MENTAL CHRONOMETRY IN THE VIRTUAL WORLD:

HOW HAND MOVEMENTS MAY PAINT THE FUTURE OF EXPERIMENTAL PSYCHOLOGY.

Vom Fachbereich Sozialwissenschaften
der Rheinland-Pfälzischen Technischen Universität, Campus Kaiserslautern
zur Verleihung des akademischen Grades
Doktor der Philosophie (Dr. phil.)
genehmigte

D i s s e r t a t i o n

vorgelegt von
Omar Jubran

Tag der Disputation:	Kaiserslautern, 14.05.2024
Dekan:	Prof. Dr. Michael Fröhlich
Vorsitzende/r:	Prof. Dr. Thomas Schmidt
Gutachter/in:	1. Prof. Dr. Thomas Lachmann
	2. Prof. Dr. Cees van Leeuwen

DE 386

Juni 2024

Dedication

This work would not have been possible without the help of many people, to whom I dedicate it in no particular order.

To my cherished family, my dearest mother, Wisam, and my beloved father, Fahmi, whose boundless love and sacrifices propelled me to where I am right now. To my three brothers Abed-Alrahman, Ahmad, and Hamza, and my little sister Rawan. I left my home ten years ago, yet I never felt away. In the loneliness of life on the other side of the sea, you were my comfort.

To professor Thomas Lachmann, my supervisor, whose guidance, support and encouragement helped make my own path within science. Your belief in my abilities, fueled my determination, and for that I am forever grateful. Your work ethic will continue to inspired me wherever I go, and I will look back at my time in Kaiserslautern with much pride to have worked with you. In addition, I express my gratitude to professor Thomas Schmidt and professor Cees van Leeuven that provided guidance throughout the path. Your expertise, time and insight shaped my academic pursuit.

To Maximilian Wolkersdorfer. The friend that stood by me during good and bad times. Our late night walks, our long discussions, and your advice I hold deer to my heart. I survived this journey in part thanks to you.

To my friends Laís Muntini and John Gamboa. You were by side throughout this whole journey, writing manuscripts and raging against the machine. Our late night chats, and intense discussions about everything life kept me sane throughout these years.

To the true mensa warriors, Felix, Chris, Nils and Fenia who made every day at the TU Kaiserslautern, and later at the RPTU much brighter. Our weird lunchtime discussions have scared generations at the university and I wouldn't have had it any other way.

Finally, to the extraterrestrial being that came into my life in the midst of the pandemic, the person that changed the course of my life, my wife Boran. While my journey with the PhD is coming to a close, my journey with you is just beginning. You have given me hope, love and support. I know that no matter where I go, you will be there by my side, and I will be by yours. I am forever grateful.

Preface

In his book, “The Structure of Scientific Revolutions” Kuhn outlines the hallmark of scientific achievement called paradigms (Kuhn, 1970). Paradigms have two essential characteristics: they are unprecedented and leave open problems ripe for scientific exploration. Paradigms for Kuhn are necessary for scientific advancement, singular, grand, and groundbreaking.

Experimental psychology may have shifted with the introduction of the subtraction method by Donders and the additive factors by Sternberg (Donders, 1969; Sternberg, 1969). Another such leap happened with the introduction of brain imaging and physiological data (D. E. Meyer et al., 1988). These achievements can indeed be paradigm shifts described by Kuhn, but one could argue that the field is still in the pre-paradigmatic phase.

In this work, my initial goal was the search for a new paradigm! A shift in science, one that is grand. I did realize, however, that perhaps that is not something I am capable of. Nor is it what the field needs. Experimental psychology is relatively young and doess not necessarily need a new paradigm but rather more tools and techniques to operationalize its current theories and concepts.

Thus, this work instead explores Virtual Reality and its application in psychological experiments. It also explores the utility of hand movements and distributional analysis, culminating in the introduction of Spatiotemporal Survival Analysis (StSA), hoping that these methods advance the status quo of science.

Table of Contents

List of tables

List of figures

Notes

Chapter 3 was published as an ACM proceeding for ECCE 2022.

Chapter 4 was submitted for publication for behavior research methods.

First part of chapter 5 was submitted for publication as an ACM proceeding and full publication for ECCE 2024.

Design and data collection for the second part of chapter 5 has been done within the scope of a master's thesis I supervised by Cora Trefney-Rankin.

The formatting for chapters 3-5 has been adjusted to fit within the current thesis, including figure numbers and reference style.

Chapter 1: Introduction

Can we measure cognitive processing in ecologically valid settings?

Classical experimental psychology has long relied on controlled laboratory settings to quantify mental operations. These settings traditionally are extremely minimal, eliminating any source of noise or external effects. Experiments are performed on computer screens and responses are collected through a keyboard, using mental chronometry as its primary tool. Mental chronometry relies on Error Rates (ER) as well as Reaction Times (RT; also known as Response Times). RT have emerged as the principal tool through which insights into cognitive functions is gained. The origin of this notion can be traced back to the experiments conducted by Franciscus Donders in 1868 (Donders, 1969).

However, classical experiments in psychology fall short of mimicking real-life scenarios. Real life scenarios and mental processes are multifaceted, complex, and rarely occur in isolation. Some experiments try to remedy this issue by conducting studies outside of the laboratory. These, however, fall short of the control of the lab. In addition, replicating psychological effects in general is compromised by a variety of issues in experimental control, data collection, and analysis.

Virtual reality (VR) technology presents fresh opportunities for executing psychological experiments. This technology offers experimentally controlled environments, as well as continuously tracked responses using motion tracking controllers. VR combines immersive visual presentation with advanced techniques for continuous monitoring of responses by tracking of hand movement trajectories. VR also provides standardized equipment, suitable for between-lab replications. This ensures consistency and comparability across different research settings.

The primary focus of this work was on exploring the utility of VR and continuous response tracking in psychological experiments. Continuous tracking elucidates the fine-grained dynamics of decision-making. Distributional methods, such as Survival analysis (SA) and my newly developed method, Spatiotemporal Survival Analysis (StSA) served as the primary tool to analyze these responses. Studying the time-course of behavior in classical paradigms can deepen our insight into the cognitive processes, such as working memory and response conflict. This methodology not only sheds new light on classical paradigms but also establishes a groundwork for future research aiming to unravel the complexities of cognitive processes in experimental contexts.

This work also sets out to demonstrate how it is possible to preserve the legacy of prior well established experimental paradigms in VR. In a series of experiments, VR was used to replicate findings in tasks related to visual perception (chapter 4) and working memory (chapters 5 and 6). These experiments demonstrate that, not only can classical results be replicated in VR, but that this technology enables a more fine-grained understanding of human behavior, relying on continuous tracking of response behavior (demonstrated by chapter 5). Finally, this work concludes with remarks on the future of experiments in VR and the utility of Survival Analysis using continuous movement trajectory data.

Chapter 2: Theoretical background

2.1 Mental Chronometry

Mental chronometry aims to assess the temporal dynamics of response during a cognitive task (Posner, 2005). Understanding the temporal dynamics of mental operations within the human brain has been a driving force for researchers since the mid-19th century (D. E. Meyer et al., 1988) Central to this understanding, is the study of response times (RT), also known as reaction times (Donders, 1969).

Donders achievement was marked by the introduction of what is known as the subtraction method (Donders, 1969). In his experiments, he compared participant's reaction times to a simple stimulus (i.e. a recognition task), with reactions to a choice task, where participants responded to two different light stimuli, each by pressing the appropriate button. He found that participants were slower in the choice task and attributed this difference to the decision-making process; given that the perception and response processes were the same in both tasks. This is what he called pure insertion, which assumes that each cognitive component produces an activation that is the same irrespective of context (Friston et al., 1996).

Donders employed this method to discern the difference between simple reactions and more complex decision-making processes, which laid the foundation for subsequent developments in the study of mental chronometry. While Donders's work is certainly a cornerstone of mental chronometry, some scholars contend that Hermann von Helmholtz's investigations into simple reaction-time procedures in frogs could have been a precursor to Donders's endeavors. Helmholtz's groundbreaking work, carried out between 1850 and 1853, might have indirectly influenced Donders's experimental design and conceptual framework (Helmholtz, 1948).

Expanding on this lineage, Saul Sternberg introduced the additive factors method, a seminal framework that dissected mental operations into sequential stages. Sternberg postulated that RTs could be deconstructed according to these stages, each contributing to the overall response time (Sternberg, 1969). This proposal facilitated a more nuanced understanding of cognitive processing by allowing researchers to disentangle the various components contributing to the time it takes to respond to stimuli, where Analysis of Variance (ANOVA) is used to analyze the effects of experimental factors on the mean RT. In essence, Sternberg's additive factors method evolved Donders's initial concept, while relaxing his postulation of pure insertion.

2.2 Response Conflict Paradigms

Response conflict paradigms are the tasks of choice for studying the time-course of response activation and inhibition. For this work, I employed both response priming and Eriksen flanker tasks to investigate response activation in Virtual Reality. In response priming (Vorberg et al., 2003), a very short task-irrelevant prime (about 17 ms) is followed by a target that requires a speeded two-choice response. When the prime is mapped to the same response as the target (congruent trials), responses are accelerated, and error rates are reduced. When the prime is mapped to the opposite response (incongruent trials), it slows down responses and provokes errors. Response priming increases with prime-target SOAs of up to 100 ms for primes and flankers alike. In Flanker tasks (C. W. Eriksen & Schultz, 1979), the congruent or incongruent task-irrelevant item is placed adjacent to the stimulus, rather than preceding it like a prime.

In priming tasks, the time-course of the underlying response conflict is by now reasonably well understood. The prime generates a feedforward burst of response activation (T. Schmidt et al., 2011; Vorberg et al., 2003) that completely dominates the fastest responses, which invariably

follow the prime instead of the target (Panis et al., 2020; Panis & Schmidt, 2016). Because incongruent primes activate the wrong response, the target must overcome a response conflict that deepens as the SOA is prolonged. We have used a variety of methods to trace this response conflict, including the microanalysis of response time distributions and hazard rates(T. Schmidt et al., 2022; Wolkersdorfer et al., 2020), monitoring of pointing movements that initially detour towards the misleading prime (T. Schmidt, 2002; T. Schmidt et al., 2015), monitoring of response forces (F. Schmidt et al., 2014), lateralized readiness potentials (Vath & Schmidt, 2007) and mathematical modeling via accumulator models (T. Schmidt & Schmidt, 2018). A particularly fascinating aspect of response priming is that priming effects are independent of the degree and time-course of visual masking of the prime (especially metacontrast masking), making it ideally suited for the study of unconscious perception (Biafora & Schmidt, 2020; T. Schmidt & Biafora, 2024; T. Schmidt & Vorberg, 2006).

2.3 Virtual Reality: experimental control and trajectory tracking

Because human perception and behavior are affected by a wide range of elements, the matter of isolating variables has always been a concern of cognitive scientists seeking to make valid and controlled investigations. The representativeness of the results gathered through the confining conditions of the laboratory, their generalizability, and applicability are addressed as the 'real-world or the lab' dilemma (Holleman et al., 2020). Hence, adding naturalistic conditions to the experimental designs, with interacting features, from different perceptual modalities, sets a big challenge. Especially considering the precision and equivalency of rendered experimental conditions throughout the necessary repetitions and between the subjects. In this sense, VR comes as a viable tool for cognitive research as it seeks to simulate world events in a realistic and immersive manner (Abdillah et al., 2020) and to mimic the natural process of perspective projection

(Scarfe & Glennerster, 2019). By approaching ecological validity with experimental control, research using VR takes steps into overcoming the above-mentioned dilemma, offering the possibility to approximate the in-laboratory stimuli to its natural presentation, while maintaining the control of a scenario created by design.

VR techniques have rapidly advanced in the past few years. VR enables the presentation of realistically rendered 3-dimensional objects in controlled experimental setups (Hoffman, 1998; Wilson & Soranzo, 2015). Presenting realistically rendered objects enhances active manipulation tasks, while VR optimally supports continuous observation of behavior in space and time by tracking hand movement trajectories. In addition to trajectory tracking, VR also supports the continuous tracking of gaze data in 3D free viewing using eye trackers embedded in Head Mounted Displays (HMD). HMD and VR controllers provide synchronized complementary data that can be used for analyses in paradigms relying on both.

This work set out to develop methods of understanding and analyzing trajectory data, namely using the novel method Spatiotemporal Survival Analysis (StSA) (Jubran et al., Submitted). Before VR, trajectory tracking was mostly performed on pointing movements. For example, in a task where participants were asked to draw a line between two points on a paper, Woodworth demonstrated that movements at the beginning of task execution tended to be faster and less accurate, and slower and more accurate as the participants got closer to the endpoint (Woodworth, 1899). This provided information about the time required for movement planning and execution as well as speed-accuracy tradeoffs. Since then, many studies have used pencil & paper, digitizing tablets, stylus, and computer mouse tracking for trajectory recording (Elliott et al., 2001).

Mouse tracking is a powerful tool for investigating real-time cognitive processes such as prediction. Studies typically involve a binary forced-choice task, where participants respond to stimuli by deciding between two buttons on a screen (T. Meyer et al., 2023) . During the response process, the mouse trajectory is recorded, aggregated, and later analyzed as a time series (Schoemann et al., 2021).

However, mouse studies are in general restrained by design factors, i.e., response dictation, mouse sensitivity, and starting procedure (Elliott et al., 2001). In addition, the relation between the movements and their effects on the screen, for instance grasping an object, remains conventional and artificial, restricting their utility (Niu et al., 2022). Finally, mouse tracking has limits in combination with eye tracking, since it is difficult to make mouse movements without following the cursor on the screen, limiting the viability of combining mouse movements with eye tracking. There is little effort in combining eye-tracking and mouse movements despite its potential benefits.

When it comes to investigations combining eye and hand movements, Hayhoe and colleagues (2003) used a head-mounted eye-tracker to record participants' natural behavior, i.e. while making a sandwich (M. M. Hayhoe et al., 2003). They aimed to understand the temporal dynamics of natural vision. Examining the patterns of eye-hand movement, they observed that spatial memory representations are needed for task execution, as well as planning and coordination. Hence, hand movements can be made independently from eye movements based on scenic memory. Other task-oriented studies helped describe the role of eye movements in the execution of every day visually guided behaviors that involve body movement, unraveling that eyes are positioned where is best for the spatio-temporal demands of the task to be done, being closely related to the time course of the task). Fixations were characterized to be guided by highly task-specific information,

relying on computations, e.g., fixating ahead of the impact point of a bouncing ball (M. Hayhoe & Ballard, 2005).

2.4 Spatiotemporal Survival Analysis (STSA)

Distributional analyses of behavioral performance measures across trials have become more relevant in recent years(Allison, 1982; Proctor et al., 2011; Ridderinkhof, 2002; Wolkersdorfer et al., 2020). Survival Analysis (SA) is one of the most diagnostic approaches focused on analyzing time-to-event data [28]. In the first step, time is divided into discrete time bins t. The hazard function $h(t) = P(T=t|T\geq t)$ plots the conditional probability of response occurrence, under the condition of no previous responses, against the passage of discrete time. The survivor function $S(t) = 1\ P(T \leq t)$ provides the probability that a response does not occur before the end of bin t. In other words, the percentage of the total trials without a response by the end of that bin. Using linear interpolation, one can obtain the estimated median RT, which is equivalent to S(t).50. In the last step, the conditional accuracy function ca(t) = P(correct=t) plots the conditional probability of a correct response in bin t against the passage of discrete time (Pachella, 2021; Wickelgren, 1977), for a detailed overview, see (Panis et al., 2020).

Spatiotemporal Survival Analysis (StSA) (Jubran et al., Submitted) is a novel method that aims to analyze time to events in continuous movement trajectories similar to SA. In STSA, the analysis pipeline consists of multiple steps. In the first, preliminary step, average movement trajectories for each condition are calculated. In the second step, single trials for all participants in each condition, without any censoring of the data are visualized. In the third step, the raw data is visualized by dividing time and space into discrete spatiotemporal bins, where the number of trajectories visiting each bin is summed up. These numbers are presented in a high-resolution heat

map of responses, showing the frequency distribution of trajectories visiting each spatiotemporal bin.

In the fourth and final step of the analysis, STSA is conducted using predefined spatiotemporal bins. An event in this case is the first instance where a trajectory's absolute value surpasses a spatial threshold, subsequently using its direction to determine accuracy. SA is then conducted by going over each spatiotemporal bin, updating the risk set with every bin first in space and then in time. The results are again presented in heat maps, where higher hazard rates show a higher probability that a trajectory will visit a certain spatiotemporal bin. These heat maps are then scaled according to the maximum hazard across all conditions. Similarly, conditional accuracy is calculated for each spatiotemporal bin and represented in heat maps as well.

Chapter 3: Research tools and methods

3.1 Behavioral measures

Throughout this work, different discrete (Error rates and Reaction Times) and continuous measures (Hand movements) of behavior were employed. Hand movements were calculated and presented as movement trajectories that start from a neutral position, between both response choices, and end by favoring one of them. Trajectories represent the difference of hand position between correct and incorrect choice as shown in Figure 1.

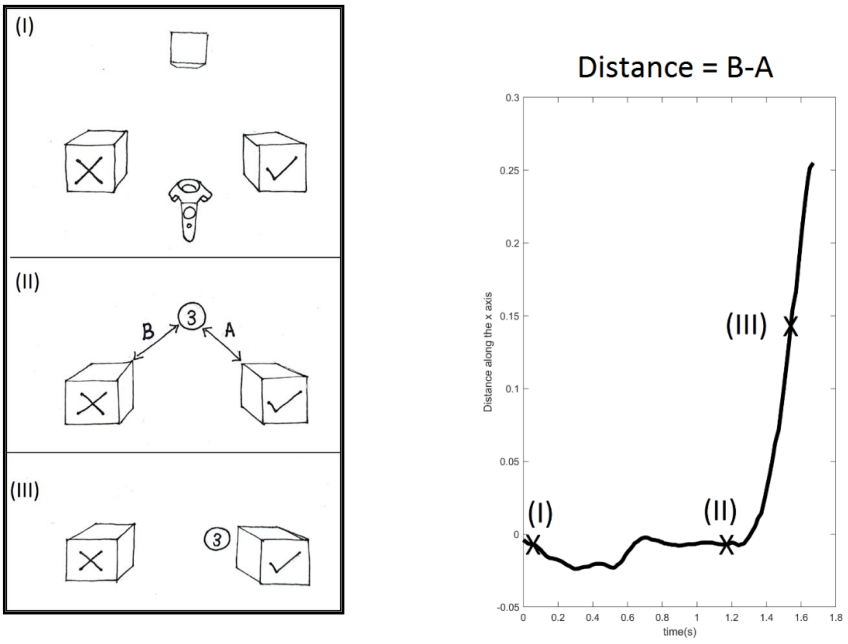


Figure 1. Trajectory calculation: The check-marked box represents the object, and the x-marked box represents the distractor. The movement trajectory is then calculated as the difference between the distance to the distractor (B) minus the distance towards the target object (A), resulting in a positive trajectory if the participant is getting closer to the target object, and a negative trajectory if the participant is getting closer to the distractor.

Given that both choices are on the same axis, this translates to displacement along the x-axis. Thus, moving the hand closer to the object will result in positive trajectories, while moving the hand closer to the distractor will result in negative ones.

An advantage of continuous response tracking is the ability to plot individual trials over time. To analyze these hand movement trajectories, the steps outlined by the Spatiotemporal Survival Analysis (StSA) pipeline (Jubran et al., submitted) were followed: First, hand movement trajectories were averaged, as seen in Figure 2. This allows for a quick overview of the data and is analogous to descriptive statistics. The second step in the analysis is plotting individual trials for all participants and all conditions, as shown in Figure 3.

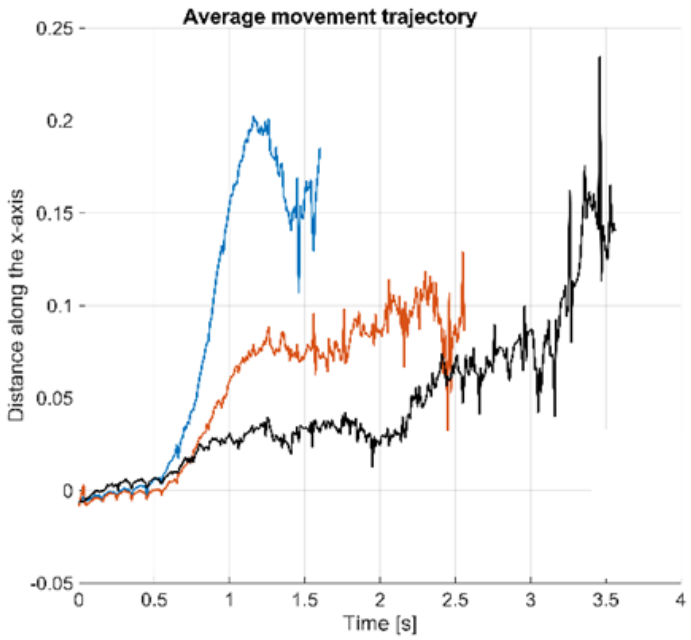


Figure 2. Averaged trajectories.

In the third step, individual trials are presented as log-transformed heat maps for better visual clarity of the response distributions as shown in Figure 4. Heat maps consider individual differences as opposed to averaged trajectories. Individual participant data can also be plotted as heat maps.

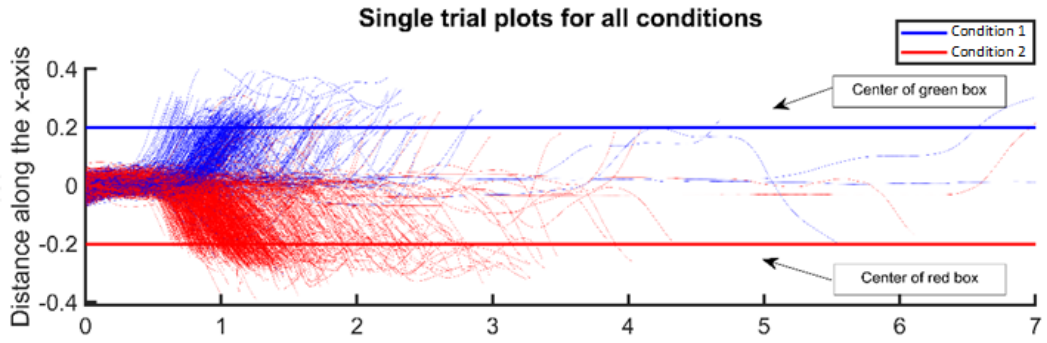


Figure 3. Individual trials for all participants and all conditions.

