at the University of Ljubljana during a lesson. Approximately in the middle of the experiment, the university had an emergency drill, and all the students were asked to leave the building immediately. They were allowed to come back approximately an hour later, and after that, they continued with the experiment. As a result, many participants did not even finish the experiment, and due to the interruption, we can not assign any real meaning to the collected data.

Based on our experience, we decided that doing the study during a lecture has too many risks, which we originally did not consider, such as interrupting unexpected events and students being much more distracted by one another. Therefore, we changed our methodology for our second attempt at collecting data.

Despite the unsuccessful data collection, we still wanted to get feedback from the participants and make further adjustments based on their answers. Especially insightful were the strategies they used, which helped us to improve the game even more.

Several students mentioned that before the speech rectangle disappeared, they regrouped the items to the correct order, with drinks first and food second, and remembered it like that. That means they did not have to manipulate the items after order was no longer present, which makes the game more likely to reflect short-term memory.

We made the following change to enforce the internal manipulation with no stimuli present. Each item is shown individually for 1.5 seconds, and the items appear sequentially one after another. This means that the entire order is never shown all at once, which is more consistent with other working memory tasks we use.

During the discussion, we noticed that even though all participants put drinks first and food second, some also kept the order within each category, and some did not. For example, if the order was egg, tea, and toast, the first group of students would consider the answer tea, toast, and egg as incorrect, while the second one as correct. The first approach is more strict, and the second one is more loose. The game implementation was aligned with the approach of the second group, allowing for any order of the items as long as all drinks were put onto the tray before the food.

This confusion results from miscommunication among the thesis authors, which may arise because both approaches to the task should reflect working memory. After further discussion we decided to change the game implementation to only allow the answers that keep the ordering of items also within each of the categories. Even though we did not find sufficient empirical evidence to support or oppose our decision, we suspect the more loose approach may eventually slip to recognition of items presented as part of an order (Stern & Hasselmo, 2009).

Next, we changed the layout of the items in the left panel to have drinks at the top and food at the bottom. This way, we use the design to suggest the correct order of item groups, as drinks are supposed to be put on a tray first. The new layout is aligned with the mapping principle of good interaction design proposed by Norman (2013).

Despite our extensive testing of game settings for Mac computers (see Section 2.4), a few participants using Mac encountered a new issue. After they went through the instructions with no problems, the customer appeared, but no order was shown. Additionally, the submission button did not work.

We are unsure about the real cause of this issue, as other MacOS users did not have the same problem. Interestingly, we encountered the same issue after some time. While solving the issue for ourselves, we learned that MacOS puts seemingly dangerous files under quarantine. The walk-around is a bit more technical and includes overriding the quarantine through the terminal using a specific command, which worked for us. Therefore, we concluded there may have been some (new) system updates, forcing the quarantine on the game application. We created a document for the participants with detailed instructions on how to solve this issue and included it in the instructions.

3.1.3 Second data collection

3.1.3.1 Participants

As a result of the first unsuccessful data collection, we needed to find a different group of participants for our study and repeat the measurement. We chose psychology students in their first year at the University of Ljubljana. Our sample had 70 participants, from which eight were excluded. One was excluded as they claimed to have cognitive impairments, four because their mother tongue was not Slovene, one because they only filled out the first questionnaire, and two because they did not submit their Restaurant Game results. The final sample consisted of 62 participants ($57F, 5M$) who were between 19 and 22 years old ($M = 19.5, SD = 0.646$). The participants were divided into two groups, with 28 participants in Group A and 34 participants in Group B.

3.1.3.2 Procedure

After our experience from the first data collection, we decided to change the procedure. As the author of this thesis could not do the study in person, we opted for an in-home option. Prof. Podlesek, the consultant of this thesis, contacted the chosen group of students via the university's online system and also during the lecture and asked the students to participate in our study in their free time. Each student received step-by-step instructions to follow, same as in the first data collection (see Appendix A).

There were few differences in the procedure compared to the first data collection. Firstly, as participants did the experiment in their free time, they digitally checked a box after reading an informed consent. Secondly, we included a question about the assigned group in the first questionnaire. Lastly, the items in the game were presented individually, and the order had to be assembled while keeping the order of the items within categories. Other than that, there were no changes in the procedure.

3.1.3.3 Statistical analysis

The collected data was combined into jamovi (jamovi project, 2024) from Google Forms, PsyToolkit, and Dropbox. The first included data from the questionnaires. The second provided data from Digit span (backward), Shape-filling, and Corsi block-tapping (backward) tasks. The last had log files from the Restaurant Game. Several graphs were generated using Python (library plotly)(Plotly Technologies Inc., 2015).

The final data for each participant included research code, group, gender, age, if they have cognitive impairments, and if Slovene is their mother tongue. Then, task data included memory span measured by the Digit span task (backward), average response time in pure and mixed conditions in the Shape-filling task, memory span measured by the Corsi task (backward), and for the Restaurant Game, the span and final score (total amount of scored points). Additionally, we calculated the switch cost as a ratio of response times for mixed and pure conditions in the shape-filling task. We chose ratio over difference as it is a more suitable measure for generalization. Finally, the data included user experience reports on engagement and motivation for all tasks.

The results of Corsi task showed that several participants scored zero due to different issues encountered. These participants were excluded from all analyses concerning the Corsi task and experience as their results were not representative.

As the data were not normally distributed, the validity was tested with Spearman's correlation test on three pairs of tasks. Each pair consisted of the Restaurant Game and one of three PsyToolkit tasks. To compare the experience among the four tasks, we chose Friedman's analysis of variance for engagement and motivation reports separately. Additionally, we summarized the qualitative data reported on the user experience.

3.2 Reliability

3.2.1 Participants

We reached out to the participants of our second data collection again after approximately one month. We needed the same participants to play the game at least twice after a time delay to test test-retest reliability. We contacted them via their online university portal with detailed instructions. We collected data from 33 participants $(30F, 3M)$ who were between 19 to 21 years old $(M = 19.5, SD = 0.564)$.

3.2.2 Procedure

To test the test-retest reliability of the Restaurant Game, we let the participants play the game for the second time within an interval of 20 and 39 days after the validity part of the study. In the same way as in the validation part of our study, participants were asked to participate in their free time. They received detailed step-by-step instructions, which were a game-relevant subset of the instructions used within the first part of the study. The participants were asked to download the game, play it, and upload the log file into Dropbox. Additionally, we asked them to put the number two after their research code to prevent mixing the results with the previously collected data.

3.2.3 Statistical analysis

We added the data about the final span and scored points from this part of the study into the jamovi table containing the final data from the validation part. We analyzed the correlations and compared the mean values of game results and final scores between the rounds to test the reliability.

Chapter 4

Results

To simplify the presentation of the results, we refer to the tasks by shorter names: Digit for the Digit span task (backward), Shape-filling for the Shape-filling task, Corsi for the Corsi block-tapping task (backward), and Restaurant for the Restaurant Game. The summary of data is available on the following link: https://davinci.fmph.uniba.sk/~dlugosova24/DiplomaThesis_RestaurantGame/MainResearch/Data/.

Four participants scored zero in Corsi, which did not reflect their visual capacities. We excluded them from any analyses concerning Corsi and analyses of quantitative data on user experience, as the zero scores could bias their experience.

4.1 Psychometric characteristics

4.1.1 Group differences in results

The participants were divided into two groups with different order of the Restaurant and PsyToolkit tasks. Before we analyzed the data properly, we needed to test whether there were any significant differences between the groups that should be considered.

The Shapiro-Wilk normality test revealed that data for both groups in Digit and Restaurant, and Group B in Corsi were not normally distributed (see Table 4.1). On the average, the participants in Group A had higher results in the Digit task ($Mdn = 7, IQR = 1.25$) compared to Group B ($Mdn = 6, IQR = 1$). In the case of Shape-filling, there was a very slight difference, with Group A scoring more ($M = 1.82, SD = 0.268$) than Group B ($M = 1.8, SD = 0.224$). For the Corsi, there were no clear differences between the groups ($Mdn_{\mathrm{A}} = 6$, $IQR_{\mathrm{A}} = 2.00$, $Mdn_{\mathrm{B}} = 6$, $IQR_{\mathrm{B}} = 1.50$). The same was true for the Restaurant ($Mdn_{\mathrm{A}} = 5$, $IQR_{\mathrm{A}} = 1.25$, $Mdn_{\mathrm{B}} = 5$, $IQR_{\mathrm{B}} = 1.00$).

The check of assumption of normality within the comparison tests showed the data from Shape-filling was not normally distributed ($W = 0.961, p = .046$), and the p-value for Digit was above the 0.05 ($W = 0.963, p = .057$). Thus, we included

		Participants	Shapiro-Wilk (W)	Shapiro-Wilk (p)
Digit	Group A	28	0.913	0.024
	Group B	34	0.929	0.029
Shape-filling	Group A	28	0.927	0.053
	Group B	34	0.975	0.607
Corsi	Group A	27	0.937	0.105
	Group B	31	0.889	0.004
Restaurant	Group A	28	0.909	0.019
	Group B	34	0.890	0.002

Table 4.1: Results of normality test for group data in each task

	Test	Statistic	p-value	Effect size
Digit	Mann-Whitney	392	0.22	0.178
	Student's t-test	-1.12	0.27	-0.285
Shape-filling	Student's t-test	-0.37	0.71	-0.094
	Mann-Whitney	465	0.88	0.023
Corsi	Mann-Whitney	409	0.88	0.023
Restaurant	Mann-Whitney	430	0.5	0.097

Table 4.2: Group differences for each task

the results of both tests, Mann-Whitney and Student's t-test, for these two tasks. The additional Levene's test showed that groups in both tasks had homogenous variances ($F_{\text{digit}}(1, 60) = 3.3, p_{\text{digit}} = .074, F_{\text{shape-fill}}(1, 60) = 0.268, p_{\text{shape-fill}} = .607$).