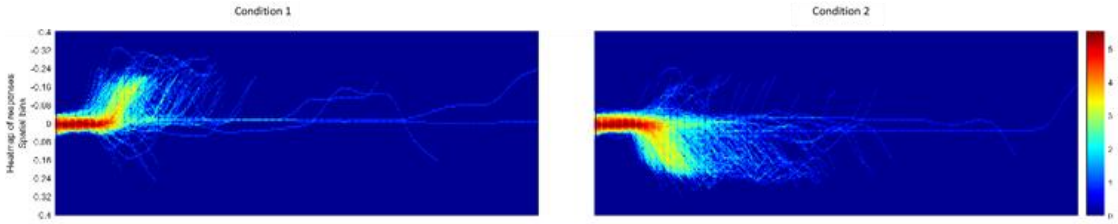


Figure 4. Heat maps that show data from all participants for both conditions.

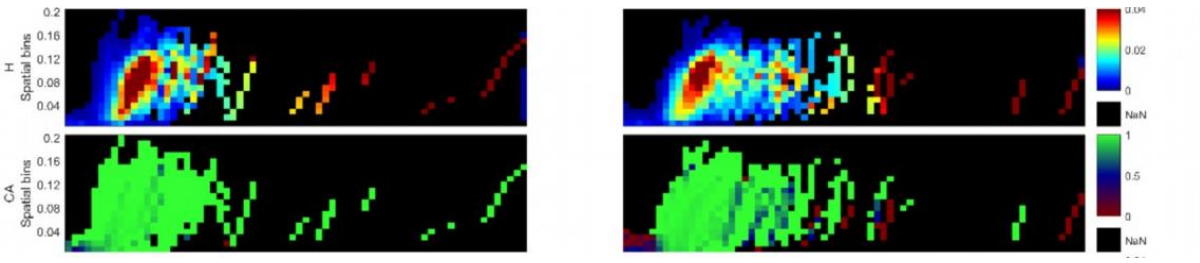


Figure 5. Hazard rate plots(H) and conditional accuracy (CA) plots for all participants in all conditions and trials

The final step of StSA is performed as follows: Each trial is divided into 20 measurements; each measurement representing the first time a trajectory crosses a spatial threshold in either direction

(each with 5% of total distance, incrementally increasing from 0). The direction of the crossing (accuracy) and time are calculated at each measurement and stored in a spatiotemporal bin, marked by the combination of discrete time recorded and crossed spatial threshold. These measures are then used to calculate hazard rates (H) and conditional accuracy (CA) values in each spatiotemporal bin shown as pixels in the graph. CA, in this case, considers the average accuracy of all trials in a certain spatiotemporal bin. H is defined as the conditional probability that a response occurs within a spatiotemporal bin given that it did not occur before; these are shown in Figure 5. Trajectory analysis (StSA), including the visualizations of H and CA in plots, could enable inferences about the fine-grained dynamics of individual differences in cognitive tasks that aggregated RT cannot show.

3.2 Statistical Analysis:

Studies conducted utilized within subject designs, emphasizing a small number of participants (Small-N designs). To achieve optimal power, the number of trials per participant and condition were maximized, and incorporated into the calculations of statistical power (Baker et al., 2021; Murphy et al., 2004; P. L. Smith & Little, 2018). Individual differences in the fine-grained dynamics of cognitive processing in studies were then analyzed using ANOVAs and Survival Analysis when applicable.

3.3 **Experimental power and measurement precision**

Apart from power, measurement precision at the level of individual participants in single sessions was controlled. Precision was calculated as s/Vr (Eisenhart, 1963), where s is a single participant's standard deviation in each cell of the design, and r is the number of repeated measures per cell and subject.

Chapter 4: Reproducing Classical Priming, Flanker, and Lexical Decision Tasks in VR

Between ecological validity and experimental control

Abstract:

Advances and easier accessibility of Virtual Reality (VR) technology are opening up new opportunities for researchers to conduct laboratory experiments under more ecologically valid yet controlled conditions. In the current study, we investigated the feasibility of utilizing VR in experimental psychology using three classical paradigms: a flanker task, a lexical decision task, and a response priming task with basic geometric shapes. To this end, we compared performance in desktop versions of these tasks with that of VR versions. Both versions were matched for visual angles of the stimuli and for the response device, which was a VR controller. For the lexical decision and the flanker tasks, results were in line with behavioral patterns reported in the literature, reproducing the classical effects i.e., faster responses for consistent trials compared to inconsistent ones in both versions of the task. However, we did not find the classical priming effects, possibly due to technical limitations causing a higher than expected SOA variance. Regarding the different versions, for none of the three tasks, did we find any difference in effects or interactions. These results show that VR is a viable medium for experimental research using the controllers, as long as visual angles and response devices are controlled. We present work in progress, with preliminary data for the lexical decision task for which data collection is not completed yet.

Introduction:

Virtual Reality (VR) technology opens new outlets for conducting psychological experiments involving visual input, leading to expedient methodological prospects. These include an increase in ecological validity while preserving control on experimental factors such as luminance, visual angles, and ergonomic validity of response devices. However, the employment of VR for psychological experiments also comes with certain constraints; such as conditions of stimulus presentation and response collection using wireless VR controllers. These are not necessarily disadvantages but they raise questions about the comparability of the results to those obtained using traditional methods in terms of validity and reliability of mental chronometry measures. We are testing the comparability between desktop monitor setups and VR across three classic and well-established experimental paradigms of experimental psychology; a flanker task (B. A. Eriksen & Eriksen, 1974), a visual response priming task (T. Schmidt & Vorberg, 2006), and a visual Lexical Decision Task (LDT; D. E. Meyer & Schvaneveldt, 1971). These three tasks are primarily visual perception tasks, with high sensitivity for visual noise on the one hand, while being easy to implement in VR on the other.

The Flanker task

Eriksen and Eriksen proposed a seminal paradigm to investigate visual selective attention based on response conflict (B. A. Eriksen & Eriksen, 1974). It allows measures of how task-irrelevant flanking stimuli influence the processing of a visual target to be taken. For that purpose, in each trial, a target stimulus is presented, surrounded by adjacent task-irrelevant flankers. These are either calling for the same response as the target, i.e. consistent flankers, or the opposite response, i.e. inconsistent

flankers. By varying the consistency between flankers and target, a typical pattern of results is reported in the literature: responses to consistent trials are faster and more accurate in comparison to inconsistent ones. Widely replicated for different kinds of material (Lachmann & Van Leeuwen, 2004; Stoffels & Van Der Molen, 1988; Van Leeuwen & Lachmann, 2004), this is referred to as the Flanker Effect.

Response Priming

Response priming is another well-established paradigm (Fehrer & Raab, 1962) that has been used to explore early visuomotor processing (Klotz & Neumann, 1999; Klotz & Wolff, 1995; Vorberg et al., 2003). Experiments in the field involve target identification tasks, in which participants have to respond as quickly and as accurately as possible to a target stimulus preceded by a (masked or unmasked) prime stimulus. Prime and target can either call for the same response, i.e. consistent primes, or the opposite response i.e. inconsistent primes. The typical pattern of results reported in the literature is that consistent trials show faster and more accurate responses when compared to inconsistent ones. The differences between consistent and inconsistent trials define the Priming Effect. This effect was shown to depend on stimulus-onset-asynchrony (SOA), i.e. the delay between the onset of the prime and the onset of the target, increasing for SOAs of up to about 100 ms (F. Schmidt et al., 2011; T. Schmidt, 2002; Vorberg et al., 2003) and decreasing for SOAs larger than this.

Lexical Decision Task

The third well-established paradigm (D. E. Meyer & Schvaneveldt, 1971) is the LDT. Word recognition research in the last 50 years has relied on the use of the LDT to evaluate the facility with which words are activated and retrieved during their recognition (Balota & Chumbley, 1984;

Wagenmakers et al., 2008) and how different variables can interact during word recognition processes (Larsen et al., 2006) by exploring differences on reaction times and accuracy. In the LDT a variety of letter strings are presented. These represent either words (e.g., SONG), or non-words (e.g., CNSO), and participants are instructed to classify them as such, as quickly and accurately as possible (Carney et al., 2015). Typically, word recognition times and accuracy rates are affected by the word likelihood of the non-words, in addition to a wide number of factors at multiple linguistic levels (Aguasvivas et al., 2020).

Methods

Participants

For the priming and flanker tasks; twenty university students (6 males, ages 20-30 years) participated in the study at the University of Kaiserslautern (TU Kaiserslautern), Germany. They all had normal or corrected-to-normal visual acuity and hearing. No speech, neurological, or reading impairments, or known pathologies. They all gave their written consent before participation. In the case of the LDT, we expect to record data from a similar group of participants.

Stimuli

For the flanker task, we relied on arrows as our primary experimental stimuli in compliance with (Stoffels & Van Der Molen, 1988). We only used two experimental conditions: consistent when flanking arrows and target pointed in the same direction, and inconsistent when flanking arrows and target pointed in opposite directions. In the case of the priming task, we opted for basic shapes; which were a square or a circle presented at three different SOAs (50,70,90 ms), SOAs were selected to be less than 100ms to observe the increasing priming effect as shown in Figure 6. In the case of the LDT, 100 high-frequency and highly accurate recognizable words and 100 highly accurate

recognizable non-words from the Spanish lexical decision mega-study (ESPALEX) will be chosen (Aguasvivas et al., 2020).

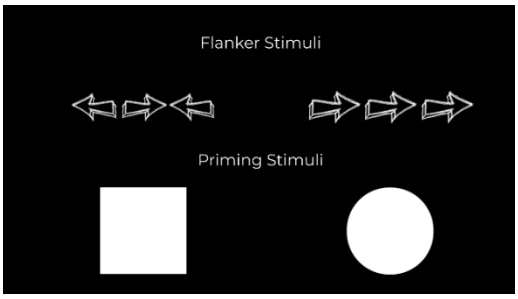


Figure 6. Stimuli used for different conditions in the Flanker Task (congruent on the right, incongruent on the left) and the Priming Task.

To ensure a naturalistic virtual experience within a real-world scenario, we chose a common and relatable setting: a classroom environment, where the stimuli were placed on a whiteboard. In this vein, targets were selected considering real classroom conditions, imitating chalk on a classroom board. However, to reduce disfluency within LDT stimuli, words will be presented in white ink avoiding chalk simulation.

Experimental Setup

This study compares the performance of each task in two different versions: VR, and PC, as part of an international collaboration between the University of Kaiserslautern in Germany and Nebrija University in Spain, the flanker task and the priming task were both conducted in Germany, while the design and data collection related to the LDT will be carried out in Spain.

The experimental tasks and surrounding environments are matched in both the PC and the VR versions; where stimuli are presented on a blackboard in a classroom environment (see figure 7). In the PC version the scene was statically presented on a monitor, while in VR, the participant was

immersed in the virtual environment and had the freedom to look around while remaining seated in their chair. The tasks were carried out in counterbalanced order and each started with a practice block of 20 trials. This was followed by 3 experimental blocks of 50 trials in the Flanker task, and 6 blocks of 50 trials in the Priming task, finally the LDT will have two blocks of 100 trials. This insured at least 50 trials per condition for each task, which combined with 20 participants should give the experiments enough statistical power according to (Baker et al., 2021).

Figure 7. Environment used for the Flanker and the Priming Tasks (left) and LDT (right).

The experimental procedure of each task is shown in Figure 8; in all three tasks, a fixation cross was presented for 1000ms at the beginning of each trial. This was followed by a target stimulus that was presented until a response was given. This target was preceded by a prime presented for 17ms, and a variable SOA in the Priming task.

In all tasks, participants were instructed to respond as quickly and accurately as possible using the trigger buttons on the controllers. Concerning the Flanker task, they had to respond with the hand corresponding to the direction of the arrow presented as the target stimulus, ignoring the flanking stimuli. As for the priming task, participants had to respond with their right hand if they saw a circle, and their left hand if they saw a square as the target. Finally, for the LDT, participants will be

instructed to answer with the right hand if they identified the string of letters as a word, or the left hand if they identified it as a non-word.

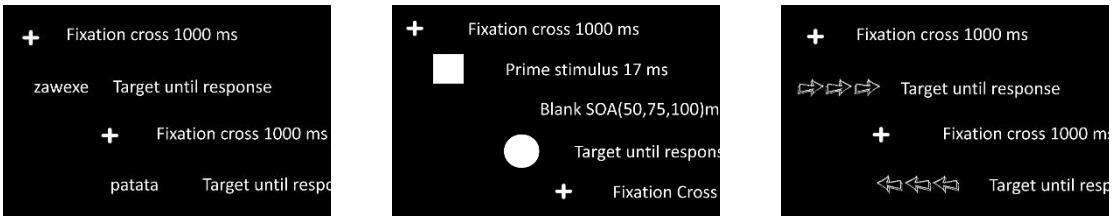


Figure 8. Experimental Procedure for all three tasks: flanker (right), priming (middle), LDT (left)

Apparatus

Tasks were executed on a high-end gaming laptop computer running Windows 10 Enterprise 2016 LTSB 64-Bit operating system with an Intel® Core i7-8750H processor. The stimuli in the PC version were presented on a monitor (40cm x 30cm) with a screen resolution of 1920 x 1200 and a refresh rate of 90 Hz, running g-sync technology to ensure a stable framerate. Participants were seated 70 cm away from the monitor. For the VR condition an HTC Vive VR headset with 1080x1200 per eye resolution, and a refresh rate of 60Hz was used. Both versions of the task used HTC Vive controllers. The experiment was programmed using Unity Software, version 2018.417f.

Visual Angle calculation

For both PC and VR versions, the visual angles were matched. To measure these angles in VR, we taped a transparent tape on the inside lenses of the VR headset, marked with lines in set distances, afterwards we observed the physical size of stimuli in the environment, seated at the appropriate distance; this, in combination with the distance between the lens and the eye was used to calculate the visual angle. This meant that the visual angle perdures as long as the participant is seated in the

same position. This angle was then replicated in the PC version using appropriate distancing and stimuli sizes.

Results:

Flanker task

Results for the flanker task are shown in figure 9. We carried out a 2 (Consistency: consistent and inconsistent) x 2 (Version: VR and PC) repeated-measures ANOVA for RTs in the flanker task; after removing trials longer than 1 second and inaccurate responses. The analysis showed a significant main effect for Consistency ($F(1,19)=336$, $p< .001$) where consistent trials were shorter than inconsistent trials for both VR and PC, and a significant main effect of Version ($F(1,19)=7.1$, $p=.015$) where VR RTs were shorter than PC.

Figure 9. Reaction times (left) and reaction time distributions(right) for the Flanker Task.

Priming Task

Results for the flanker task are shown in Figure 9. A 2 (Consistency: consistent and inconsistent) x 3 (Stimulus-onset-asynchrony: 50ms, 70ms, and 90ms) x 2 (Version: PC and VR) repeated-measures ANOVA analyses for RTs were performed after removing trials longer than 1 second, and inaccurate responses. Our main results show that RTs in the priming task showed a significant main effect for

Consistency (F(1,19) = 37.6, p < .001) where consistent trials were shorter than inconsistent trials for both VR and PC, and a significant main effect of SOA (F(2,19) =114.7, p=.015) where reaction times were shorter for longer SOAs. No significant effect was found for version (F(1, 19) = 0.30, p=.587). Finally, a significant interaction between Consistency and SOA was found (F(2,38) = 3.60, p=.001) which is typical of a priming task; longer SOAs had a decreased priming effect.

Figure 10. Priming Task RTs and reaction time distributions for both.

Discussion

Our goal is to test whether the findings of three classical experimental paradigms can be replicated in VR and thereby examine the applicability of VR as a valid setting in psychological research. In the present study, we tested this for three such classical paradigms: Flanker, Priming, and Lexical Decision. Data collection of the flanker task and the priming task are completed and the first analyses are presented here. The lexical decision task is in progress.

For the flanker and the priming task, we found, both for the classical desktop monitor and the VR version, the results typically observed in the literature for these paradigms, i.e., responses for consistent trials were faster and more accurate than for inconsistent ones with no significant interactions between version and consistency. In the case of the priming task, longer SOAs show faster reactions than shorter ones, with no interactions between version and SOA. Given that both

versions share the same experimental settings, i.e. same visual angles and response devices, our results suggest that the tasks function similarly in both versions, with VR having a speed advantage.

A typical finding that we were unable to replicate in our study is the increase in priming effect with increasing SOA (F. Schmidt et al., 2011). However, it must be taken into account that the SOAs in our study were up to 100ms; which may indicate that the actual SOAs were longer than intended. This can be due to the lack of synchrony between the PC and VR versions, or the high computational load associated with running the experiment simultaneously on PC and VR, causing stimuli presentation times to be less accurate. These effects are only present in the priming task due to its timing-sensitive nature, where stimulus presentation times and SOAs play a huge role in the observed effects. Nonetheless, the same patterns of behavior are observed in both the VR and PC versions, suggesting that these effects have less to do with the versions, and more to do with the setup.

Response distributions in both tasks show a similar pattern to that of reaction times, where it points out that VR controllers are a viable input device for such experiments. One thing to note is that due to the sampling frequency and the wireless nature of these controllers; the digitizing of some reaction times was favored in certain time bins in comparison to others.

In general, our study shows that it is indeed possible to replicate the typical results of classical paradigms under VR conditions. Thus, VR proves to be a useful technology for conducting psychological experiments in the laboratory with higher ecological validity, as long as viewing angles and response devices can be controlled. VR controllers as a response device can widen the possibilities in research. The constraints associated with VR as a method for conducting psychological research can even be turned into advantages in terms of flexibility and control. Yet,

this behavioral test is only a first step toward attempting the replication process of cognitive effects, opening the door for future research with more complex paradigms to explore this new field of more controlled settings.

Chapter 5: Spatiotemporal Survival Analysis: Elucidating Working Memory Estimates in Virtual Reality

Abstract

Behavioral experiments are typically limited to on-screen stimuli and key-press responses. Virtual Reality (VR) allows stimulus presentation in an immersive 3D environment and facilitates spatiotemporal trajectory tracking of response movements. We assessed these benefits by comparing a VR experiment to its classical counterpart. In an N-back task, working memory load effects on response times (RT) were reproduced in the Halfway and Arrival times of VR response trajectories. Yet, both the RT and VR data flout the assumptions of the customary ANOVAs, suggesting effects contingent on response dynamics. Overall consistency according to survival analysis cross-validates classical RT and VR results, whereas Halfway times resemble RT more than Arrival times. Survival analysis showed some surprising effects, namely a bias for sameness in early responses that is resolved in later ones. These observations led us to propose spatiotemporal survival analysis of response trajectories. It revealed transitions in the decision-making process, detailing not only when, but also where in the trajectories effects such as the early response bias arise. While classical and VR-based results are mutually consistent, spatiotemporal survival analysis of VR-based response tracking data enhances our understanding of response processes.

Introduction

In laboratory experiments on human perception and cognition, the basic principle of eliminating nuisance variables has become associated with minimalism in stimulus and response design. Stimuli are reduced to their bare essentials and responses to key presses, which are analyzed in terms of error rates (ER) and response times (RT). Experimental procedures have long been computer-based, using the monitor to display visual stimuli and the keyboard to collect responses. Performing repetitive tasks in such an impoverished environment is not something we do naturally. This raises the question of ecological validity, or whether classical laboratory results generalize to more realistic conditions. Field studies in which participants observe and actively manipulate three-dimensional objects could provide clues about this. However, such studies inevitably lack the rigor and control of the laboratory.

The emergence of virtual reality (VR) technology allows for the presentation of realistically rendered three-dimensional objects while preserving the benefits of controlled experimentation (Hoffman, 1998; Wilson & Soranzo, 2015). Unlike previous efforts to simulate 3D environments using devices such as 3D computer monitors, stereoscopes, and haploscopes, VR allows participants to navigate freely within the VR space (Banks et al., 2016). Cutting-edge head-mounted display technology optimizes the sense of presence, or 'being there', in virtual space (Scarfe & Glennerster, 2019). This is widely considered to be a critical factor for participants to respond naturally, mirroring their responses in the real world(Draper et al., 1999; Slater et al., 2009).

Moreover, VR technology enhances in particular the utility of active manipulation tasks. It provides an optimal platform for continuous observation of behavior in space and time through precise trajectory tracking. Trajectory tracking promises a wealth of information beyond a single response

to a stimulus, providing insight into decision-making, response preparation, and execution. Our current study aims to demonstrate how these advantages could benefit experimentation, by comparing classical and VR versions of the same task.

If trajectory tracking in VR is to move perception and cognition research beyond the status quo, an assessment of the consistency with classical results is crucial. We will conduct a behavioral experiment to compare performance in a classical laboratory setting on a personal computer (PC) with a VR version of the same task, namely the N-back task. Introduced in 1958 by Wayne Kirchner, this is a classical paradigm for estimating working-memory capacity (WMC; A. Baddeley, 1992; Conway et al., 2005; Kirchner, 1958; Wheeler & Treisman, 2002). Participants are presented with a series of items and are probed at random intervals to indicate by keypress whether the item currently on display matches the one presented N steps back. The value of N, that a participant still responds to correctly, is a direct estimate of WMC. We compare ER and RT from the classical version of the N-back task with analogical measures obtained from the movement trajectories in the VR-based version. On both, we perform distributional analyses (i.e. survival analysis) to gain insight into the fine-grained dynamics of response activation as well as the time course of cognitive processing. We observe consistency between both tasks. However, questions emerge that can only be answered by observing the VR trajectories. To elucidate the trajectory tracking results, we introduce the method of Spatiotemporal Survival Analysis (StSA). We claim that this method allows us to go beyond classical results in understanding decision processes.

Comparison of classical and VR-based experiments: the case of working memory

From a traditional perspective, the complex, dizzying, dynamic and gamified environments of VR might appear full of transient distractions that obfuscate measurement. For experimental

psychology to advance with the benefits of VR without sacrificing its legacy of prior achievements, it is crucial to establish whether legacy laboratory results hold up in VR environments. For optimal comparability of classical and VR-based experiments, it is prudent to proceed one step at a time, rather than straying away too quickly from classical paradigms. This means keeping randomized trials with distinct stimuli and responses and let participants perform a VR version of a classical task in the context of an enriched stimulus environment.

In a recent study, VR controllers were shown to be reliable devices for measuring response timing and accuracy in flanker and priming tasks (Jubran et al., Submitted). For the current N-back task, both serve as common dependent variables. As task difficulty increases (i.e., higher N), classically both RT and ER typically rise (for an overview see Meule, 2017). To examine the consistency between the classical N-back task and its VR counterpart, we utilized ER in both a classical and a VR version of the experiment, while for the latter we obtained analogues of RT from the captured three-dimensional response trajectories, namely the Halfway and Arrival times. The former is the time it takes to reach the midpoint of the response process; the latter is the time it takes to complete the response. While Arrival Times align naturally with RT, Halfway Times offer an arbitrary yet model-free insight into the ongoing response process. We initially compare results between classical and VR versions using traditional analyses of variance (ANOVA).

Distributional analyses of response latency measures across trials have become more relevant in recent years (Panis et al., 2020; Proctor et al., 2011; Ridderinkhof, 2002; Wolkersdorfer et al., 2020). Survival Analysis (SA) is one of the most diagnostic approaches (Panis et al., 2020) as well as a standard tool in many scientific disciplines when analyzing time-to-event data (Allison, 1982; Singer

& Willett, 2003). It provides insights into the temporal dynamics of response activation, hesitation, and indecision. We will apply SA to our RT, Halfway and Arrival times.

Insights from these analyses are still based on discrete measures of behavior and decision-making. In order to demonstrate the utility of response trajectory tracking in VR as a rich source of information about the processes leading to the response, we need to move beyond single events such as RT. Smith and colleagues (2022) used a button-release-and-press method. Following a target presentation, participants release a start button before pressing a response button (K. Smith et al., 2022). This allowed the authors to record two measures instead of one, besides the RT a response initiation time at which the start button is released. While this measure provides additional insight into the cognitive processes leading to the response, it still falls short of capturing the entirety of the response process, as enabled by trajectory tracking.

Trajectory tracking studies date back to before 1900 (Woodworth, 1899). The earliest studies used pen & paper, while from the 1970ies, digitizing tablets, stylus, and mouse tracking were utilized (for a review, see Elliott et al., 2001). These studies were mostly limited to pointing movements on two-dimensional surfaces (e.g., F. Schmidt & Schmidt, 2010; T. Schmidt, 2002). In one study, pointing movement trajectories have provided information about the time required for movement planning (Paulignan et al., 1991). Participants pointed in sequence to several visual targets, with the authors determining response initiation latency to be around 100 ms. Notably, when a new target appeared during the execution of a movement, participants' arrival at the target was delayed by a further 100-150 ms. This delay also manifested when changes to target size and position were introduced (Brenner & Smeets, 2005; Day & Lyon, 2000). The findings led to the conclusion that movement planning occurs within 100-150 ms from stimulus onset.

In addition to delays, bottom-up distractors in different locations may result in spatial deviations in trajectories. In Schmidt (2002), the novel item consisted of a distractor preceding the target stimulus. In Cressman et al., 2007 the distractor was presented after movement initiation. In both studies, early trajectory deviations towards the distractor occurred, which were later readjusted towards the target. The initial deviations in trajectory effects are akin to early errors in classical response conditions (Wolkersdorfer et al., 2020). Conversely, different effects could also be obtained from top-down processing. Rheem et al., 2018 evaluated deviations from an optimal trajectory as an indicator of WM load. They found that in higher load conditions, participants responded more slowly but, remarkably, with less deviation from the optimal route to the target. The result is akin to response inhibition as observed, e.g. in eye move-movement trajectories in Go/No-go tasks (Camalier et al., 2007). These contrasting effects illustrate that trajectory tracking can be diagnostic regarding the nature of the response mechanism.

Trajectory tracking in VR also bears relevance to an increasing concern in the field of experimental psychology: that of statistical power (Bishop, 2019). Statistical power refers to the ability of an experiment to detect an effect in a sample distribution (Baker et al., 2021). Well-known determinants are the number of participants as well as the number of trials per participant and condition. Incorporating the latter into the calculations of statistical power (Baker et al., 2021; Murphy et al., 2004) allows for small-n designs, i.e., designs with small number of participants (P. L. Smith & Little, 2018). Small-n designs enable data reproducibility at the level of individual participants, rather than through averaging across participants (P. L. Smith & Little, 2018). Another factor with the same effect is measurement precision within trials (Eisenhart, 1963). Movement trajectories potentially offer great precision, because the spatiotemporally resolved character of tracking data allows the extraction of multiple data points from each trial. Trajectory data can be

analyzed as series, transitions could be identified, and segments assigned to distinct stages, such as movement initiation, subsequent redirections, and completion of the trajectory (Elliott et al., 2001; F. Schmidt & Schmidt, 2010; Woodworth, 1899).

Despite these benefits of trajectory tracking, the relation between the movements and their effects on the screen, for instance grasping an object, has remained conventional and artificial, restricting their utility (Niu et al., 2022). Mouse tracking studies are also restrained by design factors, namely response dictation, mouse sensitivity, and starting procedure (for an in-depth review, refer to Schoemann et al., 2019). In contrast, by allowing response effects to appear naturally, VR opens for experimental psychology the perspective of ecological validity. Hence, response tracking in VR could advance our understanding of response processes beyond what classical methods contribute. The present study serves to establish this claim.

Methods

Participants

Twenty right-handed university students (6 males, ages 20-30 years, mean age of 26.3 years, SD = 2.5 years) from the German city of Kaiserslautern took part in the study. They all had normal or corrected-to-normal vision. According to self-reports, participants had no impairment of speech and reading as well as no diagnosis of psychological or neurological disorders. They all gave their written consent before participation. The experiment was approved by the ethics committee of the university following the ethical standards of the institution and with the 1964 Helsinki Declaration.

Stimuli

We used the display of Figure 11 (upper half). In the classical version, the display was a static image, lacking the yellow fixation box or the two response boxes. In the VR version, the display constituted

a dynamic, 3-dimensional environment. Displayed objects have realistically rendered surfaces, details, and cast shadows, enhancing realism and immersion in their virtual environment (Steinicke et al., 2005).

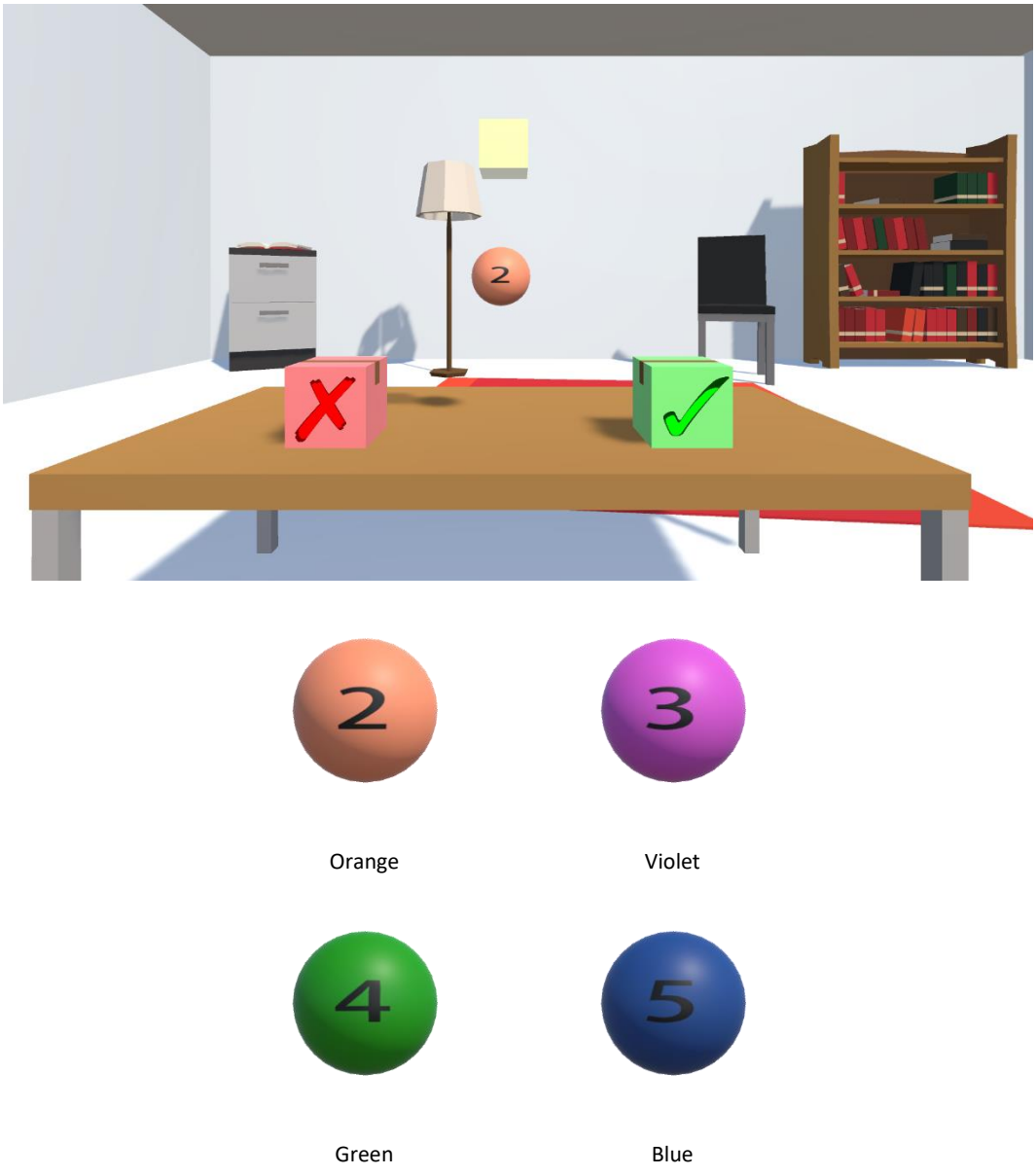


Figure 11. Visual display, showing one stimulus, fixation box, as well as the two response boxes (top half) and stimuli (bottom half) for both the classical and VR versions of the experiment. Note that in the VR version, the stimulus and the fixation box never appear simultaneously, and that in the classical version, the boxes do not appear at all.

Numbers ranging from 2-5 were the target stimuli in both the classical and the VR versions of the experiment. They appeared as labels of 4 different colored balls, Orange, Violet, Green, and Blue, as shown in the lower half of Figure 11. Colors and numbers were always matched, e.g., the blue ball always had the number 5. In both classical and VR versions, targets were placed on top of a virtual desk.

Apparatus

The study was conducted in a dimly lit experimental room. Tasks were executed on a high-end gaming laptop computer running a Windows 10 Enterprise 2016 LTSB 64-bit operating system with an Intel® Core i7-8750H processor. The stimuli in the classical version of the experiment were presented on a monitor (40cm x 30cm) of a PC with a screen resolution of 1920 x 1200 and a refresh rate of 90 Hz, running g-sync technology to ensure a stable frame rate. Participants were seated 70 cm away from the monitor. The VR version used an HTC Vive pro-VR headset with 1080x1200 per eye resolution and a refresh rate of 90Hz. The experiment was programmed using Unity Software, version 2020.1.17f1. In VR, hand movement trajectories were recorded throughout trials using VIVE Pro controllers.

Design and Procedure

Classical and VR versions consisted of equivalent sequences of events as follows: In each trial, a randomly chosen stimulus was presented and participants responded as fast and accurately as possible whether the current stimulus was a Match (identical) to the one presented N trials ago. Each trial had a 33% chance of being a Match. The value of N increased from 1-4 after every second block. Participants completed two sessions in counterbalanced order, one on a PC (the classical version) and one in VR. Each session lasted about an hour and had seven experimental blocks of 40

trials each. A session began with one practice block of 20 trials, also with a 33% chance of being a Match, where participants indicated whether the target stimulus was a green ball. Sessions were performed on different days, approximately 1 week apart on average.

In the classical version, a trial started with the appearance of the display with one of the target stimuli presented in the middle of the screen until response, participants pressed one of two response keys (letters F and J on a keyboard) using left or right index finger to indicate Match or non-Match, respectively. Key mapping was fixed for each participant and counterbalanced across participants. The stimulus disappeared after a key was pressed and the next trial started after a one second interval.

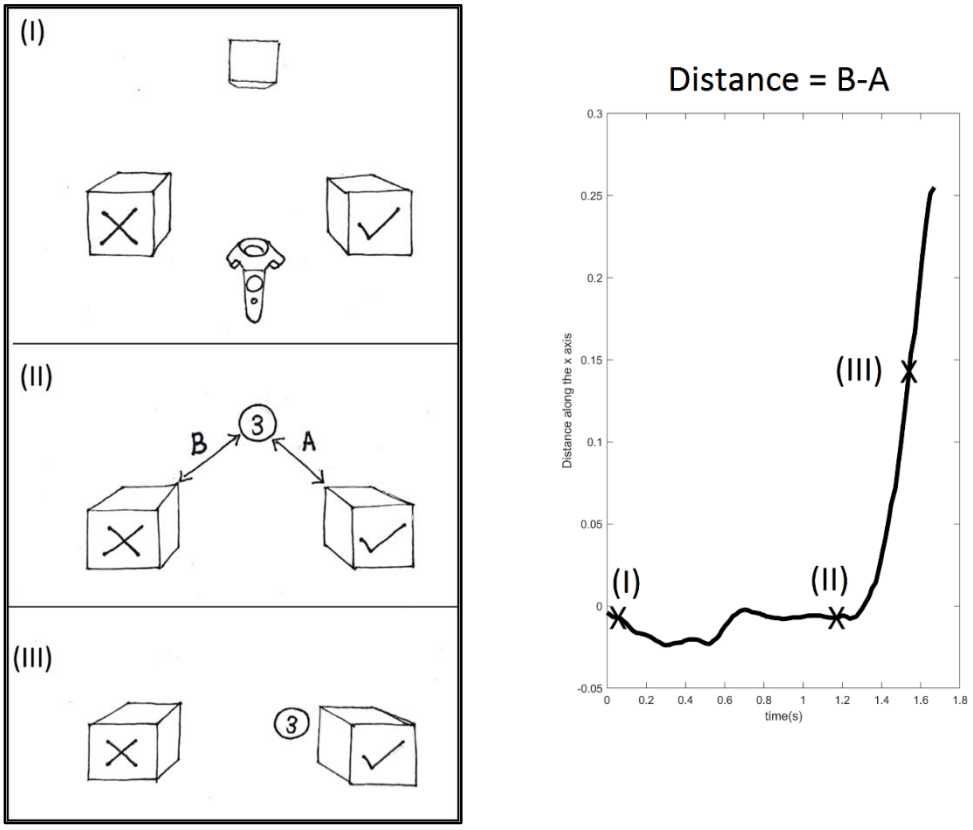


Figure 12. Response procedure in the VR version of the experiment (left). In stage (I), a transparent box appears, and the controller starts in a neutral position between the two response boxes, the checkmark indicates "Match", and the cross indicates "non-Match". In stage (II) the stimulus is captured; A and B represent the distances between the stimulus and the two boxes respectively. In (III) the stimulus is moved closer to one of the boxes. The stages (I, II, and III) of the response procedure depicted as a function of the stimulus position with respect to the response boxes (right). The movement trajectory is calculated by subtracting the distance between the two response boxes (B-A).

For the VR version, the procedure is shown in Figure 12. Participants started each trial with a transparent box, presented in the middle of the scene until the participant moved the controller inside the box. Next, the box disappeared and the target stimulus appeared halfway between the transparent box and the table (see the top half of Figure 11), and movement trajectory tracking started. They then moved the controller towards the stimulus and subsequently pressed and held

the trigger button to capture it. The controller became invisible once the stimulus was captured, and the participants had to hold the trigger button to maintain control of the stimulus. Using the now invisible controller they moved the stimulus in the three-dimensional virtual space to the green box if the target matched and to the red box otherwise. The target stimulus disappeared as soon as it collided with any of the boxes, and the controller reappeared in its stead. If the trigger button was released before collision the stimulus was kept in place within the environment and the controller reappeared. As a result, participants could move their hands to a resting position at any time during or after a trial. Positions of the boxes were counterbalanced across participants, to avoid any biases due to direction or the existence of shadows (left-right biases). There were no restrictions on the movements participants were allowed to make. The trial ended as soon as the ball reached the boundaries of either of the boxes. After a one second interval, the transparent box reappeared. The next trial started immediately after participants moved the controller into the transparent box again.

Analysis

Whereas all participants completed the task up to the three-back task (N3), 5 participants did not complete the four-back task (N4) in either the classical or the VR version, due to task difficulty or fatigue. This condition was therefore excluded from the analysis. Out of the 20 participants, one was also excluded from the analysis due to data corruption.

For the remaining 19 participants, RT were collected for the classical version and movement trajectories for the VR one. From the trajectories, ***Arrival times*** were calculated, which were the time starting from the moment the controller entered the fixation box until the stimulus collided with either of the boxes. Note that this interval includes the time needed for the participants to

capture the stimulus, as well as periods when participants release the stimulus before arrival, causing it to remain in the same location. We also calculated ***Halfway times***, which were recorded as the time once the stimulus was within a range of a spherical radius around the boxes equal to ½ of the distance between the two boxes. The selection of the halfway point was aimed to assess participants' performance at an intermediate stage. This offered a preliminary view of potential variations in participants' strategies and allowed exploration of any differences in their speed of approach.

Arrival and Halfway times each are associated with their own ER measure counterparts: Whereas Arrival ER inform us whether participants collided with the correct box, Halfway ER inform us whether participants at any point in time moved halfway towards it in the correct or incorrect direction.

Data analysis encompassed several stages. We started with traditional repeated-measures ANOVA, using RT and ER for the classical version, as well as Arrival times, Halfway times, and their ER counterparts for the VR version. All ANOVAs were performed twice; the first time without trimming outliers, revealing the full extent of inequalities in variance across conditions. This makes the point that, underlying these differences in variance, there might be a host of interesting processes waiting to be uncovered. The outcomes of this initial ANOVA are detailed in the supplementary materials.

Next, as is common for ANOVA, the data were trimmed for outliers using quartile outlier rejection (1.5 Interquartile range; David & Tukey, 1977) and further trimmed excluding values outside the range of the mean plus two times standard deviation. Trimming eliminates data deemed less representative of the underlying population. It reduces variance in the sample, in order to satisfy the criteria for planned analyses. However, it runs the risk of underestimating population means

and variances. Moreover, these procedures, we argue, may eliminate important aspects of the response dynamics and thus amount to throwing out the baby with the bath water. Despite the trimming procedure, we still encountered unequal variances, which violates the sphericity assumption. To address these violations in the way this is typically done, we applied the Greenhouse-Geisser correction (Greenhouse & Geisser, 1959) when necessary.

Subsequent analyses again were conducted without data trimming. To enhance our comprehension of the temporal characteristics of the data beyond the insights provided by mean RT analysis, we conducted SA. This involves investigating the time course of events, delineated by RTs and ERs for classical, alongside their respective counterparts in the VR version. In the first stage of SA, time is divided into discrete time bins t. The hazard function $h(t) = P(T=t|T\geq t)$ plots the conditional probability of response occurrence (condition: no previous responses) against the passage of discrete time. The survivor function $S(t) = 1\text{-}P(T \leq t)$ provides the probability that a response does not occur before the end of bin t. In other words, the percentage of the total trials without a response by the end of that bin. Using linear interpolation, median RT can be estimated, which is equivalent to $S(t)=.50$. In the last stage, the conditional accuracy function $ca(t) = P(correct=t)$ plots the conditional probability of a correct response in bin t against the passage of discrete time (Pachella, 2021; Wickelgren, 1977).

Lastly, we introduce our novel method, StSA. Leveraging continuous trajectories, we segment trials into many small steps, enabling the identification of multiple events throughout the trial, in contrast to relying on a single RT at the end. The StSA pipeline produces descriptive plots, trajectory distributions, as well as time to event data that contain details that are readily interpretable.

StSA consists of multiple stages. In the first, preliminary stage, average movement trajectories for each condition are calculated. In the second stage, single trials are visualized for all participants in each condition. Note that we perform no trimming of the data for the StSA. In the third stage, the raw data are visualized by dividing time and space into discrete spatiotemporal bins. Each bin counts the number of trajectories visiting each a given place in a given time. These counts are presented in a high-resolution heat map of responses, showing the frequency distribution of trajectories visiting each spatiotemporal bin.

In the fourth and final stage of the analysis, StSA is conducted using predefined spatiotemporal bins. In this stage, we identify events as the first instance where a trajectory's absolute value surpasses a spatial threshold, subsequently using its direction of travel (towards the left or right box) to determine accuracy. Each trial is divided into 20 measurements; each measurement representing the first time a trajectory crosses a spatial threshold in either direction (each with 5% of total distance, incrementally increasing from 0). The direction of the crossing (accuracy) and time are collected at each measurement and stored in a spatiotemporal bin, marked by the combination of discrete time recorded and crossed spatial threshold. These measures are then used to calculate hazard rates (H) and conditional accuracy (CA) values in each spatiotemporal bin. CA in this case considers the average accuracy of all trials in a certain spatiotemporal bin. H is defined as the conditional probability that an event occurs within a spatiotemporal bin given that it did not occur before neither in time nor space. The results for both are presented in heat maps, where higher hazard rates show a higher probability that a trajectory will visit a certain spatiotemporal bin. These heat maps are then scaled according to the maximum hazard across all conditions. Similarly, conditional accuracy is represented in heat maps as well. StSA, including the visualizations of H and

CA in plots, could enable inferences about the fine-grained dynamics of individual differences in cognitive tasks.

ANOVAs were carried out using JASP (0.17.1); discrete SA was done using Rstudio (2023.03.1, running R version 4.0.2), StSA trajectory analyses and visualizations were carried out using MATLAB (Version: 9.4.0.813654 (R2018a)).

Results

Means Analyses using ANOVA

We first analyzed the mean RT in the classical version and compared it to Halfway times and Arrival times in VR. We then analyzed accuracy -the complement of ER- in the classical version and compared them to accuracy extracted from Halfway times and Arrival times in VR.

A total of 9,120 trials were initially available for analysis. Trials at the beginning of the N-back blocks are task-irrelevant and are thus removed from the analysis, leaving 8664 trials, 4332 for classical, and 4332 for VR. The data were analyzed as follows: First, the mean correct RT and ER were inspected for the classical version, as well as the mean correct Arrival times, mean correct Halfway times, and their respective ER for the VR version. We performed three sets of analyses. First, three-way ANOVAs, with the factors trial type (Match vs non-Match), N-back (1-3), and device (PC vs VR), were performed to compare classical measures with the VR Halfway measures, one for each of the two dependent variables, correct RT and ER. Second, the same analysis was performed to compare the classical measures with the VR Arrival measures. Third, separate two-way ANOVAs were calculated for classical, VR Halfway, and VR Arrival measures each.

Next, the same ANOVAs were repeated with data trimmed for outliers using quartile outlier rejection, (1.5 Interquartile range), trials that lay outside the range were trimmed in each

experimental condition, leaving 4091 (5.56% removed) trials for classical, 4034 (6.87% removed) trials for VR Arrival, and 4028 (7.01% removed) trials for VR Halfway analyses. Further, trials outside the range of the mean plus two times standard deviation were excluded from the analysis, leaving 3884 (5.05% removed) trials for classical, 3837 (4.74% removed) trials for VR Halfway times analyses, and 3841 (4.78% removed) trials for VR Arrival times. Finally, error trials were excluded from RT (6.2% removed), Halfway times analyses (7.92% removed), and Arrival times (7.75% removed).