The differences between groups were not significant and had negligible effect size except for the Digit task, where the effect size was weak (see Table 4.2). Therefore, we continued to analyze the data as one sample regardless of the group of participants.

4.1.2 Validity

To determine the convergent validity of the Restaurant Game, we compared its results with those of the Digit and Corsi tasks. To determine its divergent validity, we compared the results of the Restaurant Game to those of the Shape-filling task.

Overall, the results in Digit ranged from 4 to 9 ($Mdn = 7, IQR = 1.00$) and were not normally distributed ($W(62) = 0.932, p = .002$). The results of Shape-filling ranged from 1.4 to 2.65 ($Mdn = 1.80, IQR = 0.349$), and the Shapiro-Wilk test showed the data were not normally distributed ($W(62) = 0.960, p = .04$). The results in Corsi ranged between 3 to 8 ($Mdn = 6, IQR = 2.00$) and were also not normally distributed ($W(58) = 0.930, p = .002$). Finally, the results for the

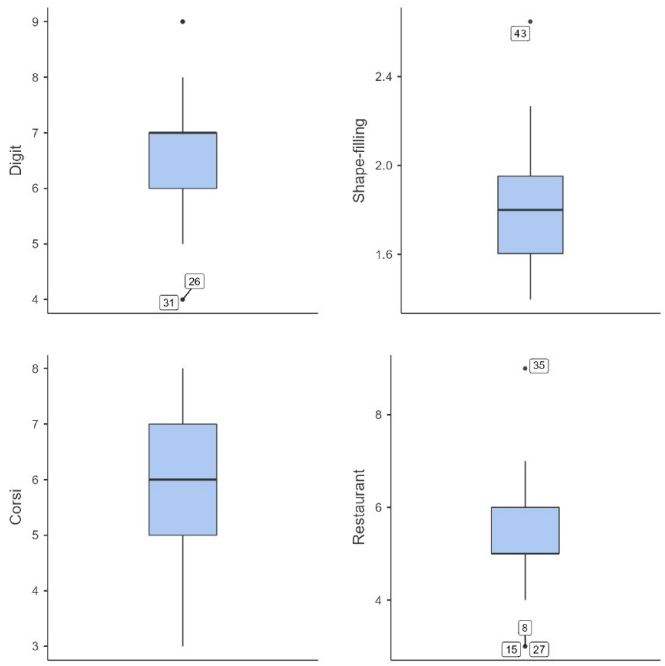

Figure 4.1: Results of the tasks.

Restaurant ranged from 3 to 9 ($Mdn = 5, IQR = 1.00$) and were not normally distributed ($W(62) = 0.906, p < .001$). The tasks showed 7 outliers altogether. We decided to keep them in, as we used non-parametric tests, and excluding these outliers would create new ones, eventually reducing the sample significantly. The results of the analyses of task results are shown in Figure 4.1.

The results of one-sided tests of Spearman's correlation coefficient (testing whether the correlation is positive) for all task pairs are shown in Table 4.3 and suggest that only Digit and Restaurant were significantly correlated with moderate effect size. Other correlations were not significant, with weak or negligible effect size.

4.1.3 Reliability

To determine the test-retest reliability of the Restaurant, we compared its data from the first data collection with its data from the second collection. The participants played the game for the second time within an interval of 20 and 39 days after the first part. The reliability analysis was done on a sample consisting of participants who participated in both parts of the study.

We decided to expand our focus from the game results to game scores (i.e., the size of the largest successfully remembered order and total number of points collected, respectively). Even though they are closely related, the scores are a more detailed measure. The results of both metrics from both parts of the study can be seen in Figure 4.2.

		Digit	Shape-filling	Corsi	Restaurant
Digit	Spearman's rho	-			
	p-value	-			
Shape-filling	Spearman's rho	0.03	-		
	p-value	0.41	-		
Corsi	Spearman's rho	0.19	-0.14	-	
	p-value	0.08	0.85	-	
Restaurant	Spearman's rho	0.45	0.04	0.20	-
	p-value	<.001	0.37	0.06	-

Table 4.3: Intercorrelations of the tasks

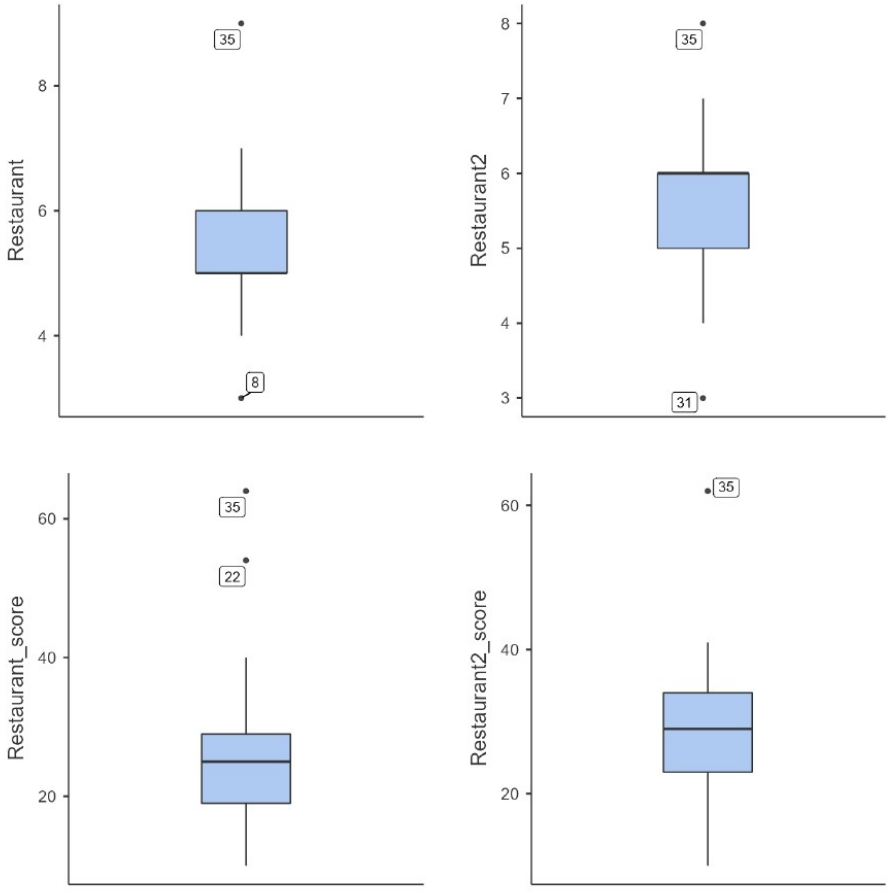

Figure 4.2: Results and final score of the Restaurant Game in the first and second part of the study.

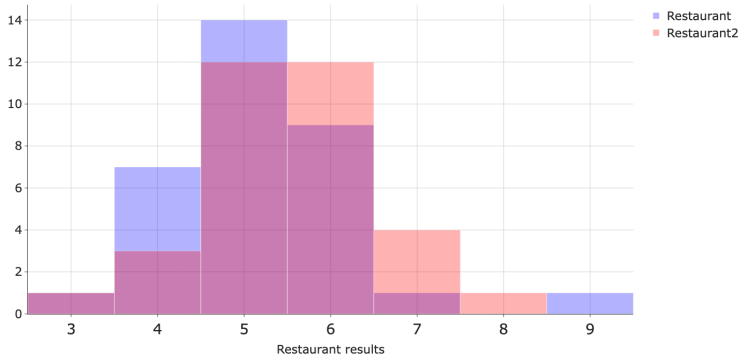

Figure 4.3: Distributions of Restaurant Game results from both parts of study.

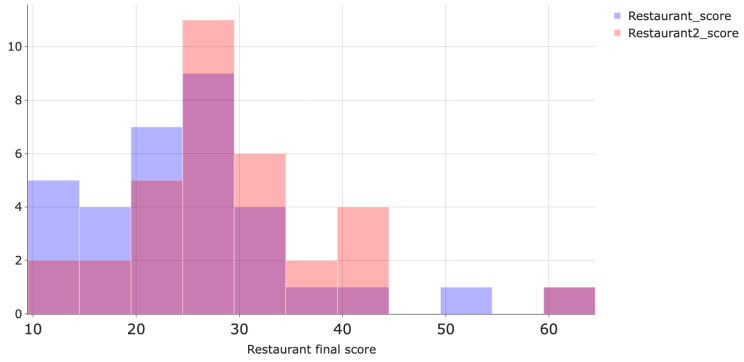

Figure 4.4: Distributions of Restaurant Game final scores from both parts of study.

The results for the first round ranged from 3 to 9 ($Mdn = 5, IQR = 1.00$). Shapiro-Wilk test showed the data were not normally distributed ($W(33) = 0.864, p < .001$). In the case of the second round, the data ranged from 3 to 8 ($Mdn = 6, IQR = 1.00$), and the data were also not normally distributed ($W(33) = 0.922, p = .021$).

When looking at the final scores, in the first round, the points ranged from 10 to 64 ($Mdn = 25, IQR = 10.0$). Shapiro-Wilk test showed the data were not normally distributed ($W(33) = 0.886, p = .002$). In the second round, the final scores ranged from 10 to 62 ($Mdn = 29, IQR = 11.0$) and the data were not normally distributed ($W(33) = 0.933, p = .043$).

To better understand the distribution of the data, we visualized them as overlapping histograms for easier comparison. In Figure 4.3, we see the game results, and in Figure 4.4, we see the final scores from both parts of the study. In both figures, we can see the second round data to be seemingly higher.

Spearman's test showed that there was a weak correlation of the game results between the two rounds of playing the game, which was not statistically significant ($rs(33) = 0.149, p = .407$). Similarly, a non-significant weak correlation could be observed for the final scores ($rs(33) = 0.150, p = .403$).