Reaction Time (RT)

We report the results of analyses that are conducted with trimming, shown in Figure 13 and 14. Those of the untrimmed version will only be reported if their significance levels differ. Full results for the untrimmed data are available in the supplementary materials. We compared the correct RT in the classical version with the correct Halfway times in the VR version (see Fig. 3, first and second column). The 2 (Device: PC vs VR) x 2 (Trial type: Match vs non-Match) x 3 (N-back: N1, N2, N3) repeated-measures ANOVA revealed a main effect of Device [$F(1, 18) = 5.505$, $p = .031$, $\eta^2 = .026$]. In addition, main effects for N-back [$F(2, 36) = 28.524$, $p < .001$, $\eta^2 = .386$] after correction and Trial type[$F(1, 18) = 11.163$, $p = .004$, $\eta^2 = .025$] were obtained. We obtained interactions between Device and Trial type [$F(1, 18) = 4.451$, $p = .049$, $\eta^2 = .003$], Device and N-back[$F(2, 36) = 11.009$, $p = .002$, $\eta^2 = .048$] after correction, as well as between Trial type and N-back[$F(2, 36) = 4.402$, $p = .026$, $\eta^2 = .006$] after correction. Finally, there was no three-way interaction between Device, Trial type, and N-back [$F(2, 36) = 0.950$, $p = .378$] after correction.

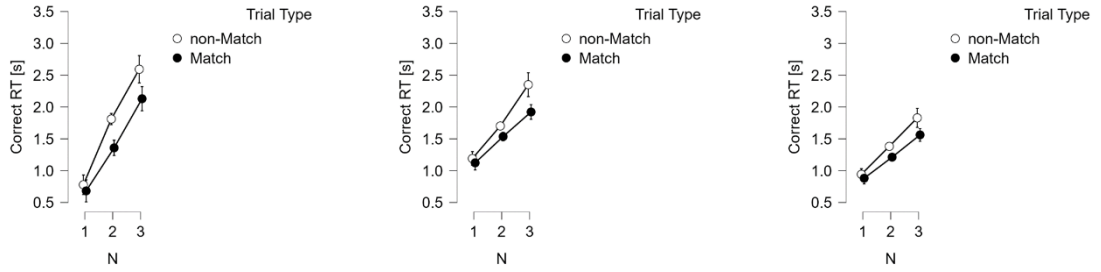


Figure 13. Averages for the classical RT (left), Halfway times (middle) and Arrival times(right). Vertical bars represent standard errors.

Next, we compared the correct RT in the classical task with the correct Arrival times in VR (see Fig. 3, first and last column). There was no main effect of Device [$F(1, 18) = 0.996$, $p = .332$], while main effects of N-back [$F(2, 36) = 26.446$, $p < .001$, $\eta^2 = .434$] after correction, and Trial type [$F(1, 18) = 12.692$, $p = .002$, $\eta^2 = .030$] were obtained. We obtained interactions between Device and N-back [$F(2, 36) = 8.267$, $p = .005$, $\eta^2 = .027$] after correction, as well as between Trial type and N-back [$F(2, 36) = 6.066$, $p = .013$, $\eta^2 = .009$] after correction. However, there was no interaction between Device and Trial type [$F(1, 18) = 2.478$, $p = .133$]. Additionally, there was no three-way interaction between Device, Trial type, and N-back [$F(2, 36) = 1.196$, $p = .307$] after correction.

To compare effects across measures, we followed up with 2 (Trial type) x 2 (N-back) repeated-measures ANOVA for each measure. Starting with the classical task, there were main effects of N-back [$F(2, 36) = 24.989$, $p < .001$, $\eta^2 = .508$] after correction and Trial type [$F(1, 18) = 13.134$, $p = .002$, $\eta^2 = .033$]. Further, there was also an interaction effect [$F(2, 36) = 3.680$, $p = .036$, $\eta^2 = .008$] after correction, which was not obtained without trimming [$F(2, 36) = 3.156$, $p = .060$, $\eta^2 = .007$] even after correction.

Concerning the Halfway times in VR, there was a main effect of N-back [$F(2, 36) = 21.623$, $p < .001$, $\eta^2 = .419$] after correction and Trial type Trial Type [$F(1, 18) = 4.608$, $p = .046$, $\eta^2 = .028$]. Further, there was no interaction effect [$F(2, 36) = 1.504$, $p = .238$, $\eta^2 = .007$] even after correction.

Finally, for Arrival times in VR, there were main effects of N-back [$F(2, 36) = 19.873$, $p < .001$, $\eta^2 = .425$] after correction and Trial type [$F(1, 18) = 7.686$, $p = .013$, $\eta^2 = .032$]. Further, there was a significant interaction effect [$F(2, 36) = 4.157$, $p = .046$, $\eta^2 = .015$] after correction, which was not obtained without trimming [$F(2, 36) = 3.352$, $p = .077$, $\eta^2 = .016$] after correction.

Accuracy

In line with RT analyses, we first compared the accuracy in the classical task with the accuracy at Halfway (see Fig. 4, first and second column). Accuracy means the percentage of correct answers in each condition, which is the complement of ER for the classical experiment, and its counterparts for Arrival and halfway in VR. We obtained no main effect of Device [$F(1, 18) = 0.018$, $p = .896$]. We obtained main effects of N-back [$F(2, 36) = 39.945$, $p < .001$, $\eta^2 = .341$] after correction and Trial type [$F(1, 18) = 10.527$, $p = .004$, $\eta^2 = .054$]. Two-way interactions between Trial type and N-back [$F(2, 36) = 5.098$, $p = .029$, $\eta^2 = .033$] were obtained after correction. In contrast, no two-way interactions between Device and N-back [$F(2, 36) = 0.089$, $p = .841$] even after correction nor Device and Trial type [$F(1, 18) = 2.727$, $p = .116$] were obtained. A three-way interaction between Device, Trial type, and N-back [$F(2, 36) = 0.079$, $p = .826$] was not obtained, even after correction.

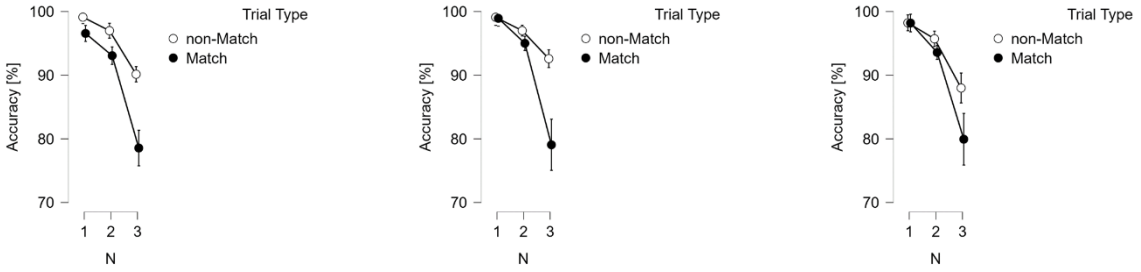


Figure 14. Average accuracy for classical (left) VR Halfway (middle) and VR Arrival (right). Vertical bars represent standard errors.

We then compared the accuracy in the classical task with the accuracy on Arrival in VR (see Fig. 4, first and last column). Again, there was no main effect of Device [$F(1, 18) = 1.471$, $p = .241$]. However, we obtained main effects of N-back [$F(2, 36) = 35.973$, $p < .001$, $\eta^2 = .328$] after correction and Trial type [$F(1, 18) = 17.096$, $p < .001$, $\eta^2 = .079$]. We obtained no interactions between Device and Trial type [$F(1, 18) = 0.410$, $p = .530$] nor between Device and N-back [$F(2, 36) = 0.035$, $p = .909$] even after correction. However, we obtained an interaction between Trial type and N-back [$F(2, 36) = 10.574$, $p = .003$, $\eta^2 = .061$] after correction. There was no three-way interaction between Device, Trial type, and N-back [$F(2, 36) = 0.751$, $p = .424$] even after correction.

Again, we followed up with 2 (Trial type) x 2 (N-back) repeated-measures ANOVAs for each measure to compare results across devices. Concerning the classical task, we obtained a main effect of N-back [$F(2, 36) = 35.240$, $p < .001$, $\eta^2 = .399$] after correction. We obtained a main effect of Trial type [$F(1, 18) = 15.494$, $p < .001$, $\eta^2 = .107$] and an interaction [$F(2, 36) = 7.046$, $p = .006$, $\eta^2 = .047$] after correction.

Next, we considered the accuracy at Halfway in VR. We obtained a main effect of N-back [$F(2, 36) = 24.212$, $p < .001$, $\eta^2 = .325$] after correction. However, we did not obtain a main effect of Trial type [$F(1, 18) = 3.697$, $p = .070$, $\eta^2 = .026$], this effect was however present in the untrimmed data [$F(1, 18) = 4.462$, $p = .049$, $\eta^2 = .033$]. Additionally, there was no interaction [$F(2, 36) = 1.796$, $p = .196$].

Finally, for accuracy on Arrival in VR we obtained main effects of N-back [$F(2, 36) = 22.088$, $p < .001$, $\eta^2 = .318$] after correction, Trial type [$F(1, 18) = 13.815$, $p = .002$, $\eta^2 = .075$] and an interaction [$F(2, 36) = 8.131$, $p = .007$, $\eta^2 = .078$] after correction.

Summary and Discussion

As N increased, responses became overall slower and less accurate across all measures. Additionally, responses in Match trials were faster than non-Match trials in all measures. Similarly, Match trials were more accurate than non-Match trials in the classical task and in VR in the Arrival times and the untrimmed Halfway times. This was not the case, however, for the trimmed ones. Higher values of N were associated with slower and less accurate responses. This effect was more pronounced in the classical version than in VR conditions. While the N-back effects are present in the Halfway times, it is more pronounced in the Arrival times. Overall, we may conclude that the Arrival times in VR consistently replicate all classical effects of the N-back task (Meule, 2017).

Remarkably, this consistency occurs, despite the large differences with the classical response times due to the additional requirements of the VR response procedure. These differences are most marked for N = 1; for larger Ns, they are absorbed in the VR response execution times. This suggests that the decision processes in the VR condition may overlap with response execution. Because of this overlap, we should be able to interpret deviations or delays at different times during the

trajectory as indicative of a decision-making processes (Cressman et al., 2007; Rheem et al., 2018; F. Schmidt & Schmidt, 2010).

The Halfway times of trajectories in all Ns are up to .7 seconds faster than the classical response times. Nevertheless, their accuracy is clearly above chance levels, even rivaling classical RT and Arrival times. This accuracy indicates that a substantial amount of decision-making has been done at this initial stage of response execution, further indicating that response execution and decision-making overlap during this task.

For classical response conditions, the overlap between decision and response execution is typically considered to be minimal. Participants make decisions after accumulating sufficient information. Drift-diffusion models (DDM) represent the accumulation as a random walk with drift, continuing until one of the alternative decision boundaries is reached, prompting response initiation (Ratcliff, 1978; for an overview see Ratcliff et al., 2016). DDM can address a variety of characteristic phenomena, including speed-accuracy tradeoffs; faster responses are less accurate than slower ones (Ratcliff, 1978, 1981). Furthermore, DDM show robust performance when response deadlines are introduced, limiting the amount of information the participant is allowed to accumulate before decision-making (Karşılar et al., 2014). The fixed decision boundaries in DDM, however, cannot accommodate real-time changes in the decision process that occur during response execution, like those observed in VR trajectories (Cressman et al., 2007; Rheem et al., 2018; F. Schmidt & Schmidt, 2010).

We may raise the question, however, to what extent it is justified to assume only minimal overlap between decision-making and response execution in the classical task. On one hand, the motion component in the classical task is extremely limited. On the other hand, key press responses are

liable to bottom-up distraction and top-down inhibition just as well as VR motion trajectories. So the DDM may face limitations for classical responses as well. This may raise doubts on the standard procedures of trimming the data, as these may systematically eliminate such effects from the results.

Variance Tables for the means analysis

To estimate the extent of selective elimination and its effects, we tabled the variances of the means (Table 1), across conditions (Table 2) and across individual participants and conditions (Table 3).

Tables 1 and 2 show that across all conditions, variance tends to increase with task difficulty: lowest for N1, larger for N2, and still larger for N3. Non-matches show a larger variance than Matches, except notably for N1. Trimming reduces the variance in all the conditions. However, we still observe differences of variances across conditions. In addition, this procedure has little to no effect on sphericity (comparing assumption checks tables in the supplementary materials). Violations of sphericity could be corrected by Greenhouse-Geisser but call the results of the ANOVA into question (see Armstrong, 2017); It suggests that such data are not a great match for ANOVA. ANOVA assumes that error variance is additive to the task effect. In fact, the noise is more likely to be multiplicative. The more difficult the task, the more room for diversions, hesitations, and self-corrections that would increase the noise. Table 3 shows that the effect of increasing variance depends on individual participants; some show much larger increases with task difficulty than others, suggesting that not all individuals are subject to the same degree of hesitation, self-correction, etc., as some will find the N3 task still relatively easy while others struggle.

Table 1: Variance of the means

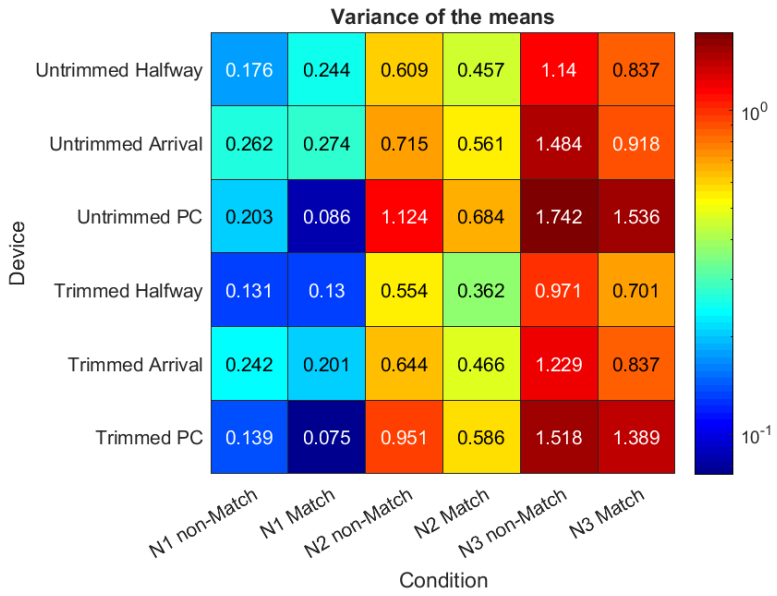


Table 2: Average variances calculated for each condition

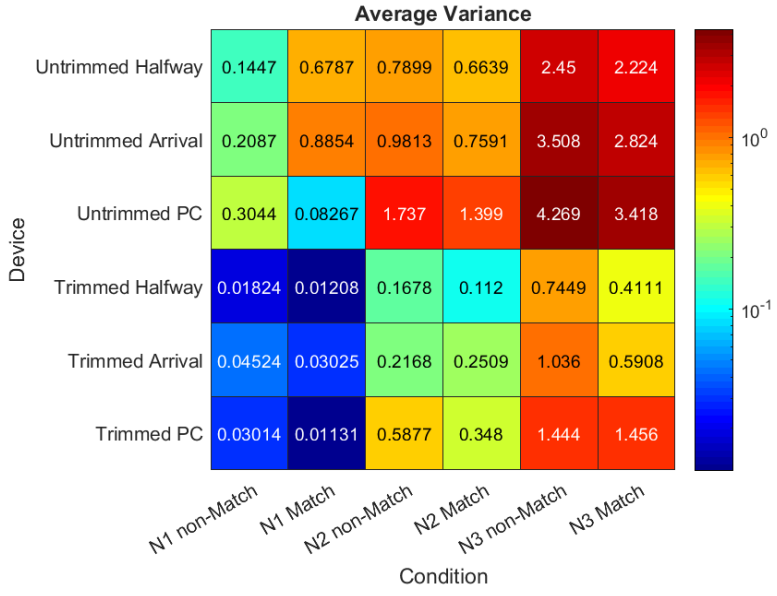
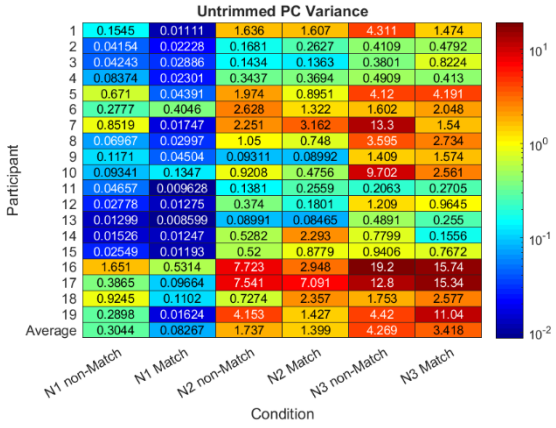
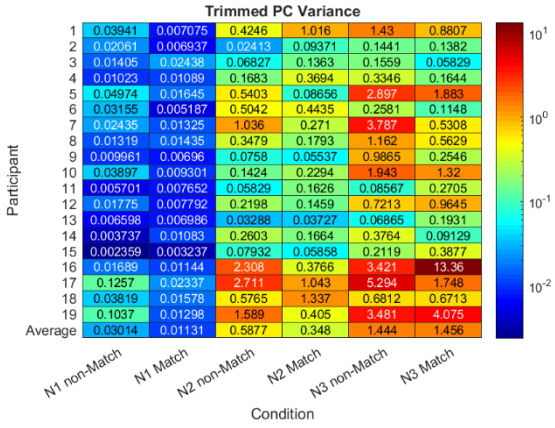
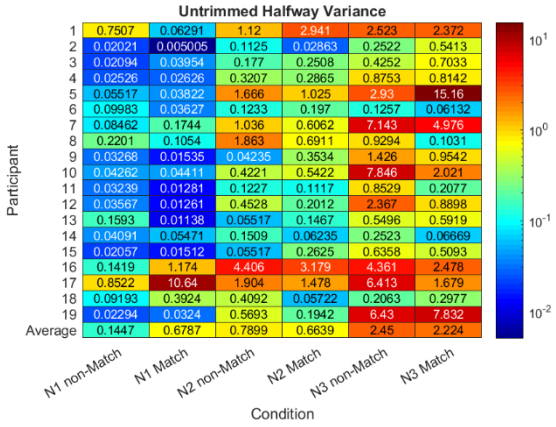
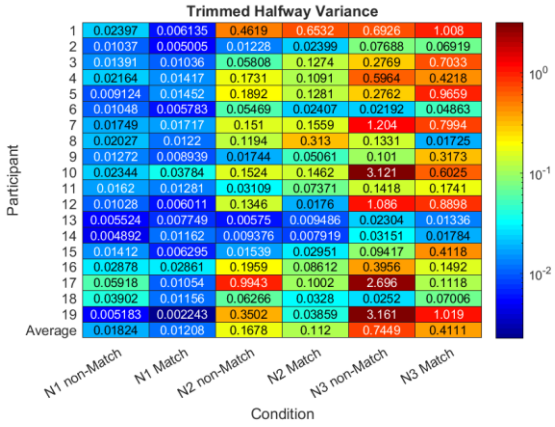


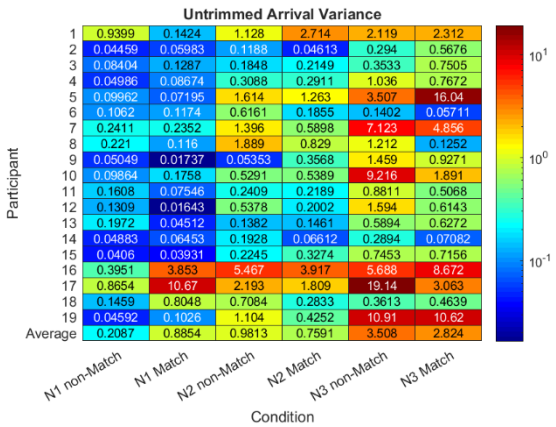
Table 3: Variances calculated for each participant and condition

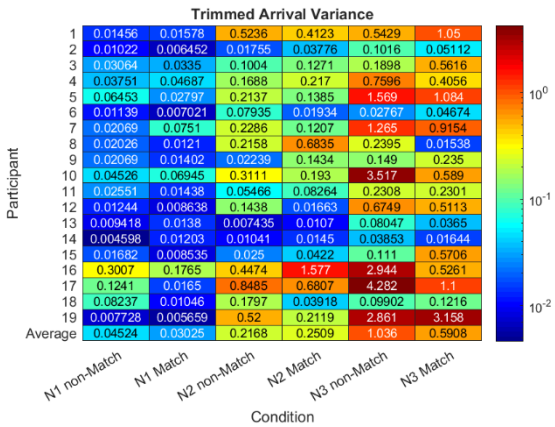












Note that the concerns about increasing variance apply equally to the VR trajectory times and the classical RT. In the ANOVAs it was shown that for the trajectories, there is substantial overlap between decision-making and response execution. Here, even though movement is extremely restricted in the classical condition, the increases in variance with N are of the same order of magnitude as those of the trajectories. This indicates that, rather than noise, they should be considered, as for the trajectories, as showing that decision-making is not distinct in time from response execution.

Survival Analysis (SA)

We calculated sample-based estimates of h(t), S(t), and ca(t) for each experimental condition (trial type (Match vs non-Match), N-back (1-3)), for each of the measures (RT, Halfway times, Arrival times). To obtain a high temporal resolution, we divided the first 7,000 ms after item onset into 70 discrete temporal bins of 100 ms (indexed t = 1 70; e.g., bin 37 = bin 3,700 = (3600,3700]). Accordingly, trials after 7,000 ms were right-censored observations, as only a limited number of responses occurred after and were thus seen as uninformative. Note, that this is not equivalent to outlier removal and trimming when dealing with mean data and its analysis. In contrast, SA includes information on right-censored trials and doesn't run the risk of underestimating the means and variances (for more detail, see Panis, et al., 2020).

We first present SA results for the classical, PC-obtained RT, then the VR Halfway times, and finally the VR Arrival times.

Survival Analysis for the classical Data

In N1, the first responses occur after about 400 ms in both Match and non-Match conditions. However, responses occur at a higher rate in the Match compared to the non-Match condition

(Figure 15, panel A). For the next 800 ms, hazard rates stay higher in the Match compared to the non-Match condition. Note that 95% of responses occur within the first 1,700 ms (Figure 15, panel D). Remarkably, the earliest responses are 100% correct in both Match and non-Match conditions (Figure 15, panel G). However, as time progresses, responses become less accurate in the Match condition (0% accuracy at (1300,1400]) but stay at around 100% accuracy in the non-Match condition, indicating that as time progresses a bias towards non-Match emerges for the remaining responses. Note, however, that most responses have already occurred at this point (Figure 15, panel D).

As N increases, first responses continue to occur after about 400 ms (Figure 15, panels B and C). However, hazard rates flatten out compared to N1. This is evidence that the decision-making process takes more time as task difficulty increases. Because of that, it now takes the first 5,300 ms in N2 for 95% of responses to occur, and 6,900 ms in N3 (Figure 15, panels E and F). Note that hazards remain higher in the Match compared to the non-Match condition throughout almost the entire distribution. While conditional accuracy develops similarly compared to N1 (early correct responses, followed by the emergence of a non-Match bias for the remaining trials, Figure 15, panel H), N3 shows a remarkable difference (Figure 15, panel I). Earliest responses here are 0% correct in the non-Match condition, and 100% correct in the Match condition. This indicates an early bias for Match, or in other words, participants responding fast in N3 are heavily biased towards Match responses. After 800 ms this shifts to a bias towards non-Match, as responses continue to be more accurate in the non-Match condition. Note, that overall accuracy decreases with N, as more responses occur at lower conditional accuracy (as seen in $S(t)$ and $ca(t)$ in Figure 15).