		Group	Mdn	IQR	Shapiro-Wilk (W)	Shapiro-Wilk (p)
Engagement	Digit	A	6	2.00	0.852	.001
		B	6	1.00	0.866	<.001
	Shape-filling	A	6	2.00	0.831	<.001
		B	6	1.00	0.847	<.001
	Corsi	A	6	2.00	0.898	.012
		B	6	2.00	0.862	<.001
	Restaurant	A	6	1.00	0.810	<.001
		B	6	1.00	0.838	<.001
Motivation	Digit	A	6	2.00	0.849	<.001
		B	6	1.75	0.858	<.001
	Shape-filling	A	6	2.00	0.787	<.001
		B	6	1.00	0.864	<.001
	Corsi	A	5	2.50	0.881	.005
		B	6	2.00	0.833	<.001
	Restaurant	A	7	1.00	0.742	<.001
		B	6	2.00	0.848	<.001

Table 4.4: Results of the summary statistics for engagement and motivation report for group data in each task

4.2 User experience

We assumed the newly developed Restaurant Game would provide a better experience than the classic assessments. We asked participants to self-report their engagement and motivation to test this assumption.

4.2.1 Group differences in user experience

The participants were divided into two groups, with the tasks arranged in different order. Although this division was intended to eliminate the order effect, we were interested to see if there were any differences between groups in the experience reports.

Table 4.4 shows the summary statistics of reports on engagement and motivation for both groups for all tasks, suggesting some potential differences in motivation. The Shapiro-Wilk normality test revealed that none of the data were normally distributed.

The differences between groups were not statistically significant and had weak effect size, even though the difference in engagement in the Corsi task was almost significant (see Table 4.5). We continued to analyze data regardless of the group of participants.

		Statistic	p-value	Effect size
Engagement	Digit	428	0.477	0.101
	Shape-filling	382	0.160	0.199
	Corsi	299	0.055	0.286
	Restaurant	430	0.491	0.100
Motivation	Digit	397	0.244	0.167
	Shape-filling	375	0.137	0.212
	Corsi	340	0.207	0.188
	Restaurant	374	0.125	0.214

Table 4.5: Group differences in engagement and motivation for each task

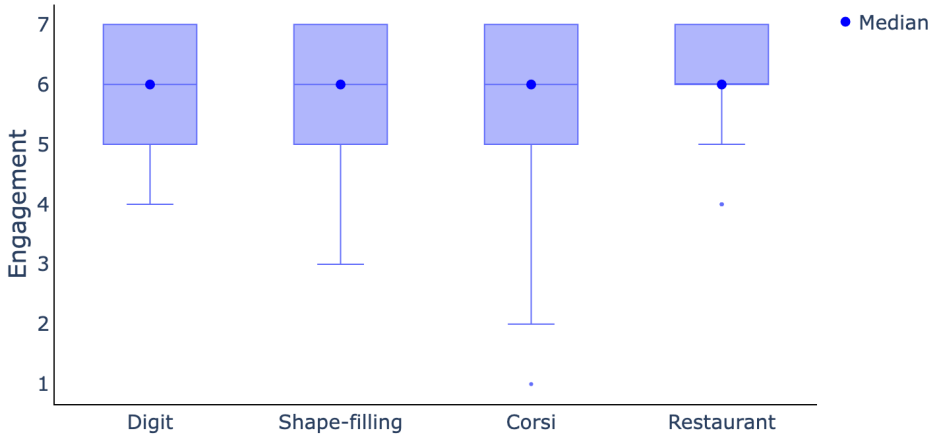

Figure 4.5: Engagement for each task.

4.2.2 Engagement

In the final questionnaire, participants evaluated their engagement for each task on a seven-point Likert scale. Among the three PsyToolkit tasks, the highest engagement was reported for the Shape-filling, the slightly lower in the Digit, and the Corsi was the least engaging one. The engagement of the Restaurant Game was higher than in the PsyToolkit tasks. We report on the mean, as median did not suggest any differences. The results of the analysis are shown in Table 4.6 and Figure 4.5.

Friedman's analysis of variance (ANOVA) showed a significant difference in engagement among the tasks ($\chi^2(3) = 9.98, p = .019$). Further pairwise comparisons showed that the engagement was significantly higher for the Restaurant than the Corsi, and there were no significant differences for other pairs of tasks (see Table 4.7).

	Min	Max	Median	IQR	Mean	SD
Digit	4	7	6	1.75	5.88	0.931
Shape-filling	3	7	6	2.00	5.9	0.931
Corsi	1	7	6	1.75	5.47	1.44
Restaurant	4	7	6	1.00	6.12	0.88

Table 4.6: Summary statistics on engagement for each task.

		Engagement		Motivation	
		W	p	W	p
Restaurant	Digit	1.655	0.100	2.562	0.011
Restaurant	Shape-filling	1.568	0.119	2.562	0.011
Restaurant	Corsi	3.224	0.002	2.904	0.004
Digit	Shape-filling	0.087	0.931	0	1.000
Digit	Corsi	1.568	0.119	0.342	0.733
Shape-filling	Corsi	1.655	0.100	0.342	0.733

Table 4.7: Differences in engagement and motivation among the tasks

4.2.3 Motivation

In addition to reporting on engagement, we asked participants to self-report their motivation for each task on a seven-point Likert scale. The motivation among the PsyToolkit tasks was highest for Corsi, slightly lower for Shape-filling, and slightly lower again for Digit. Similar to the engagement results, the Restaurant was the most motivating overall. We report on the mean, as median did not suggest any differences. The results of the analysis are shown in Table 4.8 and Figure 4.6.

Friedman's ANOVA showed a significant difference in motivation among the tasks ($\chi^2(3) = 10.4, p = .015$). Further pairwise comparisons showed that motivation was significantly higher in Restaurants compared to all PsyToolkit tasks (see Table 4.7).

	Min	Max	Median	IQR	Mean	SD
Digit	3	7	6	1.75	5.62	1.21
Shape-filling	3	7	6	1.00	5.66	1.15
Corsi	3	7	6	2.00	5.69	1.22
Restaurant	4	7	6	1.00	6.16	0.875

Table 4.8: Summary statistics on motivation for each task.

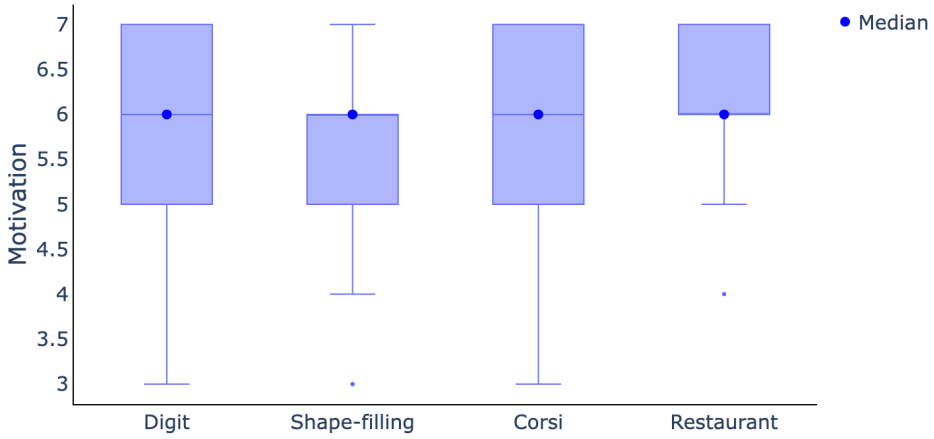

Figure 4.6: Motivation for each task.

4.2.4 Qualitative reports

In addition to reports on engagement and motivation, we asked for additional comments on the Restaurant and overall evaluation of the participants' experience among the tasks. These are the qualitative data that complement the quantitative ones primarily used to test the hypothesis on user experience.

Overall, 14 participants provided an answer. In six cases, the answer included a positive evaluation of the Restaurant and its instructions. Another four participants reported technical issues with the game. However, to our knowledge, these issues did not prevent the participants from playing the game. Another four answers included suggestions for further game development, such as including the possibility of taking a break or explaining the scoring system in more detail.

When asked about the overall experience while doing the tasks, 44 participants positively reported their experience with the tasks. From these, 10 stressed out they liked the Restaurant more than the other tasks, 6 positively commented on the short duration of the study, and 4 praised clear instructions of the tasks. Among the positive comments, participants mentioned they liked when they got feedback and could see their memory capacity, and that they felt motivated and attracted by colorful animations in the Restaurant. One of the participants reported: "I enjoyed the restaurant task way more than any tasks before in the Psytoolkit, it had great visuals and sort of a purpose, which made me more motivated for it than the tasks before."

On the other hand, approximately 25% of the reports included something negative, such as difficulty of the tasks, insufficient rewards, complications with the Corsi, or short appearance of stimuli. Some reported decrease in their attention or tense feelings. One explicitly mentioned they felt they would be stupid if they did not perform well.

4.3 Additional analyses

As we gathered a lot more data, we used them for more specific analyses to help determine further research interests and questions. Below, we describe some of the more interesting and relevant results.

4.3.1 Alternative final scores

Determining the participant's performance is a metric that we defined in the beginning. However, this metric could be chosen differently, so we considered using a different one. What eventually interested us were the game results, which we believed to reflect the working memory. When we wanted to estimate the game result based on the final scores, we could not calculate it precisely. The reason for that is that the score of 20 could be achieved by scoring $2, 2, 3, 4, 4, 5$ with result 5 but also by scoring $2, 3, 4, 5, 6$ with result 6. Therefore, we used an approximation of the game results calculated as the square root of the final scores.