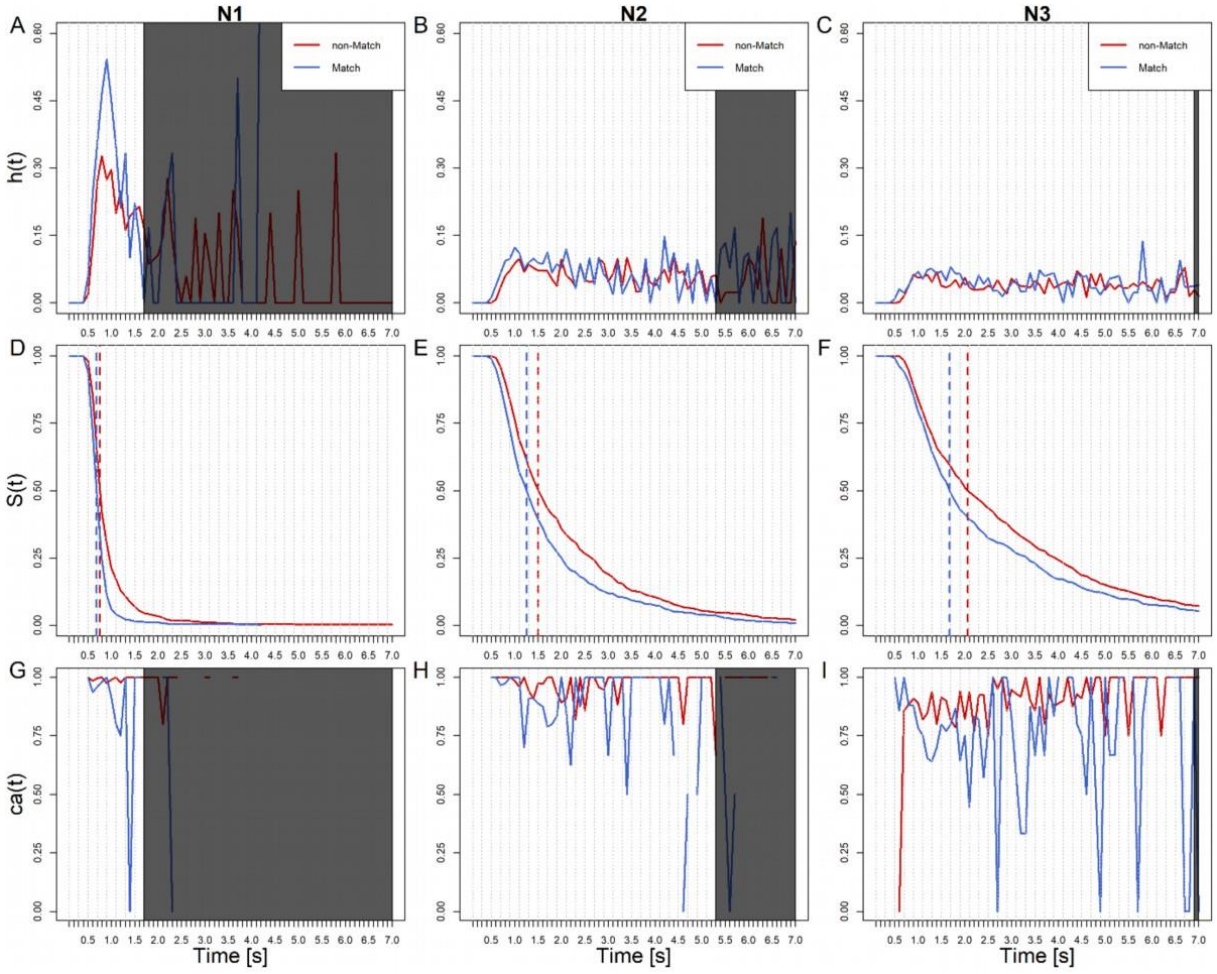


Figure 15. Sample-based estimates of h(t), S(t), and ca(t) aggregated across all participants in the classical measure, for the first 70 bins (or 7,000 ms) after item onset. Bin width equals 100 ms. Each column represents a different N-back condition (in ascending order). Blue lines represent Match trials, red lines non-Match trials. Dotted vertical lines represent median RT estimates for Match and non-Match trials. Note that the greyed areas in h(t) and ca(t) plots obscure remaining time bins that contain less than 5% of trials (S(t>T) = .05). This is to highlight the more informative left side of the distributions.

Survival Analysis for Halfway VR Data

We now consider the VR Halfway times (Figure 16). Overall, the distributions are similar to those of the classical RT. In general, hazard rates remain higher for the Match compared to the non-Match condition. N1 again shows virtually perfect performance in the earliest events but does not reveal a late non-Match bias (Figure 16, panel G). In contrast to the classical results, the VR Halfway times reveal an early Match bias already in N2, which then shifts again to a non-Match bias after about 1,000 ms from item onset. This effect can also be found in N3. Note that overall, the 95% criterion of events occurring is reached earlier than in the classical RT. Halfway to completion of a movement in VR is simply reached faster than the completion of a classical key press.

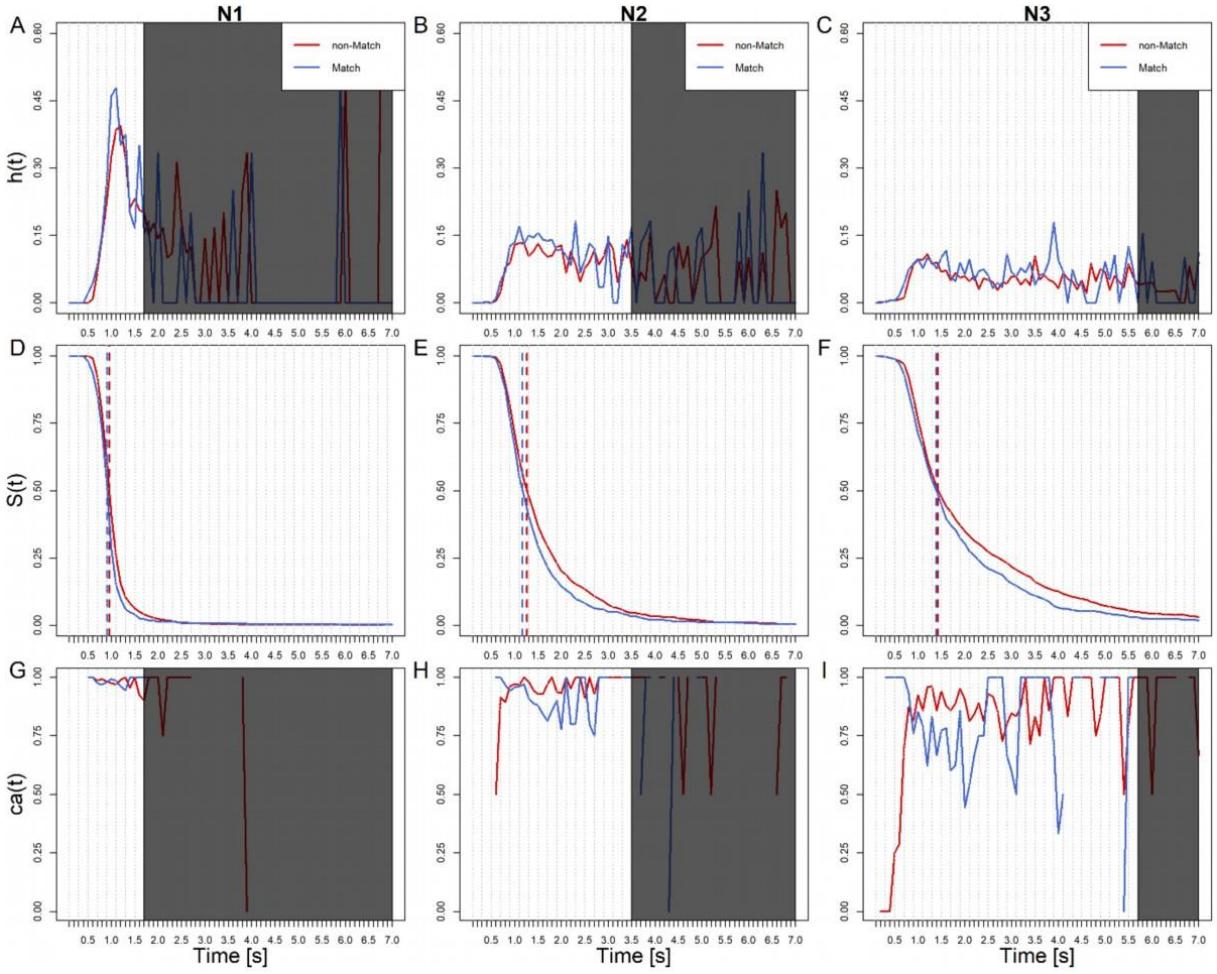


Figure 16. Sample-based estimates of h(t), S(t), and ca(t) aggregated across all participants in the VR Halfway measure, for the first 70 bins (or 7,000 ms) after item onset. Bin width equals 100 ms. Each column represents a different N-back condition (in ascending order). Blue lines represent Match trials, red lines non-Match trials. Dotted vertical lines represent median RT estimates for Match and non-Match trials. Note that the greyed areas in h(t) and ca(t) plots obscure remaining time bins that contain less than 5% of trials (S(t>T) = .05). This is to highlight the more informative left side of the distributions.

Survival Analysis for Arrival VR Data

Finally, we consider VR Arrival times. Events generally occur at a slower rate in Arrival times than in the classical RT and VR Halfway times (Figure 17). Regarding N1, the effects are again similar: early accuracy at about 100% (Figure 17, panel G). Interestingly, hazard rates initially increase at the same time (~600 ms) for both Match and non-Match conditions. However, hazard rates start to be higher in the Match condition only after 900 ms from item onset (Figure 17, panel A). This is also apparent in N2 (Figure 17, panel B). In contrast to both classical RT and VR Halfway times, there is no indication of an early Match bias, neither in N2 nor N3 (Figure 17, panels H and I). Instead, early arrivals are always at the correct location, followed in the Match condition by a decrease in accuracy (again an indication of a late bias towards non-Match). Thus, while classical RT and VR Halfway times show an early bias towards Match, participants can correct for this bias in the VR Arrival times. In other words, initial movement (VR Halfway times) and classical button presses (RT) capture early biases, while participants use their full movement trajectory in VR to overcome those biases upon arrival.

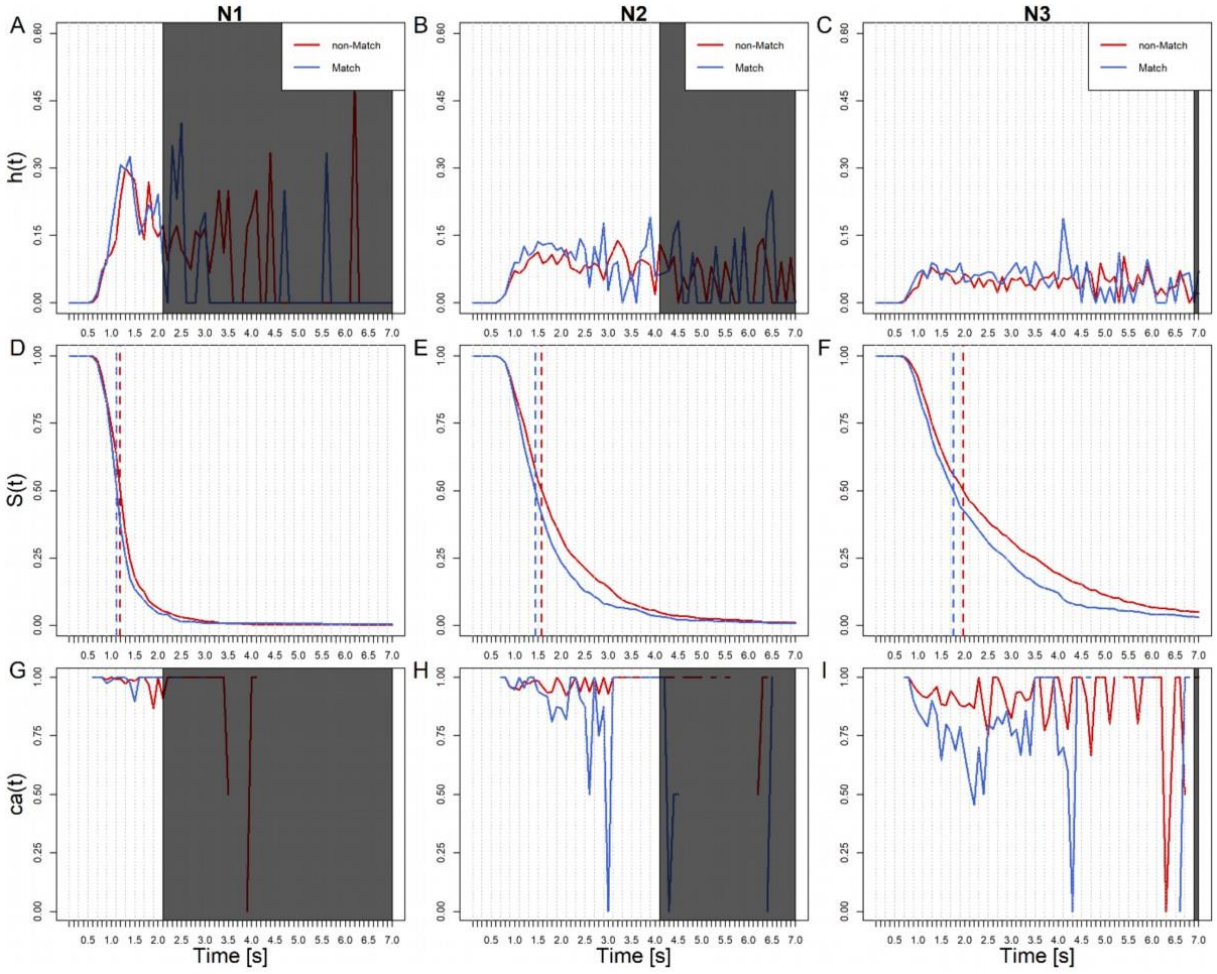


Figure 17. Sample-based estimates of h(t), S(t), and ca(t) aggregated across all participants in the VR Arrival measure, for the first 70 bins (or 7,000 ms) after item onset. Bin width equals 100 ms. Each column represents a different N-back condition (in ascending order). Blue lines represent Match trials, red lines non-Match trials. Dotted vertical lines represent median RT estimates for Match and non-Match trials. Note that the greyed areas in h(t) and ca(t) plots obscure remaining time bins that contain less than 5% of trials (S(t>T) = .05). This is to highlight the more informative left side of the distributions.

Summary and Discussion

SA confirmed that responses in Match trials were faster than non-Match trials across all devices and measures, responses in Match trials are generally less accurate, and as N increases, response latencies increase and accuracy decreases. In addition, SA establishes differences in response behavior (mainly in hazard rates, but also present in conditional accuracy) across all measures. In particular, SA revealed evidence for an early bias (<800ms) towards Match for higher values of N in both Match and non-Match conditions, in the RT and Halfway times. This is despite the fact that only 33% of trials were Match trials. From 800ms onwards, there is a non-Match bias across all measures. Note that 800ms is the time when the earliest responses are recorded for Arrival times, which indicates that participants have already overcome early Match biases before arrival. Match responses are consistently faster than non-Match responses. Later responses become more accurate in both conditions, showing the end of the response bias. It is interesting that classical RT are more similar to Halfway than to Arrival times in this respect. The occurrence of these biases during the response process is shown here not to be exclusive to the VR response trajectories, but occurs in the classical conditions just the same, even though response execution requires only minimal movement. To understand the origin of these biases, we applied our novel StSA method.

Spatiotemporal Survival Analysis (StSA) pipeline, Stage 1: Average Trajectory Data

Figure 18 shows average trajectory data for each condition. In all three N conditions, in the first 500 milliseconds, average trajectories keep close to the origin. This happens, despite SA showing a strong Match bias for early trials in this period. Averaging trials runs the risk of losing relevant information, such as early biases. Around 500 ms, a marked deviation from the origin begins to differentiate Match and non-Match trials. Note that in the SA, this point in time marks the transition from Match to non-Match bias. As non-Match trials are a 66 % majority, this bias is manifested as a

difference in slope between Match and non-Match trials. Non-Match trials move faster to response completion in this stage. Beyond 700 ms, the trials still seem to move to completion at different rates, more slowly for the Match condition than for the non-Match condition. But in fact, this is true only for the few trials that have not yet come to completion. These are fewer in the case of Match than in Non-match trials, as the latter tend to be slower.

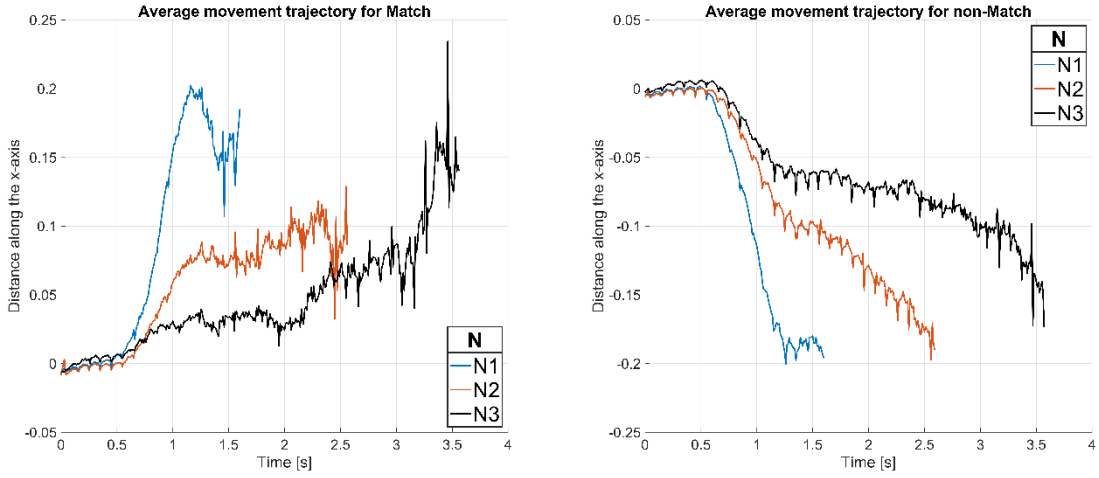


Figure 18. Averaged data across N for Match (left) and non-Match (right).

In both Match and non-Match conditions, the slope of deviation becomes less steep with increasing N. This difference in slope is maintained during most of the response process, showing that participants move at lower speeds throughout the more difficult conditions. The N1 average has little noise in both Match and non-Match conditions, suggesting that most trials share similar characteristics. In N2, the early part of the trajectory has little noise in the Match condition, while later parts are noisier, suggesting a higher variance in Arrival times. There is less noise in the non-Match condition resulting in a more compact distribution of trial Arrival times. For N3, this variance between trials becomes more apparent, causing the average to be close to zero for a longer time.

The findings suggest that as the difficulty level increases, participants exhibit slower and more variable responses. This shows the importance of considering both speed and variance to comprehensively understand participants' reactions across varying difficulty levels. In the next stage, we plot individual trials, exploring the differences between conditions.

Stage 2: Single-trial data

Figure 19 shows movement trajectories for all trials in all conditions with all data included, also the error trials and those rejected as outliers in the ANOVA. Match trials are shown in blue while non-Match are shown in red. The blue and red horizontal lines (between 2.2 and -2.2) indicate the location of the centers of the green and red boxes, respectively. Note that arrival times were registered as soon as the trajectory collided with the box, not when it reaches the center. As a result, trajectories have slightly different endpoints in space. All the trials start at a neutral position between the boxes. If participants do not move or move on the vertical plane while not getting closer to either box, their spatial position remains at zero. At varying times, participants initiate movement to either the left or the right box. In some trials, we notice *hesitation* and *self-correction* in movement; hesitation is indicated by slowing down or prolonged hovering at the same spatial location before moving closer to either box. Self-correction amounts to switching the direction of the movement mid-trial.

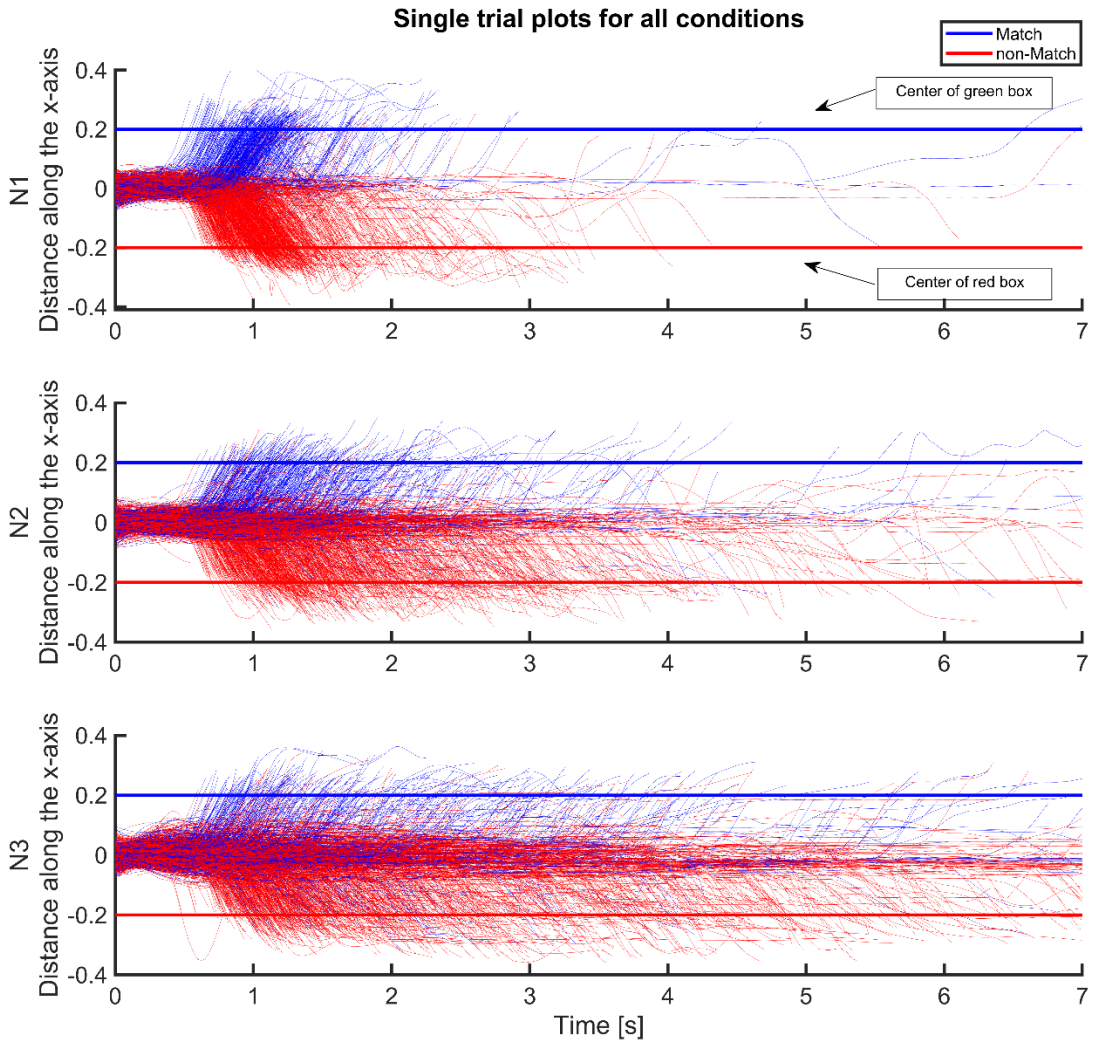


Figure 19. Single trial data for the different conditions. non-Match trials are shown in red, and Match are shown in blue. Red and blue straight lines are the centers of each box.