To show why a squared root is a suitable option, we start by defining the range of final scores for each game result. If the participant scored just once for each order size, which was a necessary requirement to increase the order size, they would score $\frac{n-1}{2} * (n + 2)$. If they assembled all orders correctly, they would get $(n - 1)^*(n + 2)$ points. These two calculations determine a minimum and a maximum for final scores for each game result (n). The approximation of both is n^2 as it has the same order of magnitude as the values of the original range. In order words, it is still a power of n. From this approximation, we easily calculate the game result as a square root and get a more continuous metric than the original game results.

After calculating the approximate game results, we got new data to analyze. The new data for the first round ranged from 3.16 to 8 ($M = 5.06, SD = 1.03$), and their distribution was not statistically different from normal ($W(33) = 0.950, p = .133$). The final scores of the second round ranged from 3.16 to 7.87 ($M = 5.37, SD = 0.88$), and their distribution was not statistically different from normal ($W(33) = 0.960, p = .255$).

In this case, Pearson's test showed a statistically significant correlation with a moderate size effect ($r(33) = 0.402, p = .02$). This result differs from the one obtained from a non-parametric test and would show the expected correlation. In other words, the alternative final scores would support our hypothesis on the game's reliability. However, we want to explicitly state that this result came from an approximation.

The advantage of this approximation was that after transforming the data, their distribution was not statistically different from normal. Originally, the data were not distributed, and we had to use the non-parametric test. This type of test usually requires more data points than parametric ones to reach significance.

We conducted an a priori power analysis with IBM SPSS Statistics software to calculate the necessary sample size for correlation analysis. The power value we required was 0.8, and the significance level was 0.05. We determined the correlation coefficient to be 0.5, the lower boundary of large effect size. If we expected a lower correlation, the sample size necessary to observe it would be even higher. The power analysis showed that using Pearson's correlation coefficient required at least 29 participants and using Spearman's correlation coefficient required 33. Our sample of 33 participants may seem to fulfill the criteria. However, it is important to note our sample had four outliers. Therefore, our sample was probably not sufficient for testing either of the correlation coefficients. However, we had a slightly higher chance of getting a significant Pearson's correlation coefficient than getting a significant Spearman's correlation coefficient, which was also the case in our empirical results.

4.3.2 Gaming and waitressing experience

In this section, we examine the potential research direction considering the influence of previous gaming and waitressing experiences on the Restaurant results. The reports on waitressing and gaming were gathered as qualitative data, which we transformed into quantitative data. For the gaming experience, we used $0, 1,$, and 2, which correspond to no recent experience, plays sometimes, and plays regularly. For the waitressing experience, we also used $0, 1$, and 2 corresponding to no experience, short-term experience waitressing or other related experience, long-term experience.

This division resulted in 44 participants with no recent gaming experience, 12 with some, and 6 with regular gaming experience. For the waitressing, there were 31 with basically no (recent) experience, 25 with some, and 6 with long-term experience. To achieve reasonable group sizes, we decided to distinguish only between participants with no experience and those with some, regardless of extent. The resulting groups consisted of 44 and 18 participants for gaming and 31 and 31 participants for waitressing.

It is important to state that we consider the analyses and results in this section to have just informative value. The reasons are that we transformed the self-report data of each participant, which is subjective and may not correspond to the quantitative answers the participants would provide if asked. Additionally, our samples were quite small, and in the example of gaming, they had significantly different sizes. Therefore, we analyzed the data just to explore possibilities for future research.

A summary of statistics for all four tasks is shown in Table 4.9 for gaming groups and in Table 4.10 for waitressing groups. Results suggest no clear advantage of either experience. We chose appropriate tests for between-subject comparisons based on the normality of the data. The only case of data with distributions that were not statistically different from normal in both groups was in Shape-filling for gaming experience.

	Expe-rience	Min	Max	M	SD	Mdn	IQR	Shapiro-Wilk (W)	Shapiro-Wilk (p)
Digit	No	4	9	6.73	1.28	7	2.00	0.940	0.024
	Yes	5	9	6.50	1.20	6	1.00	0.858	0.012
Shape-filling	No	1.40	2.27	1.77	0.21	1.75	0.32	0.974	0.424
	Yes	1.47	2.65	1.90	0.29	1.91	0.33	0.943	0.326
Corsi	No	3	8	5.85	1.23	6	2.00	0.937	0.028
	Yes	4	8	5.83	0.99	6	1.00	0.914	0.099
Restaurant	No	3	7	5.23	1.03	5	1.00	0.913	0.003
	Yes	3	9	5.28	1.23	5	1.00	0.817	0.003

Table 4.9: Summary statistics of results for each of the tasks depending on the recent gaming experience

	Expe-rience	Min	Max	M	SD	Mdn	IQR	Shapiro-Wilk (W)	Shapiro-Wilk (p)
Digit	No	5	9	6.84	1.34	7	2.00	0.892	0.005
	Yes	4	9	6.48	1.15	7	1.00	0.929	0.041
Shape-filling	No	1.40	2.27	1.80	0.21	1.81	0.26	0.983	0.878
	Yes	1.46	2.65	1.81	0.27	1.77	0.45	0.917	0.020
Corsi	No	4	8	5.86	1.03	6	2.00	0.901	0.010
	Yes	3	8	5.83	1.28	6	1.00	0.927	0.046
Restaurant	No	3	9	5.58	1.18	6	1.00	0.886	0.003
	Yes	3	7	4.9	0.87	5	1.00	0.852	$<.001$

Table 4.10: Summary statistics of results for each of the tasks depending on the waitressing experience

			Statistic	p-value	Effect size
Mann-Whitney	Digit	Gaming	340	0.370	0.143
		Waitressing	428	0.450	0.109
	Shape-filling	Waitressing	461	0.791	0.041
	Corsi	Gaming	354	0.924	0.017
		Waitressing	414	0.923	0.016
	Restaurant	Gaming	389	0.909	0.019
		Waitressing	287	0.004	0.404
T-test	Shape-filling	Gaming	-1.98	0.053	-0.553

Table 4.11: Differences in task results based on gaming and waitressing experience

Levene's test showed that the gaming groups in Shape-filling had homogenous variances ($F(1.60) = 0.793, p = .377$). The result of the Student's t-test suggested the differences between gaming groups were of moderate effect size but not significant, even though the difference was almost significant ($t(60) = -1.98, p = .053, d = -0.553$).

In other cases, we used the Mann-Whitney test. The results suggested that participants who did not have the waitressing experience had better results in Restaurant than those who had waitressing experience (see Table 4.10), and this difference was significant with a moderate effect size (see Table 4.11). Additionally, participants with gaming experience performed better in Shape-filling than participants without any recent experience (see table 4.9), and this difference was almost significant with a moderate effect size (see Table 4.11).

4.3.3 Strategies

When analyzing strategies, we used answers from all 70 participants who filled in the final questionnaire during the second data collection. Even if they did not complete one of the tasks or did not fulfill the language requirements, they could still report on their experience, and we had no reason to exclude their answers in this analysis.

Of the 70 participants, only 18 reported they did not use any strategy during any of the tasks. Of those who used some strategy, only 3 participants mentioned any strategy during the Shape-filling. These strategies were only remembering for which stimulus the participant is supposed to press one of the keys and saying out loud the part of stimulus they need to focus on (e.g., saying word square when a square is shown and the condition is focused on shape), which was mentioned twice. These participants used strategies for other tasks as well, which means almost 75% of participants did use some strategy to help them memorize longer sequences.

The number of participants mentioning any strategy for the Corsi was 4, for the Digit, it was 31, and for the Restaurant, it was 35. Additional 8 answers were not

categorized as it was not clear to which task the answers belonged. However, they mentioned using a visual representation of the information in 3 of the cases and repetition of the presented information in 5 of the cases.

We start with the Corsi, as the other two working memory tasks had more similar strategies. In this case, 3 people tried to create and remember a path of the squares corresponding to the sequence. Interestingly, one participant reported they tried to use their piano skills to remember the sequence of squares.

For the Digit, 11 participants reported they tried to group the digits, for example, into pairs, dates, or according to the phone number structure. Repeating the stimuli was mentioned 19 times, from which 9 corresponded to repeating the information aloud, 6 to repeating it verbally in mind, and 1 to visually repeating the information. There were 3 participants who explicitly mentioned they repeated the sequence of digits in the forward order aloud several times while inputting the digits from the last to the first. Two other participants mentioned more unique strategies and those were using for each digit a finger on their hand and using the keyboard to create a path among the digits. Interestingly, the second was one of our arguments for not having the Digit and Corsi one after another.

Lastly, for the Restaurant, only 4 participants mentioned some type of grouping strategy, for instance, based on the food type like breakfast food or creating meals. Repeating the stimuli was altogether mentioned 24 times from which 12 participants repeated the order out loud, 7 in their minds, and same as in the Digit, 3 participants explicitly mentioned repeating the order in the original order while assembling the answer drinks first, food second. The last 2 participants sang the dishes or used the first two letters of each item to create something like a song. Another group of strategies includes mental visual representation, which was reported by 6 participants, out of which one imagined their family making the order. The last 2 participants mentioned more unique strategies. The first one was picking one finger on their hand for each of the items, which was already mentioned among the strategies for the Digit. The second one was creating a story behind the order.

Chapter 5

Discussion

In this chapter, we interpret the results and discuss the study's potential limitations. As this was just a pilot validation study, we end this chapter with quite an extensive section on possibilities for further research and game development. This last section provides guidance to achieve our ultimate goal, which is to develop a new training method that is more aligned with theoretical findings and limits the shortcomings of the current measuring and training options. We continue to use the shorter names of the tasks for simplicity.

5.1 Interpretation of results

The study consisted of two parts. In the first, the participants were asked to fill in the questionnaire with their demographics and completed four tasks (Digit, Shape-filling, Corsi, Restaurant). Finally, they reported on their experience in the second questionnaire. In the second part of the study, participants were asked to play the Restaurant Game again.

In the first part of the study, we divided the participants into two groups that differed in the order of the tasks. The aim was to test if the order influenced the results. The analysis showed no significant differences between the groups, suggesting the order of the tasks did not influence the task results. Further on, we analyzed the data from both groups combined into one sample.

To test the Restaurant's convergent validity, we studied correlations between its results and the results of backward versions of the Digit and Corsi tasks. We expected positive correlations, which were supported only in the case of the Digit, even though the p-value in the case of Corsi was close to the alpha value .05. To test the Restaurant's divergent validity, we studied correlations between its results and the results of the Shape-filling. In this case, the results were not significant with negligible effect size, which was in accordance with our expectations.

We expected Corsi to have a significant positive correlation with the Restaurant. Even though the results do not convincingly support our hypothesis, the same results were obtained for Corsi with Digit. We know Corsi and Digit both reflect working memory, but despite that, their correlation was not significant. Therefore, the insignificant correlation between Corsi and Restaurant may not discredit Restaurant as a working memory measure.

We know that Digit results depend on the span of the phonological loop, and the Corsi results on the span of the visuo-spatial sketchpad. In the former, participants typically verbalize presented digits, which is why the phonological loop is a better candidate for their processing. In the latter, the information about the spatial location of the visual stimulus that cannot be verbalized is processed. Therefore, another explanation for the low, insignificant correlation between the tasks may be the different nature of underlying processes and the fact they reflect different components of the working memory model by Baddeley and Hitch (1974). There was a moderate correlation between Digit and Restaurant, which shows the similarity in the processes underlying both tasks. Therefore, a similar explanation to the one used for Digit and Corsi could be used to explain the low correlation between Corsi and Restaurant. However, the relatively low correlation also suggests that a lot of processes are unique to the tasks.

Opposed to the Digit and Corsi, Restaurant relies on both types of information. The order is presented as a sequence of item names, which is verbal information. If we do not assume visual strategies mentioned only by 3 of our participants, the remembered information is more likely verbal and stored in a phonological loop. This may explain why the correlation is higher for Restaurants with Digit than Corsi. However, when assembling the order, the participants had to choose the correct icons, which is visual information. Therefore, there's some visual processing, even though it may not require storage. Additionally, the game requires some language processing to understand what is being ordered and information binding to account for the "translation" from the item's name to its icon.

The information binding, which is an essential process in the game, could happen in the episodic buffer. This component of the working memory model is assumed to focus on integrating information from different sources (Baddeley, 2000). In our case, binding of the visual and verbal representation of an item is necessary, which is a cross-modal integration. This type of integration was suggested to happen in an episodic buffer with resources of the central executive, compared to the unimodal binding that was suggested to happen before entering the episodic buffer (Allen et al., 2012; Nobre et al., 2013). To conclude, the Restaurant includes processes that are believed to define episodic buffer, and therefore, the game may even reflect this component.

We briefly summarize our findings to answer the research question about validity of Restaurant Game. When testing convergent validity, both correlations were positive, as

we expected. However, one did not reach statistical significance and had a lower effect size, which can be explained. The hypothesis on divergent validity was supported.

In a final questionnaire, the participants reported on their subjective experience during the tasks. We chose engagement and motivation as specific measures to reflect it. In both cases, the overall differences among the tasks were significant. While the Restaurant Game made participants feel more motivated than any of the other tasks, the difference in engagement was significant only between the game and Corsi. The results suggest that Corsi was the least engaging task, which may also be due to the few technical difficulties some participants encountered and its relatively raw interface.

The results support our hypothesis that motivation is higher for the Restaurant than for other tasks. These results are aligned with the results by Vermeir et al. (2020). The hypothesis on the engagement is not totally supported as only Corsi had significantly lower engagement reports. However, when we analyzed the qualitative data, the Restaurant was praised in the additional comments the most. Overall, the results show that the user experience for Restaurant Game was better than for the other tasks, indicating that the effort we put into developing the game according to the known guidelines and user feedback was fruitful.

The summary statistics (mean, range) suggest the engagement was higher in Restaurant than in the Digit and Shape-filling. However, the sample size may be too small to show statistical significance. Additionally, our sample consisted of first-year psychology students who could generally be excited to participate in a study using all the measures they had been learning about. The excitement was indicated in their verbal feedback as 44 participants reported positively on their experience with tasks. Therefore, their engagement may reflect their overall engagement with the study itself.

To test the test-retest reliability of our task, we repeated the Restaurant measurement after approximately one month. The correlation between the measures was weak and not significant, which does not support our hypothesis. A possible explanation for that may be that the game was not novel to participants in the second part of the study, which Rabbitt (2004) suggested as a reason for common low test-retest reliabilities.

Another explanation for the results suggesting that the Restaurant is not a measurement with stable results is the influence of the environment in which the participants played the game. In the first round, they also did three other tasks, while in the second, only one. Additionally, both parts of the study happened in uncontrolled settings, as the participants were asked to do the experiments in their free time. Lastly, we only had 33 participants, which limits the generalization of results.

The stability results may have been different if the participants in our sample had more diverse memory abilities or if we had a bigger sample size. Even choosing a different formula for determining the final score might have deemed the Restaurant Game stable, as explored in Section 4.3.1.

5.2 Limitations of study

As we did a pilot validation study, it had several limitations. We divide them into two categories, with the first being about the sample and the second about the methodology.

Our sample was homogenous, including only young people at the peak of their cognitive abilities, without any cognitive impairments, who speak Slovene as their first language, are psychology students in the same year at the same university, and are mostly women. The homogeneity of the sample limits the generalization of the results.

Additionally, as only around 8% of our participants were males, we can not account for the possible gender differences. Even though Grissom and Reyes (2019) suggests there are no significant differences in the performance of executive functions between men and women, the specific setting of games was not properly studied.

Methodologically speaking, we chose quite an unorthodox approach when we opted to ask participants to do the experiment in their free time. Therefore, we cannot account for the influence of the time or setting in which they did the experiment. However, we had a within-subject design, and the setting was most likely stable during the whole duration of the first part of the experiment for each participant. Therefore, we believe the confounding variables had the same influence across the tasks. In other words, if somebody did the experiment tired and their results were worse than they would be under different circumstances, the results for each task were influenced similarly. We tried to control the influence of exhausting mental resources by dividing participants into two groups with different orders of the tasks. Therefore, we believe the setting should not significantly affect the correlations and comparisons studied.

With the outlook for the future, if the Restaurant proves to be valid and appropriate for cognitive training, people will play it in non-laboratory settings. In this sense, the study was done under more probable and natural circumstances for each participant. Interestingly, this may be a limitation for the PsyToolkit tasks that were designed and previously used in the controlled environment.

Digit span task (backward) is commonly used in clinical neuropsychology to measure working memory, which suggests the task effectively distinguishes healthy people and people with frontal lobe dysfunction (Jurado & Rosselli, 2007; Miyake et al., 2000). Its validity for healthy individuals may be discussed even though the task is commonly used even in non-clinical studies. Additionally, our implementation of the task makes it possible to measure the verbal memory span only up to nine items, as the digits in a stimuli sequence do not repeat. As six of our participants remembered the sequence of nine items, we can not be sure their span is not higher.

When preparing the Restaurant Game for the study, we adjusted it to the specifics of our population to increase the user experience. However, it limits the generalization of the results. If we were to develop a game suitable for different populations, we would

need extensive UX testing that was out of our scope. Additionally, we would need to consider the research findings again. For instance, if somebody has difficulties with visual inspection, it could take them longer to identify the correct icons, which could make them forget part of the order.

Overall, the results could differ if we used different measurements, as each one taps a different set of processes. Only an in-depth understanding of each measurement would provide us with a better understanding of processes tapped by the Restaurant, on which we can now only hypothesize. Therefore, we study correlation, not causation.

The limitation of the user experience report is the predictability of our hypothesis. As our sample consisted of psychology students, they were all familiar with the standard measurements and could potentially guess the aim of the study. As they were asked to participate in the study by their professor, who is in an authority position, the students may be biased in reporting more positively on the Restaurant Game.

For the second part of the study, in which we tested the test-retest reliability, we did not consider the influence of different delays on results. The small sample size for this part of the study is also an important limitation.

Lastly, our results include outliers that we did not address. We could use robust statistical methods or choose other methods to encounter for their influence other than excluding them from the sample. Therefore, the results may be a little skewed. However, we used non-parametric tests to lower their impact on the results. Since this was a pilot study and the sample was not large, we decided not to use more sophisticated statistical methods for now.

5.3 Further research

The results of this study did not completely support our hypotheses on the validity and reliability of Restaurant Game. However, experience results suggest there is potential from the user's point of view, and the additional analysis of approximated data suggests there is potential for better reliability results. We need to understand the underlying processes better before repeating a validation study under improved conditions.

As working memory is closely linked with the updating function, studying the relationship between the game and updating tasks could provide information on the game's dependence on storage components. Another potential direction for further research is studying the game with a focus on the episodic buffer. This component of the working memory model is not as well studied as the phonological loop and visuo-spatial sketchpad, and there is an ongoing discussion on its measurements (Nobre et al., 2013). However, as it is not clear what cognitive processes underlie Restaurant Game, studying its correlations with different types of tasks may help us understand it better.

It may even be beneficial to study the game under a different framework. We chose the working memory model by Baddeley and Hitch (1974) as our reference model. However, other models could be used, among the more prominent ones is the model by Cowan (1999). His Emebedded-process model is by Baddeley, viewed as "I see Cowan's model as principally concerned, in my terminology, with the link between the CE (central executive) and the episodic buffer." The two models are not mutually exclusive but differ in terminology and the focus of research, which could bring unique insights into our understanding of the game's processes.

The big potential for further research is in the population of the game's potential users. Our research sample was homogenous, and the game should be tested for different groups of participants, such as a general healthy population of different ages, neurological patients, and children.