In N1 most trials end before 2 seconds, with little hesitation or self-correction across the trials. In N2, while most of the trials end before 3 seconds, movement initiation is more spread out, indicating increased hesitation. Self-correction is also more prominent in these trials, with many trials moving initially in the wrong direction. Finally, for the N3 condition, movement initiations are

even more spread out. Whereas most trials end before 4 seconds, many trials are still ongoing past 7 seconds. Many trials are crossing over and hovering for a long period between boxes before a decision is made.

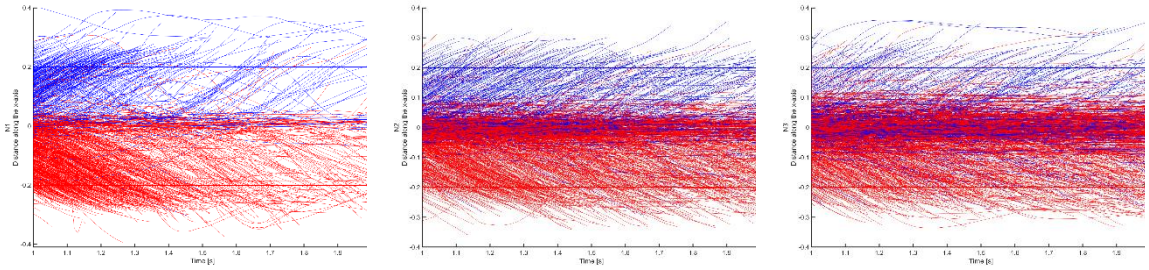


Figure 20. Zoom-in on the single trial data for the different conditions. Target trials are shown in red and non-targets are shown in blue.

Figure 20 zooms in on the single trial data between 1 and 2 seconds in all conditions. non-Match trials -shown in blue- have a clear separation from the Match trials –shown in red- in N1. As the task gets more difficult, two effects could be observed. First, more crossings between red and blue, meaning more movements in the wrong direction for both. Second, a wider concentration of trials in the middle, indicating that the more difficult the task, the longer participants tend to stay undecided.

Stage 3: Probability distribution function

Figure 21 shows the log-transformed 2D probability distribution function of the raw data in a high-resolution heat map, which represents the frequency distribution of trajectories visiting each spatiotemporal bin. The x-axis represents time bins, while the y-axis represents spatial bins, totaling 101 spatial and 701 temporal bins, resulting in 70,801 discrete spatiotemporal bins. By

incorporating data from single trials, these graphs show a finer level of detail compared to the averaged trajectory data.

In the N1 condition, for both Match and non-Match conditions, the probability remains notably high within the central region (between 0.8 and -0.8 spatially) until approximately 1 second into the experiment, indicating that for many trials, the initiation of movement starts around that time. These heat distributions typically appear narrow, resembling a single trajectory.

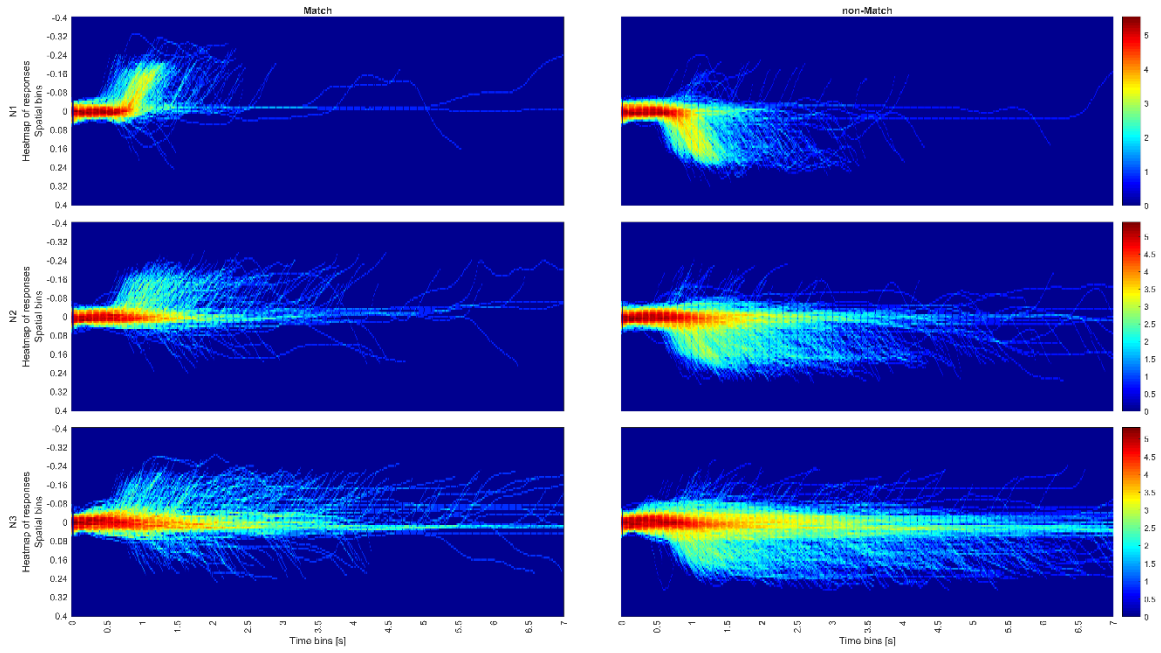


Figure 21. Log transformed 2D Probability distribution function for discrete time and space. Left Column is Targets; Right Column is Distractors. Each Row is for each N (1:3).

With increasing N, trajectories tend to linger in the center for a longer time. Specifically, for N2 and N3, the durations are approximately 1.5 and 2 seconds, respectively. This increase in trajectory duration is also associated with a more expansive distribution of responses across both space and time, most prominently observed in N3, which explains the noisy averages observed in stage 1. This

expansive distribution can indicate some overlap between decision making and movement, where participants initiate movement in some direction early in the decision making process, and either correct or confirm that movement later.

When comparing Match and non-Match conditions, we note that in non-Match conditions, the distributions are wider, indicating hesitation. Additionally, there are more deviations in the wrong direction in non-Match conditions, suggesting increased self-correction. Generally, Match trials tend to show a higher probability of deviating from the center in earlier time bins compared to non-Match trials. This effect is easier to establish here than in the average trajectories.

Stage 4: Spatiotemporal Survival Analysis (StSA)

Figure 22 shows the StSA of the trajectories in all conditions. There are twelve graphs in total, 6 for hazard (denoted as H) and 6 for conditional accuracy (denoted as CA). The x-axis shows time bins while the y-axis shows spatial bins. There are 1,400 spatiotemporal bins (20 spatial by 70 temporal). The lower resolution than in Stage 3 of the analysis ensures an adequate number of events per bin.

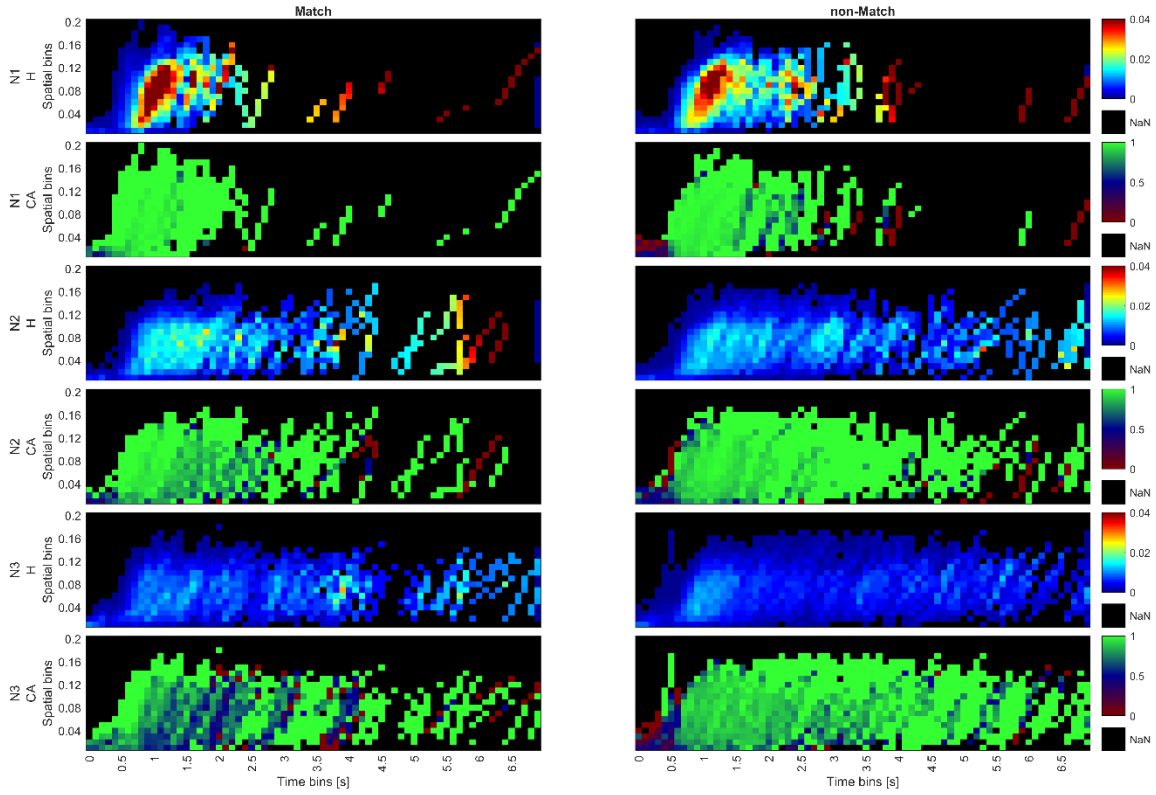


Figure 22. StSA for all conditions. Each Row is for two dimensional Hazards (denoted as H) and conditional accuracy (denoted as CA) for each N (1:3).

For both Match and non-Match trials in the N1 condition an area of interest can be identified, where hazard rates are highest between 900-1200 ms in the spatial bins between 0.04-0.12 (which corresponds to the criterion for the Halfway times in previous analyses). This shows a high across-trial consistency of movement passing the halfway point around that time. The high hazard rate, i.e. high likelihood of response occurrence around the halfway point suggests a level of confidence, confirming the utility of Halfway times in the ANOVAs and SAs as akin to RT.

Hazard rates drop rapidly after 1500 ms for Match but persist longer for non-Match within the same spatial bins, leading to higher Arrival times. Given the parallels between Halfway times and RT, persistence at halfway may be the reason for the overall higher response times in the non-Match condition. Hazard rate exhibits a slight resurgence around 1800 ms in the Match condition, whereas for the non-Match condition, the hazard rate slowly declines with time. This is consistent with the graphs from Stage 3 where response distributions of trials in the non-Match condition in time were wider.

Both the persistence at halfway and the late resurgence in the Match condition become more pronounced as N increases. For both N2 and N3, the hazard is highest around the halfway point in the 500-800 ms interval of the trial, declining sharply for the Match condition and gradually waning for the non-Match condition. For the Match condition, the resurgence of the hazard rate for N2 occurs at 2000ms and for N3 at 2500 ms. This results in two hazard peaks across all Match trials, which can be observed in Figures 23-25.

In the Match condition, high hazard is associated with high conditional accuracy. On the other hand, conditional accuracy in the non-Match condition in spatiotemporal bins with high hazard is slightly lower. Conditional accuracy in the bins at the origin (0-0.4 temporal, 0-0.4 spatial) tends to be highly accurate in the Match condition, but highly inaccurate in the non-Match condition. This shows the strong Match bias is starting from movement initiation. This bias can be observed past 400ms with increasing N and is overcome only if movements are initiated later in time.

These phenomena are more conspicuous in N2 and N3, where there is an initial bias towards Match that is evident in both conditions. In line with SAs, this bias is surprising considering that only 33% of trials were Matches. As a result of this bias, movements initiated earlier in time within the non-

Match condition tend to be less accurate, whereas this accuracy improves for responses initiated later. As for the Match condition, particularly fast movements early in time as well as movements later exhibit a heightened level of accuracy in comparison to those initiated during intermediate time frames.

Additionally, it's noteworthy that the closer the participants are to arrival, the higher their accuracy. This implies that as the decision process advances toward its conclusion, movements tend to be more precise.

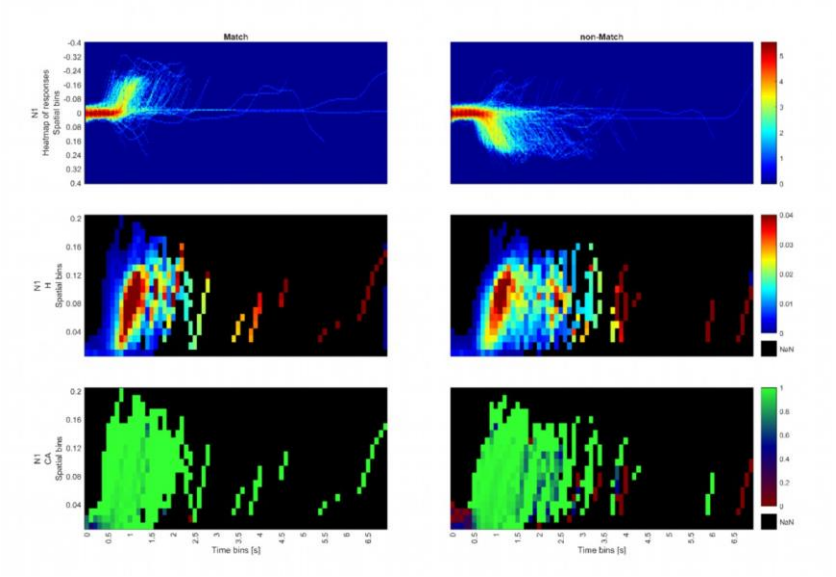


Figure 23. Combined plots for N1; Top row shows the heat map of responses; middle row shows hazard rates while the bottom row shows conditional Accuracy.

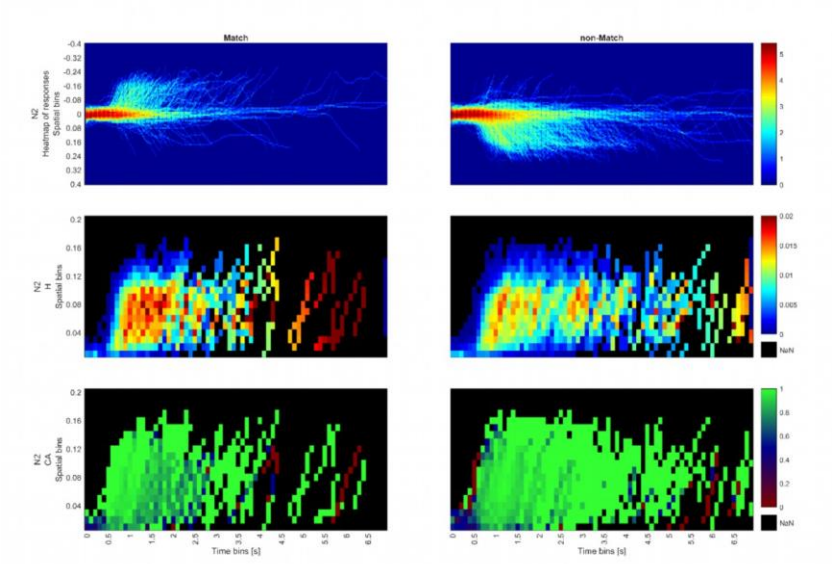


Figure 24. Combined plots for N2; Top row shows the heat map of responses; middle row shows hazard rates while the bottom row shows conditional Accuracy.

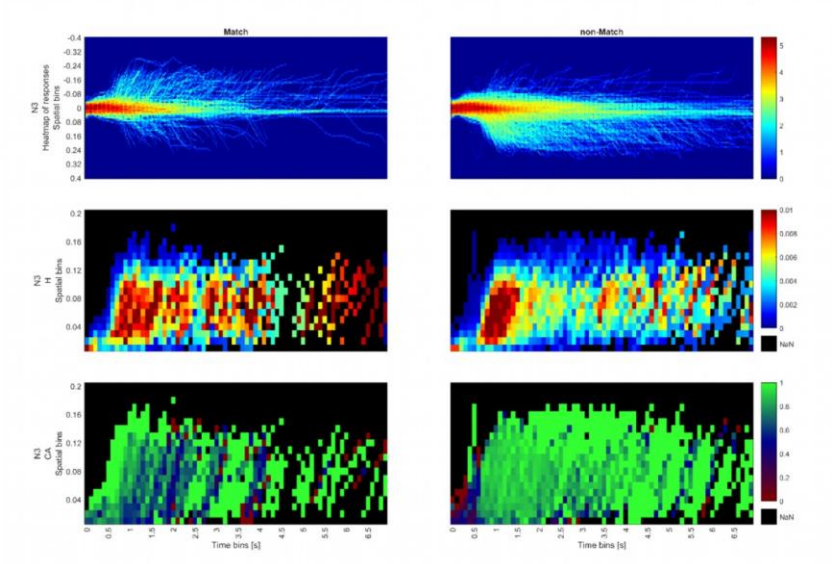


Figure 25. Combined plots for N3; Top row show the heat map of responses; middle row shows hazard rates while the bottom row shows conditional Accuracy.

Summary and Discussion

Average trajectories show the effects of the brief episodes of bias revealed in the SA. Early parts of the trajectory (<500 ms) show on average a small deviation from the origin towards the Match response box. This deviation occurs in both Match and non-Match trials indiscriminately. Later parts of the response trajectory (500-700 ms) on average reveal the opposite bias, this time differentiated in terms of the rate with which they move towards the alternative response boxes.

Single trial plots show end point variability across different trials, which explains the variability in the averaged trajectory data. In addition, spatial separation between Match and non-Match trials becomes less conspicuous with higher N values. These effects are characterized by hesitation (hovering in an undecided regions) and self-correction (changing the direction of approach). Single trial plots descriptively show these effects, while the numerical values expressed in heatmaps quantify them.

Comparing these graphs to the ANOVA and SA graphs, we can observe how the effects observed in SA start with movement initiation and continue until arrival. Like with both previous analyses, Spatiotemporal SA was able to reveal that responses in Match trials are faster than non-Match trials across all Ns. Responses in Match trials are generally less accurate, and as N increases, RT increases and accuracy decreases. As also revealed in the SA, initial responses are very accurate in N1, there is evidence for an early bias towards Match as N increases. This bias is followed by a bias towards non-Match in later responses. The spatial characteristics of this bias are revealed by the spatiotemporal SA, which shows that response initiation is associated with speed and less accuracy, while arrival is associated with precision and more accuracy. Moreover, responses generated late in time tend to be more accurate even in the early portions of the response. Overall, StSA is not only

able to reveal the classical effects shown in ANOVAs and SA but is able to reveal behavioral biases of individual responses throughout the entirety of the decision-making process. Similar to SA, StSA is not sensitive to outliers and thus can be easily generalizable to datasets with unequal variances unlike the classical ANOVA approach. Finally, StSA reveals response bias characteristics, transitions between biases to informed decisions, as well as the relationship between movement imitation times and accuracy.

General Discussion

Classical experiments lack ecological validity and rely on Response times (RT) and Error Rates (ER) as singular indicators of cognitive activity. To address these limitations, we considered the merits of Virtual Reality (VR). VR offers rich, immersive environments and facilities for tracking of response trajectories, thereby promising a rich and powerful data resource for analyzing behavior. To preserve the legacy of classical results, effects obtained in the minimalistic stimulus and response conditions of classical experiments must be reproducible in the richer environments of VR. We transferred the classical N-back task to a VR environment. For cross-validation, we compared the classical RT and ER with equally singular measures derived from the movement trajectories tracked in VR, namely Halfway times and Arrival times, and the corresponding equivalents of ER. We performed classical ANOVA, as well as Survival Analysis (SA) for RT and ER and their VR counterparts. Finally, we applied a newly proposed method for analyzing response trajectories in VR: Spatiotemporal Survival Analysis (StSA). Unlike the previous methods, StSA is an analysis specifically designed for analyzing continuous responses. We set out to demonstrate the merits of StSA by showing how it could clarify the results of the classical ANOVA and overcome violations of assumptions that are necessary for them.

Our results show that not only can the classical effects of experimental paradigms be replicated in VR, but these effects can be further explored with the continuous recording of movement trajectories. Movement trajectories provide insight here on the spatiotemporal dynamics of response activation in the N-back task, showing bottom-up effects on movement initiation, in the form of early Match biases. They also show that responses that are initiated later within the trial are free from these response biases and generally more accurate. Additionally, as trajectories approach their endpoint accuracy increases as well. These results could explain the relationship between speed and accuracy and are striking to see, given the nature of the N-back task (Meule, 2017).

(Meule, 2017) suggested that using either speed or accuracy in isolation is not enough to describe effects in the N-back task. In easier conditions, accuracy is subject to ceiling effects. They suggest using a more fine-grained analysis of the N-back task to find meaningful interpretations of the behavioral results. Indeed, our analysis does find that accuracy is quite high in most conditions, conditional accuracy however paints a different picture. Our approach provides not only a fine-grained analysis of the behavioral results, but a continuous record of behavior throughout the trial as well. These results do not require complex models or physiological data to understand the decision-making process.

The Match bias in the Classical condition, as revealed in the SA, suggests that despite response procedure differences between VR and PC, there's still an overlap in response execution and decision-making. In the classical task, participants make minimal movements with less time to correct, leading to early errors. SA reveals these early errors to be systematic and informative, meaning that trimming error trials from the RT analysis leads to a loss of valuable information on

participants' behavior. On the other hand, DDM assume a continuous accumulation of evidence leading to a decision, which oversimplifies the decision-making process (Ratcliff, 1978; Ratcliff et al., 2016). In contrast, StSA is a diagnostic tool that shows the development of decision-making and response bias in time and space.