Interestingly, we observed that participants without prior waitressing experience performed better in Restaurant Game than participants with relevant experience. As our analysis on this topic was preliminary, with too many limitations and subjectivity, this direction needs to be further studied to better understand the reason for this counterintuitive result.

Another direction of research lies in the Restaurant Game itself. Researchers may need to adjust the game's parameters to suit their needs. For example, previously, we decided to change the overall number of items but could not say what our decision's influence was on the game's overall difficulty. To define which aspects of the game should be adjustable and what the effect of each adjustment would be, we would need to do many more studies, each focusing on one of the parameters. To name just a few more, we could ask how the results would change if we changed the time for which one item in an order is shown, how they would change if we had only fourteen items, how if the stimuli were audio and not written, or what results we would get from different game versions. To name just a few examples of possible game versions, they can be forward with repeating the items in the same order in which they were placed, backward in which the items are repeated from the last item placed to the first one, or version in which drinks go first and food second but we do not enforce specific order within each of the categories.

We have already implemented a forward and backward version in addition to the versions presented throughout this thesis. The versions can be used separately or together to derive the final result. It would also be interesting to see how the results of different versions correlate in comparison to the correlations between versions of the Digit. For example, in Wechsler Adult Intelligence Scale, the Digit Span subtest scaled score is derived after administering three different task versions (Raiford et al., 2010).

In the direction of increasing the game's engagement, other game elements should be considered. In our UX testing, two participants reported being competitive and more

engaged and motivated if they could compete with themselves or others. Different game elements were shown to be effective for different groups, and their additional implementation may help to improve the overall user experience (Lumsden et al., 2016).

If the Restaurant Game proves valid and reliable for measuring working memory and has a good user experience, it could be used in future studies for different purposes. For instance, it could be used in other studies focusing on working memory, help neuropsychological diagnostics, or be further developed to serve cognitive training.

As our ultimate goal is to develop cognitive training with different modules focusing on different executive functions, further development may also focus on developing a new task for measuring and training other cognitive functions. The functions chosen next will be different and have their specifics. Despite that we believe that the experience and knowledge we gained from developing the Restaurant Game will be helpful in further development of a comprehensive diagnostic tool providing measures of different executive functions and their interrelations.

–

Chapter 6

Conclusion

In this thesis, we aimed to test the validity and reliability of the newly developed Restaurant Game and study its user experience. This study is just the first step to a much bigger goal, which is to create new cognitive training based on research and user validation while addressing the shortcomings of current options.

We started with UX testing to gather feedback on the game's original version and used the results to increase the user experience of the game. We also made adjustments based on research needs to make the game better suited for testing our hypotheses.

In this thesis, we took a more modern approach to developing a new method for measuring and potentially training executive functions. Proof of that is that we considered user experience, a field that is not often included in academic research. The research papers studying some of its aspects (mainly linked to the gamification of tasks) show promising results, which is why it was an inseparable part preceding our main research. Even though our results are insufficient for the Restaurant Game to be used as a working memory measure right now, the study showed promising results in the game's validity and potential for reliability. We believe the game development should continue. As this was just the first study of the game's psychometric properties, there are numerous limitations to our research design. However, there are even more possibilities for further development of the game and further research.

In this thesis, we showed that even though we did not have a big research team or financial support, it was possible to develop a completely new, research-based, user-supported method that addresses the shortcomings of commonly used psychological tasks. Numerous improvements could be made, but based on the experiences gained in this study, they would not require too extensive financial resources.

The human population is, on average, getting older, and many older people face health issues related to the poor functioning of executive functions. We need to focus on developing interventions that are supported by research and enjoyed by users and that effectively mitigate the negative effects of executive function decline.

Bibliography

Aben, B., Stapert, S., & Blokland, A. (2012). About the distinction between working memory and short-term memory. *Frontiers in Psychology*, *3*(301). https://doi.org/10.3389/fpsyg.2012.00301

Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of working memory: Differences in definition, degree of modularity, role of attention, and purpose. *Language, speech, and hearing services in schools*, *49*(3), 340–355.

Alexander, M. P., & Stuss, D. T. (2000). Disorders of frontal lobe functioning. *Seminars in neurology*, *20*(04), 427–438.

Allen, R. J., Hitch, G. J., Mate, J., & Baddeley, A. D. (2012). Feature binding and attention in working memory: A resolution of previous contradictory findings. *Quarterly journal of experimental psychology*, *65*(12), 2369–2383.

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology*, *8*(2), 71–82.

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., Kong, E., Larraburo, Y., Rolle, C., Johnston, E., et al. (2013). Video game training enhances cognitive control in older adults. *Nature*, *501*(7465), 97–101.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (pp. 89–195). Elsevier.

Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 5–28.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in cognitive sciences*, *4*(11), 417–423.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual review of psychology*, *63*, 1–29.

Baddeley, A. (2020). Working memory. In *Memory* (pp. 71–111). Routledge.

Baddeley, A., & Hitch, G. (1974). Working memory. *The Psychology of Learning and Motivation: Advances in Research and Theory*, (pp. 47–89).

Banich, M. T. (2009). Executive function: The search for an integrated account. *Current Directions in Psychological Science*, *18*(2), 89–94. https://doi.org/10.1111/j.1467-8721.2009.01615.x

Barkley, R. A. (2012). *Executive functions: What they are, how they work, and why they evolved*. Guilford Press.

Belleville, S. (2008). Cognitive training for persons with mild cognitive impairment. *International Psychogeriatrics*, *20*(1), 57–66.

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Bond, G. E., Wolf-Wilets, V., Fiedler, F. E., & Burr, R. L. (2001). Computer-aided cognitive training of the aged: A pilot study. *Clinical Gerontologist*, *22*(2), 19–42.

Brehmer, Y., Westerberg, H., & Bäckman, L. (2012). Working-memory training in younger and older adults: Training gains, transfer, and maintenance. *Frontiers in human neuroscience*, *6*, 63.

Brown, T. E., & Landgraf, J. M. (2010). Improvements in executive function correlate with enhanced performance and functioning and health-related quality of life: Evidence from 2 large, double-blind, randomized, placebo-controlled trials in adhd. *Postgraduate Medicine*, *122*(5), 42–51. https://doi.org/10.3810/pgm.2010.09.2200

Brydges, C. R., Fox, A. M., Reid, C. L., & Anderson, M. (2014). The differentiation of executive functions in middle and late childhood: A longitudinal latent-variable analysis. *Intelligence*, *47*, 34–43.

Burgess, P. W., Alderman, N., Evans, J., Emslie, H., & Wilson, B. A. (1998). The ecological validity of tests of executive function. *Journal of the international neuropsychological society*, *4*(6), 547–558.

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Laure, M. C., Dawson, D. R., Anderson, N. D., Gilbert, S. J., Dumontheil, I., & Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the international neuropsychological society*, *12*(2), 194–209.

Burgess, P. W., & Shallice, T. (1996a). Bizarre responses, rule detection and frontal lobe lesions. *Cortex*, *32*(2), 241–259. https://doi.org/10.1016/s0010-9452(96)80049-9

Burgess, P. W., & Shallice, T. (1996b). Response suppression, initiation and strategy use following frontal lobe lesions. *Neuropsychologia*, *34*(4), 263–272. https://doi.org/10.1016/0028-3932(95)00104-2

Buschkuehl, M., Jaeggi, S. M., Hutchison, S., Perrig-Chiello, P., Däpp, C., Müller, M., Breil, F., Hoppeler, H., & Perrig, W. J. (2008). Impact of working memory training on memory performance in old-old adults. *Psychology and aging*, *23*(4), 743.

Butler, M., McCreedy, E., Nelson, V. A., Desai, P., Ratner, E., Fink, H. A., Hemmy, L. S., McCarten, J. R., Barclay, T. R., Brasure, M., et al. (2018). Does cognitive training prevent cognitive decline? a systematic review. *Annals of internal medicine*, *168*(1), 63–68.

Caplan, B., DeLuca, J., & Kreutzer, J. S. (2010). *Encyclopedia of clinical neuropsychology*. Springer.

Chan, R., Shum, D., Toulopoulou, T., & Chen, E. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology*, *23*(2), 201–216. https://doi.org/10.1016/j.acn.2007.08.010

Corsi, P. M. (1972). Human memory and the medial temporal region of the brain.

Cowan, N. (1999). An embedded-processes model of working memory. *Models of working memory: Mechanisms of active maintenance and executive control*, *20*(506), 1013–1019.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87–114.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

Davis, J. C., Marra, C. A., Najafzadeh, M., & Liu-Ambrose, T. (2010). The independent contribution of executive functions to health related quality of life in older women. *BMC Geriatrics*, *10*(1). https://doi.org/10.1186/1471-2318-10-16

DeBattista, C. (2005). Executive dysfunction in major depressive disorder. *Expert Review of Neurotherapeutics*, *5*(1), 79–83. https://doi.org/10.1586/14737175.5.1.79

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, 9–15.

Diamond, A. (2005). Attention-deficit disorder (attention-deficit/hyperactivity disorder without hyperactivity): A neurobiologically and behaviorally distinct disorder from attention-deficit/hyperactivity disorder (with hyperactivity). *Development and psychopathology*, *17*(3), 807–825.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Diamond, A., & Ling, D. S. (2016). Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not. *Developmental cognitive neuroscience*, *18*, 34–48.

Doebel, S. (2020). Rethinking executive function and its development. *Perspectives on Psychological Science*, *15*(4). https://doi.org/10.1177/1745691620904771

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science*, *11*(1), 19–23.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Eslinger, P. J., & Damasio, A. R. (1985). Severe disturbance of higher cognition after bilateral frontal lobe ablation: Patient evr. *Neurology*, *35*(12), 1731–1731.

Fisk, J. E., & Sharp, C. A. (2004). Age-related impairment in executive functioning: Updating, inhibition, shifting, and access. *Journal of clinical and experimental neuropsychology*, *26*(7), 874–890.

Fisk, J. E., & Warr, P. (1996). Age and working memory: The role of perceptual speed, the central executive, and the phonological loop. *Psychology and aging*, *11*(2), 316.

Fournier-Vicente, S., Larigauderie, P., & Gaonac'h, D. (2008). More dissociations and interactions within central executive functioning: A comprehensive latent-variable analysis. *Acta psychologica*, *129*(1), 32–48.

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204.

Fuster, J. M. (2002). Frontal lobe and cognitive development. *Journal of neurocytology*, *31*(3-5), 373–385.

Grissom, N. M., & Reyes, T. M. (2019). Let's call the whole thing off: Evaluating gender and sex differences in executive function. *Neuropsychopharmacology*, *44*(1), 86–96.

Guarino, A., Favieri, F., Boncompagni, I., Agostini, F., Cantone, M., & Casagrande, M. (2019). Executive functions in alzheimer disease: A systematic review. *Frontiers in Aging Neuroscience*, *10*(437). https://doi.org/10.3389/fnagi.2018.00437

Harley, A. (2020). The principle of common region: Containers create groupings [Accessed: 2024-03-26]. https://www.nngroup.com/articles/common-region/

Hawkins, G. E., Rae, B., Nesbitt, K. V., & Brown, S. D. (2013). Gamelike features might not improve data. *Behavior research methods*, *45*, 301–318.

Hester, R., & Garavan, H. (2005). Working memory and executive function: The influence of content and load on the control of attention. *Memory & cognition*, *33*, 221–233.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental science*, *12*(4), F9–F15.

Hull, R., Martin, R. C., Beier, M. E., Lane, D., & Hamilton, A. C. (2008). Executive function in older adults: A structural equation modeling approach. *Neuropsychology*, *22*(4), 508.

Interaction Design Foundation. (2016a). What is human-computer interaction (HCI)? [Accessed: 2024-03-28]. https://www.interaction-design.org/literature/topics/human-computer-interaction

Interaction Design Foundation. (2016b). What is User Experience (UX) Design? [Accessed: 2024-03-26]. https://www.interaction-design.org/literature/topics/ux-design

Interaction Design Foundation. (2016c). What is User Interface (UI) Design? [Accessed: 2024-03-26]. https://www.interaction-design.org/literature/topics/ui-design

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short-and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, *108*(25), 10081–10086.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the n-back task as a working memory measure. *Memory*, *18*(4), 394–412.

jamovi project, T. (2024). Jamovi (version 2.5) [computer software]. https://www.jamovi.org

Jurado, M. B., & Rosselli, M. (2007). The elusive nature of executive functions: A review of our current understanding. *Neuropsychology Review*, *17*(3), 213–233. https://doi.org/10.1007/s11065-007-9040-z

Karbach, J., & Kray, J. (2009). How useful is executive control training? age differences in near and far transfer of task-switching training. *Developmental science*, *12*(6), 978–990.

Karbach, J., & Kray, J. (2016). Executive functions. *Cognitive training: An overview of features and applications*, 93–103.

Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. *Psychological science*, *25*(11), 2027–2037.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352.

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in cognitive sciences*, *14*(7), 317–324.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., Gillberg, C. G., Forssberg, H., & Westerberg, H. (2005). Computerized training of working memory in children with adhd-a randomized, controlled trial. *Journal of the American Academy of child & adolescent psychiatry*, *44*(2), 177–186.

Koivisto, J., & Malik, A. (2021). Gamification for older adults: A systematic literature review. *The Gerontologist*, *61*(7), e360–e372.

Kueider, A. M., Parisi, J. M., Gross, A. L., & Rebok, G. W. (2012). Computerized cognitive training with older adults: A systematic review. *PloS one*, *7*(7), e40588.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389–433.

Levine, B., Schweizer, T. A., O'Connor, C., Turner, G., Gillingham, S., Stuss, D. T., Manly, T., & Robertson, I. H. (2011). Rehabilitation of executive functioning in patients with frontal lobe brain damage with goal management training. *Frontiers Human Neuroscience*, *5*, 9.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., Munafò, M. R., et al. (2016). Gamification of cognitive assessment and cognitive training: A systematic review of applications and efficacy. *JMIR serious games*, *4*(2), e5888.

Lunt, L., Bramham, J., Morris, R. G., Bullock, P. R., Selway, R. P., Xenitidis, K., & David, A. S. (2012). Prefrontal cortex dysfunction and "Jumping to conclusions": Bias or deficit? *Journal of Neuropsychology*, *6*(1), 65–78. https://doi.org/10.1111/j.1748-6653.2011.02005.x

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology*. New York: Basic Books.

Manchester, D., Priestley, N., & Jackson, H. (2004). The assessment of executive functions: Coming out of the office. *Brain injury*, *18*(11), 1067–1081.

Marcelle, E. T., Ho, E. J., Kaplan, M. S., Adler, L. A., Castellanos, F. X., & Milham, M. P. (2018). Cogmed working memory training presents unique implementation challenges in adults with adhd. *Frontiers in psychiatry*, *9*, 390828.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer" evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*(4), 512–534.

Merholz, P. (2007). Peter in conversation with Don Norman about UX & innovation [Accessed: 2024-03-27]. *Adaptive Path*, *13*. https://web.archive.org/web/20131207190602/http://www.adaptivepath.com/ideas/e000862/

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Milyavskaya, M., Inzlicht, M., Hope, N., & Koestner, R. (2015). Saying "no" to temptation: Want-to motivation improves self-regulation by reducing temptation rather than by increasing self-control. *Journal of Personality and Social Psychology*, *109*(4), 677.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A., Shah, P., et al. (1999). *Models of working memory*. Citeseer.

Moran, K. (2019). Usability testing 101 [Accessed: 2023-08-20]. https://www.nngroup.com/articles/usability-testing-101/

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British journal of psychology*, *81*(2), 111–121.

Nacke, L. E., & Deterding, S. (2017). The maturing of gamification research.

Nielsen, J. (2000). Why you only need to test with 5 users [Accessed: 2023-05-26]. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Nielsen, J. (2012a). Thinking aloud: The #1 usability tool [Accessed: 2024-03-27]. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

Nielsen, J. (2012b). Usability 101: Introduction to usability [Accessed: 2023-05-26]. https://www.nngroup.com/articles/usability-101-introduction-to-usability/

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 206–213.

Ninaus, M., Pereira, G., Stefitz, R., Prada, R., Paiva, A., Neuper, C., & Wood, G. (2015). Game elements improve performance in a working memory training task. *International journal of serious games*, *2*(1), 3–16.

Nobre, A. d. P., Rodrigues, J. d. C., Sbicigo, J. B., Piccolo, L. d. R., Zortea, M., Junior, S. D., & de Salles, J. F. (2013). Tasks for assessment of the episodic buffer: A systematic review. *Psychology & Neuroscience*, *6*(3), 331.

Norman, D. A. (2013). *The design of everyday things*. Basic Books.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In *Consciousness and self-regulation: Advances in research and theory volume 4* (pp. 1–18). Springer.

Nouchi, R., Taki, Y., Takeuchi, H., Hashizume, H., Akitsuki, Y., Shigemune, Y., Sekiguchi, A., Kotozaki, Y., Tsukiura, T., Yomogida, Y., et al. (2012). Brain training game improves executive functions and processing speed in the elderly: A randomized controlled trial. *PloS one*, *7*(1), e29676.

Nouchi, R., Taki, Y., Takeuchi, H., Hashizume, H., Nozawa, T., Kambara, T., Sekiguchi, A., Miyauchi, C. M., Kotozaki, Y., Nouchi, H., et al. (2013). Brain training game

boosts executive functions, working memory and processing speed in the young adults: A randomized controlled trial. *PloS one*, *8*(2), e55518.

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R. J., & Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*(7299), 775–778.

Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *37*(1), 51–87. https://doi.org/10.1111/j.1469-7610.1996.tb01380.x

Plotly Technologies Inc. (2015). *Collaborative data science*. https://plot.ly

Prins, P. J., Dovis, S., Ponsioen, A., Ten Brink, E., & Van Der Oord, S. (2011). Does computerized working memory training with game elements enhance motivation and training efficacy in children with adhd? *Cyberpsychology, behavior, and social networking*, *14*(3), 115–122.

Procci, K., Chao, A., Bohnsack, J., Olsen, T., & Bowers, C. (2012). Usability in serious games: A model for small development teams. *Computer Technology and Application*, *3*(4).

PsyToolkit. (2021a). *Backward corsi task* [Accessed: 2024-02-15]. https://www.psytoolkit.org/experiment-library/backward_corsi.html

PsyToolkit. (2021b). *Multitasking* [Accessed: 2024-02-15]. https://www.psytoolkit.org/experiment-library/multitasking.html

PsyToolkit. (2021c). *Task switching* [Accessed: 2024-02-15]. https://www.psytoolkit.org/experiment-library/taskswitching_cued.html

PsyToolkit. (2022a). *Digit span task* [Accessed: 2024-02-15]. https://www.psytoolkit.org/experiment-library/digitspan.html

PsyToolkit. (2022b). *Task switching* [Accessed: 2024-02-15]. https://www.psytoolkit.org/experiment-library/taskswitching.html

Rabbitt, P. (2004). *Methodology of frontal and executive function*. Psychology Press.

Raiford, S. E., Coalson, D. L., Saklofske, D. H., & Weiss, L. G. (2010). Chapter 2 - practical issues in wais-iv administration and scoring. In L. G. Weiss, D. H. Saklofske, D. L. Coalson, & S. E. Raiford (Eds.), *Wais-iv clinical use and interpretation* (pp. 25–59). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-375035-8.10002-3

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., & Acker, J. D. (2005). Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. *Cerebral cortex*, *15*(11), 1676–1689.