One of the limitations of the N-back task and WM tasks in general, are the large individual differences within populations (Kane et al., 2007; Lamichhane et al., 2020). This is clearly reflected in our means analysis, which required extensive trimming and data manipulation to satisfy the conditions of ANOVA. SA and StSA bypass these limitations by using all available data for the analysis, only right censoring parts of reactions that are too long. Trials are grouped based on temporal and spatial thresholds, ensuring comparability of averaged trials and revealing trends that are otherwise concealed in the ANOVAs, such as the early and later biases we revealed in the analysis. In addition, SA and StSa can better reflect the distribution of trials, showing the effect of variance on results. Moreover, trajectories in general contribute to an increase in experimental power by giving more precise and continuous measurements. This enables small-n designs in WM research, helping the understanding of individual subjects.

Replicating the classical results in VR is consequential in itself since VR provides an outlet for ecologically valid stimulation using standardized equipment (Hoffman, 1998; Wilson & Soranzo, 2015). Trajectory tracking in VR offers a more naturalistic approach to data collection, where actions are more meaningful than pressing a button on a keyboard. This is achieved while preserving control of experimental factors such as luminance and visual angles. Continuous tracking of responses also lends itself to studies where comparison between mental chronemic measures and continuous physiological data is needed, such as EEG. The VR headset used has the advantage of an Eye-tracker

that is embedded within, for example, allowing easy combination between pupil data and movement.

As VR-based experiments provide more complex stimulation and responses, we might expect additional cognitive load (Sweller, 1988). Cognitive load encompasses factors such as task format, stimulus and response complexity, and time pressure, as well as environmental factors, such as noise and lighting, and individual factors, such as age, intelligence, and experience with the task (Sweller, 1988). We may wish to compare cognitive load in classical and VR-based experiments, using the classical experiment as a baseline, and arrive at a measure of cognitive load through subtraction. For instance, if the accuracy in a numerical judgment task obtained by button presses in a classical experiment is greater than in a VR environment, we could attribute the difference to the extra cognitive load imposed by the VR environment.

But from a VR perspective, just because classical experiments are extremely simplified, they do not necessarily minimize cognitive load. Key press responses are affected by stimulus-response compatibility, e.g., left-side responses to stimuli on the left and right-side responses to stimuli on the right are faster than vice versa (Hommel, 1997; Kornblum et al., 1990; Simon, 1969). In numerical judgment tasks, there are several varieties of implicit stimulus-response compatibility effects known as SNARC effects (Weis et al., 2018).

Moreover, certain aspects of classical tasks can be quite a burden. Impoverished stimuli may lead to disengagement from the task and mind wandering (Seli et al., 2018) such that participants struggle to maintain their focus on the task. Specifically, in computerized experiments, using a keyboard and mouse on a horizontal surface generates incongruence with the vertical computer

screen(Niu et al., 2022). These observations imply that it is not meaningful to use classical task estimates as a baseline for estimating cognitive load in a VR environment.

The evolving VR technology may prompt us to move beyond the concept of cognitive load, that subsumes under a single number a broad variety of situational and cognitive variables. VR, other than classical RT, offers a natural way of analyzing the spatiotemporal properties of decision-making, by offering natural facilities for tracking movement trajectories. The temporal dynamics of many processes can be researched this way, such as working memory, response conflict, perception, attention, and awareness. These designs allow the use of movement trajectories and have a long history of results that can be validated with this design. Our adapted setup, as well as the StSA analysis method, lends itself to moving forward into a more behaviorally derived understanding of human cognition.

Chapter 6: Working Memory in Virtual Reality

Abstract

In two studies we investigate working memory performance between classical computer-based setups and Virtual Reality (VR): a serial recall task with animal stimuli and a delayed matching task using color and location cues. Participants completed tasks using two different devices, a standard computer screen, and a VR headset. Overall results demonstrate overall consistent performance across PC and VR environments. Performance in the serial recall task was consistent with prior literature, showing better performance in smaller set sizes. Similarly, in the delayed matching task, performance also improved with a smaller set size. Additionally, performance in Location report was more accurate than color report. These findings suggest that VR technology can replicate traditional PC-based cognitive tasks effectively. This cross consistency between the task and response devices opens an outlet for experimental designs using VR.

Introduction

Human memory constantly engages with environmental surroundings. In the quest to understand memory, experimental research reduces the complexity of these surroundings, conducting experiments in controlled laboratory settings. These settings, however, vastly differ from the complex nature of our environment and lack ecological validity. Virtual Reality (VR) technology can bridge this gap, by presenting more realistic stimuli in naturalistic environments (Hoffman, 1998; Jubran et al., 2022; Wilson & Soranzo, 2015). (Jubran et al., 2022, Submitted) suggests that experimental results in VR correspond well to those using classical setups of computer screens and keyboards. In these studies, VR controllers were shown to be a reliable device for measuring reaction times in working memory (WM), flanker, and response priming tasks (Jubran et al., 2022, Submitted). A major advantage of using VR in experimental research is that it allows researchers to control and manipulate the environment in a way that is not possible in the real world. In addition, VR controllers can provide a more time-sensitive device than a traditional keyboard (Jubran et al., Submitted).

In the first study, we conducted a dual serial recall task, where participants are presented with a series of animal pictures in different colors and are asked to report both the animals and their colors in the presented order. Serial recall is a common paradigm used to investigate WM capacity (Bhatarah et al., 2008; Hockley, 2008). It investigates the ability to recreate a list of items in the order they were presented. Performance in serial recall tasks is affected by factors such as list length, item complexity, and distractions (Bhatarah et al., 2008; Gathercole & Alloway, 2008; Ranjith, 2012). We investigate the influence of atypical or “bizarre” colors on serial recall performance. Our second aim is to investigate the usefulness of presenting items in an immersive

setting, i.e. in VR compared to a classical setting using a computer screen. The second study focuses on feature binding using a delayed matching task (Rajsic & Wilson, 2014).

WM is commonly defined as a limited-capacity store that temporarily holds information accessible for various cognitive operations (A. Baddeley, 1992; Conway et al., 2005; Wheeler & Treisman, 2002). (A. D. Baddeley & Hitch, 1974) proposed a WM model based on rehearsal, which consists of three vital components: i) the phonological loop, which maintains verbal information in a phonological store by inaudibly articulating speech and encoding visual material, such as nameable pictures or written words; ii) the visuospatial sketchpad, which encodes visuospatial information by establishing and manipulating visuospatial imagery; iii) the central executive, which is the control mechanism of these two components (A. Baddeley, 1992, 2000; Buchsbaum, 2016). This model of WM however, does not specify how information in WM is linked to long-term memory (LTM; (Allen et al., 2009)). To address the model's limitations, (A. Baddeley, 2000) proposed a fourth component, i.e., the episodic buffer, that is responsible for chunking information and connecting WM, LTM, and the central executive (A. Baddeley, 2000; A. Baddeley et al., 2017). The episodic buffer facilitates the use of prior knowledge and experiences to aid current cognitive processing. It is also thought to play a role in the construction of mental representations or "mental models" of events or situations, which are used to guide behavior (A. Baddeley et al., 2017).

Controlled cognitive processes, including those of WM, share limited resources (A. Baddeley, 1992; Conway et al., 2005; Luck & Vogel, 1997; McLeod, 2009; Miller, 1956; Wheeler & Treisman, 2002), i.e., they compete for the same limited capacity (Kahneman, 1973; LACHMANN & LEEUWEN, 2008; Pashler, 1984). Research suggests that this also holds for attentional processes that moderate WM (Allen et al., 2009). The ability to chunk information using the episodic buffer allows individuals to

manage the limited capacity of WM and store more information in an organized and meaningful way (Miller, 1956). This can facilitate the processing and manipulation of information in WM, and potentially improve performance on complex cognitive tasks. Research also has shown that attentional processes can influence the functioning of the episodic buffer. For example, (Cowan, 2011) has demonstrated that attentional load can affect the quality of chunking in WM, with higher attentional load resulting in less effective chunking (Cowan, 2011). Other studies have shown that attentional processes can also influence the selection and integration of information in WM, as well as the transfer of information from WM to LTM (e.g.,Jonides et al., 2008)).

Color is one of the object features that play a special role in object recognition (Bramão et al., 2010; Redmann et al., 2014; Tanaka & Presnell, 1999; Therriault et al., 2009). (Redmann et al., 2019) observed that objects are recognized more readily when paired with a typical color rather than an atypical one (Redmann et al., 2019). For example, a tomato, which is an object associated with the color "red" has conceptual information of the color processed differently than a chair, which can have any color. Although the color feature has an impact on the recognition process, it has been found effective only in the early stages of object processing where object and feature segregation occurs (Bramão et al., 2010; Therriault et al., 2009). Contrasting results can be found from the bizarreness effect, where the presence of unusual or “bizarre” items within a sequence has been shown to improve memory performance (Nicolas & Marchal, 1998). (Morita & Kambara, 2022) investigated the color bizarreness effect in a free recall experiment (Morita & Kambara, 2022). Their results showed that bizarreness improved recollection of items, extending the bizarreness to color stimuli. It is still unclear, however, how object recognition and the bizarreness effect interact with serial recall.

Real-world objects are found to be easier to remember than simple stimuli such as colored squares, oriented lines, or novel shapes (Brady et al., 2016). (Markov et al., 2021) argue that these real-world objects are stored as holistic representations when the memory load is low (Markov et al., 2021). Moreover, elaboration of an item leads to stronger combination processes during encoding and thus, improved WM (Craik & Lockhart, 1972; Michelon et al., 2003). Utilizing different technologies such as VR can provide a way to further elaborate on three-dimensional (3D) stimuli and provide more context-relevant information. It also allows for a more realistic and immersive experience for participants, which may lead to more ergonomic, ecologically valid, and representative results (Chen & Wu, 2023). Furthermore, (Saint-Aubin & Poirier, 2000) found that serial recall performance is affected by long-term knowledge. They show that word lists perform better than lists of non-words because items that could be transferred into LTM after the rehearsal are remembered better in comparison to lesser-known items that might not have easy access to LTM.

Presenting items in a 3D virtual environment may also allow participants to better encode the spatial relationships of stimuli items, which could facilitate recall performance (Krokos et al., 2018). Previous research has also found that WM performance is better in VR conditions compared to other tools such as 2D video or text-book when a text and a 3D model of an object are used as learning material (Allcoat & von Mühlenen, 2018). Moreover, (Berki, 2018) found that images are remembered better in VR than in web page visualization (Berki, 2018; Namazi et al., 2021).

Studying serial recall and the factors influencing it in different contexts allows us to understand the processing and binding of visual stimuli. We investigated the bizarreness, or typicality effect in a serial recall task, where participants memorized the color and identity of colored animal stimuli. We manipulated the dimension of the stimuli (2D and 3D stimuli), the color of the stimuli (typically

colored and atypically colored), the set size (4-8 items), and the presentation device (PC and VR) in a 4x2x2x2 design. The stimuli were presented in a blocked design, with four variations, namely mostly typical or mostly atypical blocks, as well as 2D and 3D blocks. Stimuli were verified through a recognition survey conducted before the experiment. We then directed participants to perform a dual serial recall task, where they remembered animals, as well as the color they were presented with. Accuracy was assessed for each trial. We hypothesize that stimuli presented in "bizarre" or atypical colors will exhibit better recall precision compared to objects in their typical colors. In addition, 3D-presented stimuli in VR will exhibit better recall precision compared to 2D ones.

Study 1: Serial Recall:

Method

Participants

Thirty-two participants (16 assigned to VR and 16 to PC, 18 females, ages 20-30, mean age of 26.09 years, SD = 4.75) from the German city of Kaiserslautern took part in the study. They all had normal or corrected-to-normal vision. The study was approved by the Ethics Committee at the University of Kaiserslautern and participants gave written informed consent prior to participating. All procedures performed were in accordance with the ethical standards of the institution and with the 1964 Helsinki Declaration.

Stimuli

The stimuli used in the study consisted of ten different animal pictures and 3D assets, which were paired based on their colors to ensure an equal number of typical and atypical colors. The animals selected for use in the study were bear, monkey, chick, bee, duck, sheep, rabbit, elephant, cat, and tiger. Each pair represented one of five different colors: brown, yellow, white, gray, and orange.

The typical colors were selected based on the typical color variations of each animal, and none of these colors were included in the atypical color list. The atypical colors used in the study were pink, purple, green, blue, and red (see Figure 26). The study utilized six versions of each animal in both 2D and 3D. The HTML color codes were applied consistently to both the 2D images and 3D assets. The 2D items were presented on a billboard in equal sizes, while the 3D items were located at varying distances from the camera to ensure equal sizes. The 3D animals were obtained from the Unity Assets Store.

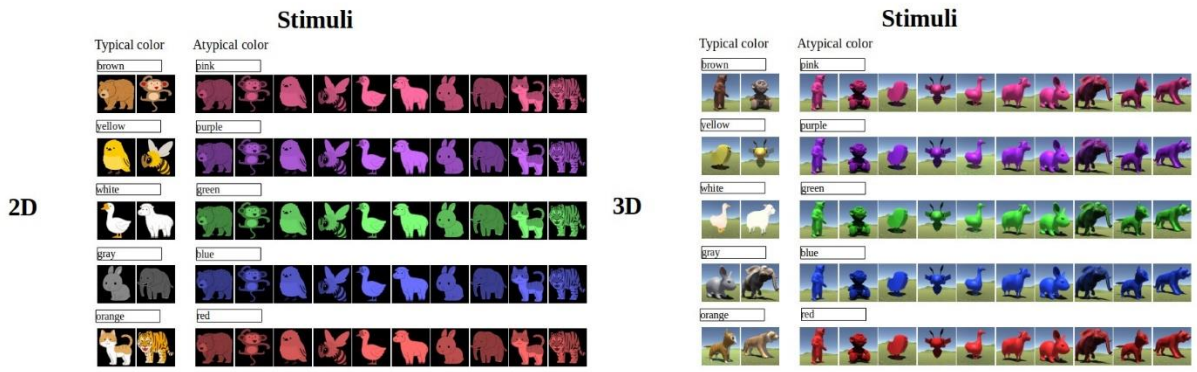


Figure 26. Experimental stimuli

Apparatus

The study was conducted in a dimly lit room at the center of cognitive science, University of Kaiserslautern, Germany. Tasks were executed on a high-end gaming laptop computer running Windows 10 Enterprise 2016 LTSB 64-Bit operating system with an Intel Core i7-8750H processor. The stimuli in the PC version were presented on a monitor (40cm x 30cm) with a screen resolution of 1920 x 1200 and a refresh rate of 90 Hz, running g-sync technology to ensure a stable framerate. Participants were seated 70 cm away from the monitor. For the VR condition, an HTC Vive pro-VR headset with 1080x1200 per eye resolution, and a refresh rate of 90Hz was used. The experiment

was programmed using Unity Software, version 2020.1.17f1. For trials in VR, response times were recorded throughout trials using VIVE Pro controllers.

Design and Procedure

Part one: Picture Rating Survey

Twenty students from the University of Kaiserslautern participated in a picture rating survey. They received course credit for their participation. The mean age of the participants was 26.2 years (SD = 4.8). The survey consisted of 240 questions, with four questions per animal picture. The survey asked to rate the pictures according to Familiarity, Image Agreement, and Visual Complexity. The survey was conducted using a standard PC under the same conditions as the later PC version of the experiment. Participants performed the survey starting with a random category of four (2D mostly typical animals, 2D mostly atypical animals, 3D mostly typical animals, and 3D mostly atypical animals) in a balanced manner to avoid any biases in the results. The first question in each survey asked for the stimulus in its typical color, followed by two questions asking for the stimulus in two randomly selected atypical colors.

Part two: Serial Recall Experiment

Both PC and VR versions included equivalent sequences of events as follows; Each trial started with a truck moving from left to right across the participant's vision. The first randomly chosen stimulus appeared as the truck passed and was presented for 3 seconds. The truck then repeated its movement from left to right, causing the stimuli to disappear as it arrived at their location and covered them and the second one to appear, and so on. (see Figure 27 for a visual representation). After the final stimulus was presented, a question panel with ten animal stimuli was presented to the participant (see Figure 28). The participants are probed then to reproduce the sequence of items presented, first by selecting the animal, and then by selecting the color,

using either a laser pointer in the VR environment or a mouse in the Desktop environment. Both the accuracy and reaction times were recorded during both color and animal selection. As was done in the survey, participants performed the experiment starting with a random condition out of four conditions (2D mostly typical, 2D mostly atypical, 3D mostly typical, and 3D mostly atypical) in a balanced manner to avoid any biases in the results. In the most typical blocks, 70% of the items were presented in typical colors and 30% in atypical colors. The opposite was true for the atypical blocks, i.e. 30% of the items were presented in typical colors, while 70% were presented in the atypical ones. Each participant completed two sessions. Each session consisted of a practice block followed by six blocks of 20 trials each, for a total of 12 blocks with 20 trials each, half using 2D images and half using 3D items. Participants were assigned to either VR or Desktop conditions and completed the experiment using the appropriate set-up. The session order was counterbalanced, and they were performed on different days, approximately 1 week apart on average.

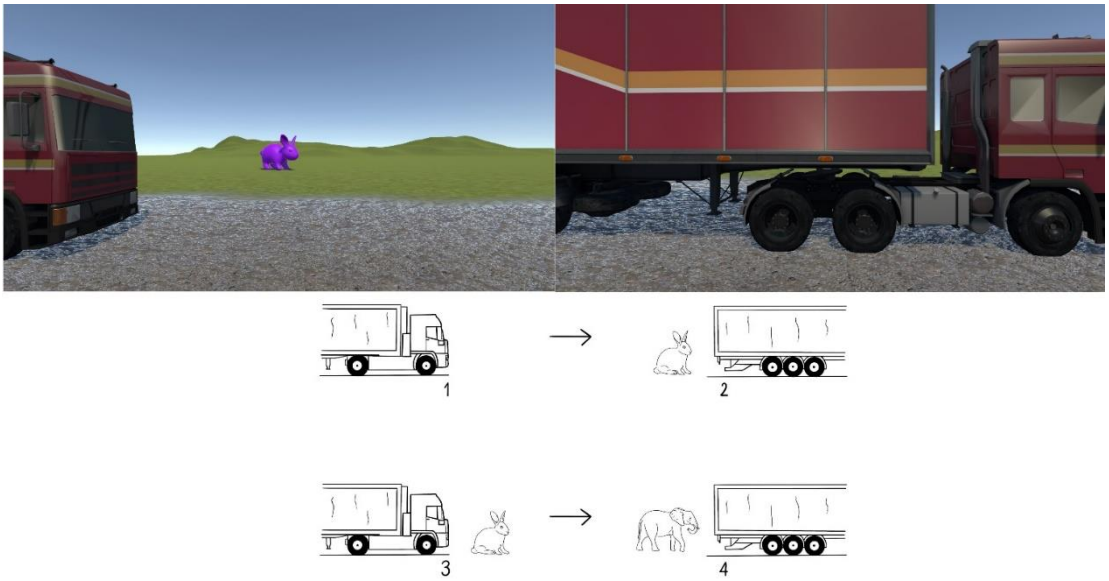


Figure 27. The Process of One Trial

Figure 28. The question panel

Analysis

Accuracy was calculated using the proportion of animals and colors remembered in their correct serial position for each trial. Animal selection accuracy was first analyzed in the classical version and compared it to animal selection Accuracy in VR. Afterwards, color selection Accuracy was analyzed in the classical version and compared it to the color selection Accuracy in VR. A total of 7680 trials were available for analysis, 3840 for PC and 3840 for VR. We performed two four-way ANOVA, for each dependent variable (animal selection Accuracy and color selection Accuracy), with the within subject factors Set size (4-8), Typicality (Typical vs. Atypical) and Animal dimension (2D vs 3D). Our between subject factor was device (PC vs VR).

Results

Survey results

In the survey, participants were probed to name animals and rate them in terms of familiarity, visual complexity, and image agreement. The analysis of results indicated that familiarity scores differed significantly across colors, $F(4, 1187) = 85.807$, $p < .001$, and dimensions, $F(1,1187) = 18.257$, $p < .001$. Familiarity scores were significantly higher for typically colored animals in comparison to atypically colored ones (M = 4.46, SD = 0.97). In addition to this, 2D images showed higher familiarity scores than 3D images (M = 3.53, SD = 1.31). Dimension had an interaction with familiarity, $F(4,1187) = 4.174$, $p = 0.00$. Post-hoc comparisons showed that typically colored images scored higher than atypically colored images in both 2D (M = 4.51, SD = 0.96) and 3D (M = 4.41, SD = 0.97) conditions. We have analyzed the scores across different animals and colors used in the study. Bear showed the greatest typicality (M = 3.75, SD = 1.36) and image agreement (M = 3.48, SD = 1.31) while monkey showed the lowest typicality (M = 3.24, SD = 1.46) and image agreement

(M = 2.94, SD = 1.47). Typicality has been found higher for blue (M = 3.35, SD = 1.27) and green (M = 3.35, SD = 1.28) than pink (M = 2.88, SD = 1.49) and purple (M = 2.79, SD = 1.51).

Serial recall experiment results

Animal selection accuracy

Selection accuracy results are shown in Figure 29. The repeated-measures ANOVA revealed no main effect of Device [$F(1, 30) = .272$, $p = .606$]. We obtained a main effect of Typicality [$F(1, 30) = 1.354$, $p < .001$, $\eta^2 = .059$], and for List Length [$F(4, 120) = 217.121$, $p < .001$, $\eta^2 = .330$] after correction. A main effect of Dimension was not found [$F(1, 30) = 1.525$, $p = .227$]. We obtained an interaction effect between Typicality and List Length [$F(4, 120) = 2.984$, $p = .022$, $\eta^2 = .003$]. There were no further significant interactions.

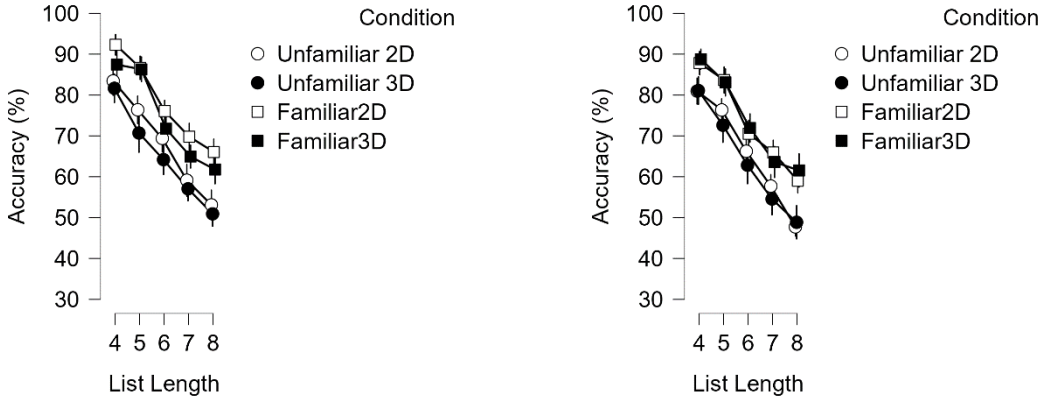


Figure 29. Animal selection accuracy results

Color selection accuracy

Color selection accuracy results are shown in Figure 30. The repeated-measures ANOVA revealed no main effect of Device [$F(1, 30) = .285$, $p = .598$]. We obtained a main effect of Typicality [$F(1, 30) = 437.044$, $p < .001$, $\eta^2 = .472$], and for List Length [$F(4, 120) = 270.783$, $p < .001$, $\eta^2 = .147$] after

correction. A main effect of Dimension was not found [$F(1, 30) = 0.\text{-}40$, $p = .843$]. We obtained an interaction effect between Typicality and List Length [$F(4, 120) = 26.074$, $p < .001$, $\eta^2 = .011$]. There were no further significant interactions.