Reynolds, C. R., & MacNeill Horton Jr, A. (2008). Assessing executive functions: A life-span perspective. *Psychology in the Schools*, *45*(9), 875–892.

Rock, P. L., Roiser, J. P., Riedel, W. J., & Blackwell, A. (2014). Cognitive impairment in depression: A systematic review and meta-analysis. *Psychological medicine*, *44*(10), 2029–2040.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of experimental psychology: General*, *124*(2), 207.

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2012). Implications of infant cognition for executive functions at age 11. *Psychological science*, *23*(11), 1345–1355.

Sala, G., Aksayli, N. D., Tatlidil, K. S., Gondo, Y., & Gobet, F. (2019). Working memory training does not enhance older adults' cognitive skills: A comprehensive meta-analysis. *Intelligence*, *77*, 101386.

Sala, G., & Gobet, F. (2020). Working memory training in typically developing children: A multilevel meta-analysis. *Psychonomic bulletin & review*, *27*(3), 423–434.

Saylik, R., Williams, A. L., Murphy, R. A., & Szameitat, A. J. (2022). Characterising the unity and diversity of executive functions in a within-subject fmri study. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-11433-z

Sbordone, R. J. (1996). Ecological validity: Some critical issues for the neuropsychologist.

Scharinger, C., Prislan, L., Bernecker, K., & Ninaus, M. (2023). Gamification of an n-back working memory task–is it worth the effort? an eeg and eye-tracking study. *Biological Psychology*, *179*, 108545.

Shah, T. M., Weinborn, M., Verdile, G., Sohrabi, H. R., & Martins, R. N. (2017). Enhancing cognitive functioning in healthly older adults: A systematic review of the clinical significance of commercially available computerized cognitive training in preventing cognitive decline. *Neuropsychology review*, *27*, 62–80.

Shahmoradi, L., Mohammadian, F., Rahmani Katigari, M., et al. (2022). A systematic review on serious games in attention rehabilitation and their effects. *Behavioural neurology*, *2022*.

Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain*, *114*(2), 727–741. https://doi.org/10.1093/brain/114.2.727

Shiffrin, R. M. (1976). Capacity limitations in information processing, attention, and memory. *Handbook of learning and cognitive processes*, *4*, 177–236.

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in psychology*, *6*, 328.

Soegaard, M. (2019). Usability: A part of the user experience [Accessed: 2024-03-26]. https://www.interaction-design.org/literature/article/usability-a-part-of-the-user-experience

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs: General and applied*, *74*(11), 1.

Stern, C., & Hasselmo, M. (2009). Recognition memory. In L. R. Squire (Ed.), *Encyclopedia of neuroscience* (pp. 49–54). Academic Press. https://doi.org/https://doi.org/10.1016/B978-008045046-9.00784-1

Stoet, G. (2010). Psytoolkit: A software package for programming psychological experiments using linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/brm.42.4.1096

Stoet, G. (2017). Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Stoet, G., O'Connor, D. B., Conner, M., & Laws, K. R. (2013). Are women better than men at multi-tasking? *BMC Psychology*, *1*(1). https://doi.org/10.1186/2050-7283-1-18

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Stuss, D. T. (1992). Biological and psychological development of executive functions. *Brain and Cognition*, *20*(1), 8–23. https://doi.org/10.1016/0278-2626(92)90059-u

Toril, P., Reales, J. M., & Ballesteros, S. (2014). Video game training enhances cognition of older adults: A meta-analytic study. *Psychology and aging*, *29*(3), 706.

Vaughan, L., & Giovanello, K. (2010). Executive function in daily life: Age-related influences of executive processes on instrumental activities of daily living. *Psychology and aging*, *25*(2), 343.

Verghese, J., LeValley, A., Derby, C., Kuslansky, G., Katz, M., Hall, C., Buschke, H., & Lipton, R. B. (2006). Leisure activities and the risk of amnestic mild cognitive impairment in the elderly. *Neurology*, *66*(6), 821–827.

Vermeir, J. F., White, M. J., Johnson, D., Crombez, G., & Van Ryckeghem, D. M. (2020). The effects of gamification on computerized cognitive training: Systematic review and meta-analysis. *JMIR serious games*, *8*(3), e18644.

Wei, J., Hou, J., Mu, T., Sun, J., Li, S., Wu, H., Su, B., & Zhang, T. (2022). Evaluation of computerized cognitive training and cognitive and daily function in patients living with hiv: A meta-analysis. *JAMA network open*, *5*(3), e220970–e220970.

West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological bulletin*, *120*(2), 272.

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: A meta-analytic review. *Biological psychiatry*, *57*(11), 1336–1346.

Willcutt, E. G., Pennington, B. F., Boada, R., Ogline, J. S., Tunick, R. A., Chhabildas, N. A., & Olson, R. K. (2001). A comparison of the cognitive deficits in reading disability and attention-deficit/hyperactivity disorder. *Journal of abnormal psychology*, *110*(1), 157.

Zabelina, D. L., Friedman, N. P., & Andrews-Hanna, J. (2019). Unity and diversity of executive functions in creativity. *Consciousness and Cognition*, *68*, 47–56. https://doi.org/10.1016/j.concog.2018.12.005

Appendix A

Procedure

GROUP A

Please follow the steps below carefully, one after another. Tick a box when the step is completed.

- ☐ Have your research code from university nearby - you will need it on several occasions.
- ☐ Some tasks may be easier to interact with using a mouse rather than a touchpad. If you have the option to use the mouse, we would recommend you use it.
- ☐ First questionnaire - https://forms.gle/DNYUUD5w5S5ZgRo77.
    - As part of the questionnaire you will find the informed consent. Make sure everything is clear to you and if you have ANY questions, please ask Prof. Podlesek (anja.podlesek@ff.uni-lj.si). If you agree to take part in this study voluntarily and have no further questions, agree by checking the box.
    - Fill in the first questionnaire asking about your research code, basic demographic information, and your experience with gaming and waitressing.
- ☐ You will now start with the part of the experiment that will run in **PsyToolKit.** In the beginning, there is an informed consent form you need to agree to. Then you will be asked to enter your research code - do so. After that you will be presented with three tasks. Please, complete all the tasks. The instructions will be part of the tasks themselves. However, if you have any questions about the tasks, do not hesitate to ask Prof. Podlesek (anja.podlesek@ff.uni-lj.si). Please start with Psytoolkit tasks now: https://www.psytoolkit.org/c/3.4.6/survey?s=3xkKx
- ☐ You will now start with a newly developed **Restaurant Game**. If you have Windows, please follow the instructions in the left column. If you have MacOS, please follow the instructions in the right column.

For WINDOWS:

- ☐ Please download the game from this link.
- ☐ **Unzip the file** by right-mouse-clicking on the zip folder and choosing the option to "Extract the file". Extract the file in a new folder. → This step is very important. If you do not extract the files, the game may not work properly or your data may not be recorded.
- ☐ Now open the extracted folder.
- ☐ Search for the file named "*WorkingMemory-Restaurant*" and double-click on it.
    - The file is an executable application. It may ask you if you are sure you want to open the file as it is from an unknown author. Please, do so.
    - Please remember - DO NOT click on the "*WorkingMemory-Restaurant*" file in the zip folder or from a web browser. In case you do this, no data will later be recorded. Be sure you run the game from the extracted folder.
- ☐ Play the game. During the game, please use the full screen mode.

For MacOS:

- ☐ Please download the game from this link.
- ☐ **Extract the files from the folder.**
- ☐ Open the folder that was extracted and find a file named *play_mac*. Double-click on it.
- ☐ If the game did not start, please do the following:
    - Go to your *Settings* and choose *Privacy&Security*. When you scroll to Security part, you should see a notification with the possibility to allow starting the game file.
    - It may happen that after entering your research code into the game, you will get a prompt asking if you want to allow the game/application to access your folder. It is because the game records important play logs (e.g., final results). **Please, give the game this access.**
    - Remember - DO NOT start the game by opening it in the web browser. In case you do this, no data will later be recorded. Be sure you run the game from the extracted folder.
- ☐ Play the game. During the game, please use the full screen mode.

- [ ] After playing the game, you will find a new text file called “YourResearchCode-Logs” in the folder where you opened the game. For example, for research code *RestGa12,* the file name is *RestGa12-Logs.*
- [ ] You will now upload this text file via Dropbox. As a name please enter JUST your research code. DO NOT ADD YOUR NAME. Additionally, the website will ask you to enter an email address - do not worry we will not be able to see it. Dropbox will use it just to send you a confirmation of successful upload. If you do not feel comfortable providing your email address, you may write down this one: dlugosova24@uniba.sk.
- [ ] Fill in the final questionnaire asking about your experience.
- [ ] If all boxes have been ticked, you have finished the experiment. Thank you for your time!
- [ ] If you want to receive research credits for your participation in this research, please fill in the form here.

Appendix B

Generator of orders in Restaurant Game

We take the list of all items and shuffle it to get a random permutation. We then take the subset of that list consisting of the first n items, where n is the current number of ordered items. We set the initial position of the first food item and the last drink item to be $n + 1$ and -1, respectively. After that, we go through the taken subset. When we come across the food, we save its position if it is smaller than the position that is currently saved. When we encounter a drink, we save its position if it is bigger than the currently saved position. Finally, we check if the position of the first food is smaller than the position of the first drink.

This comparison will be false if there is no drink or food item. The same goes if all drinks are before all foods, in which case the participant would not have to manipulate the remembered items as they would be in the correct order already. We repeat the generating process one hundred times or until we generate an order fulfilling the requirements. If no such order is generated, we generate a new random order, put a random food as the first item, and a random drink as the middle item. This enforcement is not likely to happen but ensures that every order shown to the player requires manipulation.