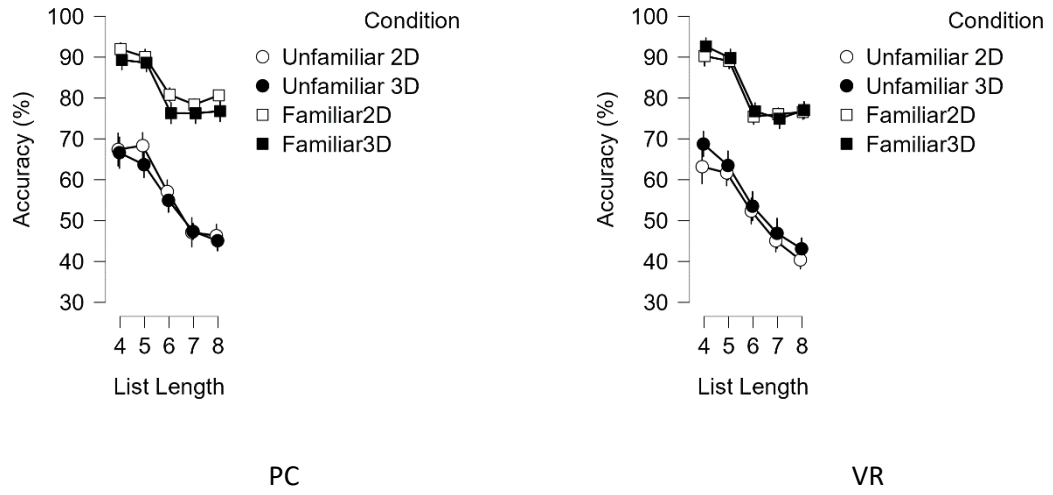


Figure 30. Color selection accuracy results

Study 2: Delayed matching task

Method

Participants

Participants were all students from the RPTU Kaiserslautern campus. There were 7 participants (4 female) between the ages of 18 and 31. We did not collect precise age, as participants were only asked to select which age bracket they fell under. We did this to preserve the privacy of participants' personal data since the age bracket is where the relevant information for our study is found. Participants received participation credit for their time. They reported no speech, neurological, or perceptual impairments. They all gave their written consent before participation.

The experiment was approved by the ethics committee of the university in accordance with the ethical standards of the institution and with the 1964 Helsinki Declaration.

Materials

There was a short demographic questionnaire asking age, gender, handedness, and for a list of any perceptual impairments that was filled out at the end of the second experiment session. Information gathered from this questionnaire was used to exclude any left-handed participants or those with any perceptual impairments.

Because one of the goals of the present study was to reproduce the results from Rajsic & Wilson (2014) we designed the stimuli outlined in their original study. However, because another one of the current experiment's goals is to validate the task as a tool for VR working memory studies, the experiment aimed to reproduce Rajsic and Wilson´s (2014) experiment in both PC and VR environments. Due to certain requirements for coding the VR condition, some characteristics of the current experiment's stimuli differ from that of the original study. However, the stimuli used in the present experiment's PC and VR conditions remain the same. This way, we can ensure that the results from the PC and VR conditions are comparable.

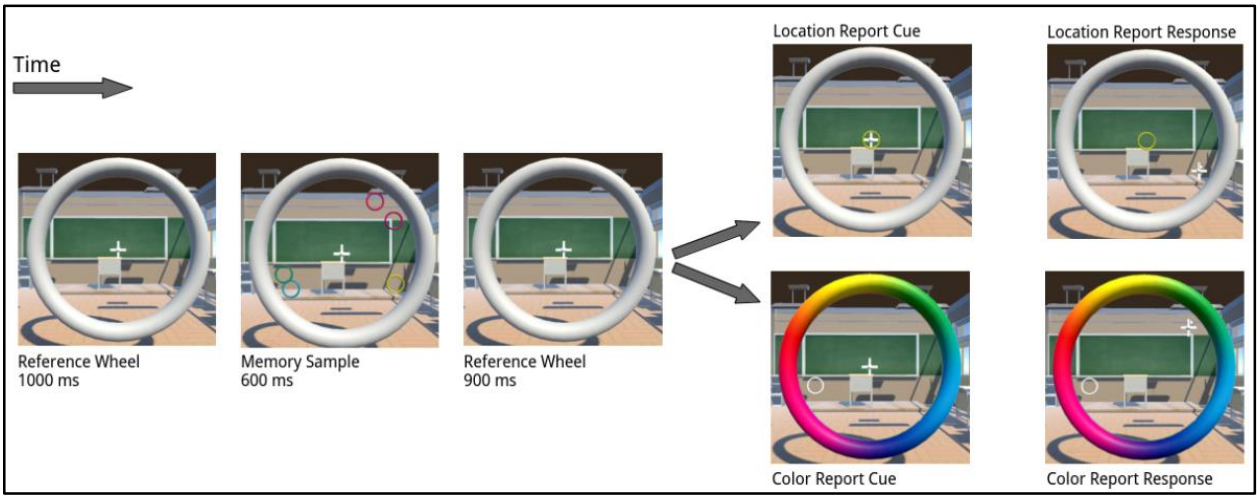


Figure 31.Representation of Experimental Display, Stimuli, and Environment

Figure 31 shows experimental stimuli and procedure. The experiment consisted of multiple steps, i.e., memory sample, retention interval, report stage, and feedback steps. There were 512 trials, each separated by a 1000 ms gap for participants to refocus. Among these trials, 12 were practice trials, optional for participants to repeat. The memory sample displays small rings in various colors from a pool of 360 colors, shown in sets of 2 or 5, with random colors for each sample. While the original study used CIE lab* color space to avoid similarity, the current one uses RGB values due to time constraints. Chosen colors must differ by a minimum RGB score of 10 to prevent color repetition. The rings appear within a circle with a radius of 0.73 m, compatible with the Unity engine and close to the original study's visual angle. The rings appear inside a reference wheel for visual consistency, except during the report stage, where a color wheel replaces it. After a 600 ms display, a 900 ms retention interval follows before participant’s report either color or location. Color report conditions display a neutral cue in the ring's target location, while location report conditions show a color cue in the reference wheel's center. Participants then respond by dragging a marker on the respective wheel. Feedback follows, displaying a black dot at the correct response location for 1000

ms. Unlike the original blank screen display used by the original study, the current experiment situates stimuli in a classroom environment for VR compatibility and consistency across versions. This choice aims to reduce disorientation and enhance VR realism, matching the lab environment. Additionally, the current experiment employs three-dimensional stimuli with highlights and shadows to improve realism.

Apparatus

The setup for this experiment was identical to that of experiment 1. In the PC condition, participants were seated 50 cm away from the monitor to match the visual angle between the PC and VR conditions. The reference wheel subtended 32° of visual angle. Participants responded via a mouse set to the maximum cursor speed offered on Windows 10 system. This speed balanced user comfort with the ability to enter quick accurate responses. The trajectory of the mouse was tracked during the report stage to provide data for future trajectory tracking analysis.

In the VR condition, participants provided their responses using the Vive controller held in the right hand. Only the right controller was tracked when recording participant responses to keep the mouse and controller inputs as similar as possible. The participant was placed in front of the reference wheel subtending 28° of visual angle during the VR condition. The experiment in both conditions was conducted while the participants were seated.

Procedure

After providing their consent, participants were given visual instructions about the task. The instructions exemplify how the stimuli are presented and explain how participants enter their response during the location report and color report sections. Participants were instructed that for each trial they would first be shown a memory sample consisting of either 2 or 5 differently

colored rings. They were informed that the rings appear inside the reference wheel in randomly selected locations and colors, and that they would enter their responses for the color or location report after a short retention interval where no stimuli are displayed.

The trials were separated into 8 blocks of 64 trials each, with the first block containing a practice section with 12 trials that could be repeated if the participant requested. There were small disparities in the number of trials participants did, due to repetition of practice blocks or unexpected program crashes.

Experimental Design

The experiment was within-subjects and compared report type (color report vs location report), memory sample set size (two items vs five items), and device type (PC vs VR) following a 2 x 2 x 2 design. The color report and location report were presented in a random order. This prevented participants from adopting strategies to avoid using cued recall based on the expected type of report. As in the original experiment, we controlled the selection of colors and locations for each memory sample so that there were no repeating items within a set. All colors in a set had a minimum difference of 10 points between their RGB values. And as in the original study, there was a space of at least 2° between each ring in each set. This allows rings to overlap without fully blocking one another.

Analysis

The response error for each trial was calculated by determining the distance between the correct target location and the participant's response location. To discover how memory for each task was affected by the set size and the device type, a 2 (Device: PC vs VR) x 2 (Set Size: two vs five) x 2

(Task: Color report vs Location report) repeated-measures ANOVA was conducted on the derived error variable.

Results

Repeated Measures ANOVA

The results of the repeated measures ANOVA are presented in Figure 32. The ANOVA revealed a main effect of Device [$F(1, 7) = 74.972$, $p < .001$, $\eta^2 = .066$], A main effect of Set size [$F(1, 7) = 77.251$, $p < .001$, $\eta^2 = .397$], and a main effect of Task [$F(1, 7) = 131.235$, $p < .001$, $\eta^2 = .413$]. The analysis also revealed a significant interaction between device and Task [$F(1, 7) = 15.977$, $p = .005$, $\eta^2 = .008$], and between Set size and Task [$F(1, 7) = 25.329$, $p = .002$, $\eta^2 = .029$]. No further significant interactions were found.

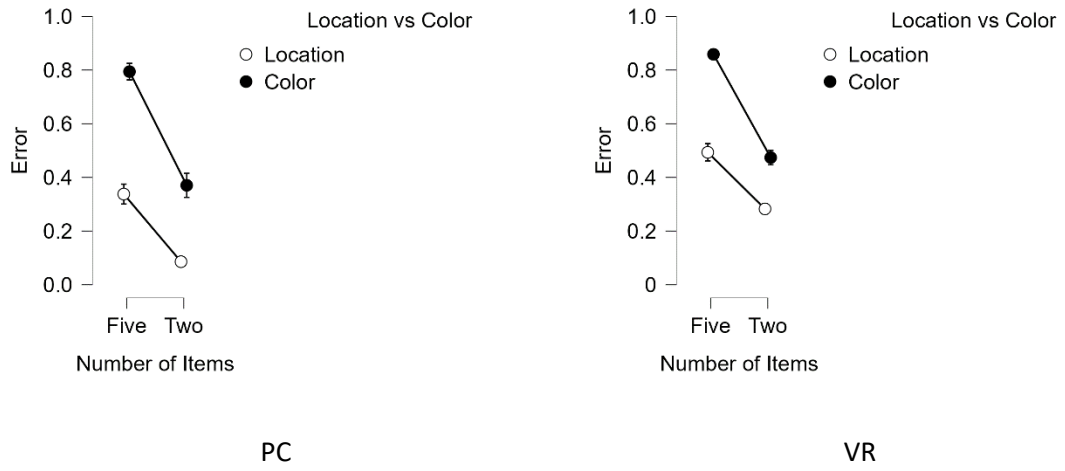


Figure 32. Color vs Location error results

General Discussion

We explored the effect of the presentation medium and the “bizarre” object effect on serial recall of visual items. The task required serial order reconstruction of typically and atypically colored animal stimuli presented in a classical setting and a VR setting. The results suggest that color does indeed affect the retrieval of both tasks, which is consistent with other studies showing how the WM system can benefit from knowledge in LTM, building stronger representations for typically colored stimuli (Brady & Störmer, 2022). The results also suggest that color information influences not only the recognition of objects but also binding and retrieval for serial recall.

Presenting animals in typical colors improved performance for both animal and color recollection, despite color being task irrelevant for the animal recollection task. The effect of typicality is much more pronounced in the color recollection task indicating increased difficulty. This increased difficulty interacts with the animal memorization task, causing worse performance. This can be due to the additional load of performing a dual task that is less prominent in the typical blocks, since participants only must remember the identity of the animal in mostly typical blocks. These results do not necessarily contrast the "bizarre" object effect. Instead, they indicate that having too many "bizarre" objects in the memory set reduces performance, due to the added difficulty of remembering both the animal identity and color identity for atypical colored animals. This shows that frequency must be considered when investigating the effect in serial recall, controlling for the number of items in the list are with “bizarre” colors.

(Brady et al., 2016) demonstrated that detailed visualizations lead to better memory performance in 3D images due to the elaboration of visionary information. Their work suggested that 3D images are closer to reality than 2D images. We predicted that 3D items would be remembered better than

2D items, despite dimensions being task irrelevant. In addition, we expected that presentation of items in VR would improve memory performance, due to the additional context, as well as increased ecological validity. Our results, however, have not shown any significant differences between 2D and 3D presentation, as well as no effects of computer and VR-based mediums.

This study is not without its limitations. Assets used in this study were limited to the free category therefore, the image quality was limited despite our consistent and valid survey results. We had 2D animals with a black background that might influence visual perception. Setting the experiment background with grass and sky could have played a role as a hint of blue and green colors to be more accurately remembered as one of the participants indicated that he got help from the background colors. Finally, the native language of the participants was not controlled. In addition to this, we did not match the animals' or colors' names in terms of the number of letters in English. Overall, future works should consider providing better visualizations with fewer distractors by taking the background set into account. A future direction for such an experiment should also be the parametric variation of the frequency of typicality. To investigate the "sweet spot" of atypical item recall advantage.

Importantly, the results show that classical tasks can be transferred well into VR setups. It is worthwhile to further investigate the effects of VR with designs that utilize the 3D visualizations. Consistency in the results between VR and computer-based indicates the viability of VR as a medium for psychological research. The results did not indicate any added difficulty due to the use of VR, which has implications for the field of cognitive ergonomics, where the assumptions of added load due to VR have not been verified. In addition, it is important to continue the development of VR hardware to overcome the limitations of motion sickness and visual fatigue (Chen & Wu, 2023).

In addition to hardware, (Chen & Wu, 2023) suggests that human-centered software designs for VR should be a priority of researchers. Such improvements to hardware and software are necessary to improve the ecologic and ergonomic validity of VR studies.

Chapter 7: Discussion

The first chapter of this work started with a question: Can we measure cognitive processing in an ecologically valid setting? The findings from the outlined experiments provide a clear answer: yes, if we preserve experimental control. This work sets out to demonstrate how this is possible to use VR for the creation of close to real-life scenarios for experimental use. The other goal of this work was to enhance our understanding of continuous cognitive processing using distributional analysis and hand movement trajectories.

Trajectories enabled a fine-grained measure of human cognition in an ecologically valid setting. We invested extensive efforts into improving the ecological validity of traditional experimentation while maintaining control over experimental variables such as luminance, visual angles, and response devices.

7.1 General discussion

In the first study (chapter 4), we replicated classical priming and flanker tasks in VR. Mental chronometry measures, i.e., reaction times and error rates, were collected using a VR controller in both a classical and a VR setup. Proper control of the response device, experimental display, and visual angles proved effective, yielding consistent results in both devices. Response distributions further demonstrated this consistency, giving us a first indication of how to use this technology.

The second study (chapter 5) demonstrated the possibilities of VR by replicating the classical N-back task in VR using hand movements. The focus of this study was not a simple replication of WM effects, but an exploration of the experimental power of hand movements. For cross-validation, we compared the classical mental chronometry measures with equally singular measures derived from the movement trajectories tracked in VR. Overall, the results showed consistency between

the classical and VR setups, replicating classical findings (Meule, 2017). Distributional analyses revealed evidence for early bias processes, typically absent in classical analyses of the task. The time-course of these processes was further detailed by StSA, allowing us to observe how early hand movements were influenced by bias, whereas late hand movements indicated decision making, mirroring known speed-accuracy tradeoff curves. StSA is unique, in that it was specifically designed for analyzing continuous responses and can overcome violations of assumptions that are necessary for ANOVA.

The third and fourth studies (chapter 6) aimed to understand WM feature binding, comparing classical and VR setups. The third study focused on the "bizarre" object effect, whereas study four focused on the binding of color and location to color stimuli. Both experiments showed the overall consistency in results between both response devices, replicating classical WM span results. In addition, study four demonstrated the utility of continuous error monitoring in VR using hand movement trajectories again. The fourth study found significant differences between the devices which are worth investigating further. These differences, however, can be attributed to physical differences in error calculation between the two devices.

Mental chronometry has been the primary tool for psychologists since the 1900s. Inferences using discrete methods such as reaction times and accuracy have answered many questions about human cognition and decision-making. If we are to drive the wheel forward in understanding human cogitation, it is necessary to improve both our experimental setups, and response collection. In this sense, VR and continuous hand movements offer a natural step forward, providing model free understanding of human behavior. This work lays the foundations for a future in experimental psychology, that can measure the fine-grained dynamics of human cognition in an ecologically valid setting.

As VR-based experiments provide more complex stimulation and responses, we might expect additional cognitive load (Frederiksen et al., 2020; Juliano et al., 2022; Sweller, 1988). However, cross-consistency in the results between the classical and VR versions of the studies indicate similar task demands. But as discussed in chapter five, VR technology should move past the arbitrary nature of Cognitive load. Instead, using novel methods like StSA, we can quantify specific mental operations, such as attention, awareness, and response conflict. Using direct action to quantify mental operations.

Continuous physiological data, such as EEG and pupillometry, were long used as a key to understand cognition. Despite their utility, they are still indirect indicators of human cognition. In addition, they are hard to comprehend without complex models and interpretations. Hand movements, on the other side, are a direct indicator of behavior. Employing StSA offers a framework for comprehending continuous hand movements, directly associating them with attention, awareness, and recognition. Moreover, VR emerges as a valuable instrument for experimental psychologists, moving beyond the laboratory, with exposure therapy and embodiment experiments for example.

Modern experimental psychology contends with emerging challenges, including experimental power, replicability, and ecological validity of experiments. These challenges are interconnected and demand a holistic approach to tackle. Projects such as the Open Science Collaboration (Open Science Collaboration, 2015) demonstrate that some effects in psychology are generally hard to replicate, creating what has been known as the “replication crisis”. The reasons behind this “crisis” have been debated in the past 10 years. For example, Francis argues that the publication bias has a great influence on our perceived effects, where scientists are racing to publish any significant findings (Francis, 2012). On the other hand, Ellis and colleagues argue that the lack of replication is

only an artifact of the bad scientific practices and lack of experimental power for example (Ellis, 2022).

VR experiments hold much promise to address these challenges. As shown by the StSA method, movement trajectories potentially offer great precision, because the spatiotemporally resolved character of tracking data allows the extraction of multiple data points from each trial. Trajectory data can be analyzed as series, transitions could be identified, and segments assigned to distinct stages, such as movement initiation, subsequent redirections, and completion of the trajectory (Jubran et al., Submitted). VR expands the scope of classical experiments by accommodating additional visual features.

7.2 Limitations and ethical considerations

The employment of VR for psychological experiments comes with certain limitations, involving comparability with classical presentation devices, and response collection devices. These are not necessarily disadvantages but raise questions about the validity and reliability of mental chronometry measures. For example, the priming paradigm and similar paradigms are very sensitive in terms of timing (F. Schmidt et al., 2011), where the effects can vary starkly if a stimulus is presented for 50ms instead of 37. In most monitors suitable for experiments, the frame duration and frame rates are extensively controlled. This can be difficult in VR, due to the sometimes-unstable hardware.

As shown by chapter 6, working memory effects can be replicated well in VR. However, quantitative differences were found in the VR condition for the delayed matching task, which could be due to differences in error calculation and should be investigated further.

VR technology also prompts some ethical concerns. Slater and colleagues argue that using VR to recreate realistic scenes, or enhance certain aspects about the surrounding world or the body of participants can lead to participants preferring the virtual world over the real world for example (Slater et al., 2020). They discuss the golden rule of reciprocity (“treat others as you would have them treat you”), and how that should guide our moral compass when designing experiences in the virtual realm.

It is important to step into VR technology incrementally, as to preserve the benefits of controlled experimentation. Comparing performance in minimal VR setups can be the gateway to introducing more complex and intricate designs, promising more ecological validity.

7.3 Personal growth

I came to this field from an engineering background. In engineering, the focus is always set on new innovations, new technologies, new methods, and analyses. It comes then as no surprise that the engineer is discussing new methods and technologies. However, this current discussion does not come just from an engineer that is excited about a new piece of technology, but rather from a now trained experimental psychologist that is trying to operationalize this technology within existing frameworks and theories.

Getting to this stage required a lot of growth, and a personal paradigm shift. One that had me asking questions finally like an experimental psychologist, not only the question of “can we, do it?” to “why should we do it?". I am thankful to all my mentors that led me to this stage, and to my colleagues that reinforced this thinking in me for the past eight years.

7.4 Parting thoughts

Looking at experimental psychology as a field, there is always a sense of minimalism in experimental designs which carries over to conservativism in data analysis and interpretations. To move the field further, we need to be more open to unexpected findings. Any experimental results are valid if we follow the methodological rigor of experimentation. Experimental psychology also needs to move away from disjointed theories to the operationalization of current ones.

This work highlights the importance of empirical research and exploratory data analysis, in varied yet experimentally controlled settings. It is always easy to introduce new theories in science; the harder task is fitting these theories within a generalized framework. The introduction of new theories without operationalization leads to two problems: (1) Many disjointed “mini” or “parallel” theories, and (2) the introduction of false theories or elements into science. It takes a few papers with convincing evidence to introduce a theory, it takes much more research and evidence to falsify it.

Bennett and colleagues demonstrated how easy it is to abuse new technology within the realm of science (Bennett et al., 2009). If you can find significant results using a dead salmon, you can find any effect if you go fishing! Thus, analyses and measures should be consistent and make sense. Working with human behavior rather than other indirect measures of cognition offers more insight into the human mind, and offers a better operationalization of current theories. Goal-directed human behavior, such as hand movements can be understood without the need for complex models and multiple correlations and statistical tests. Virtual Reality (VR) here comes as a tool for experimental control, producing valid measures.

VR moves the needle of science further, not by introducing grand theories but a new medium of conducting science. It also allows for between-lab replication using standardized equipment and environments. In addition, it allows for a finer understanding of response dynamics through continuous monitoring. In this sense, VR should not be used for VR sake, but rather as a tool. Psychologists need to start moving away from the comfort of the lab into more exploratory realms, because some answers are not found under the lamppost.

References

Abdillah, D. F., Basuki, A., & Harsono, T. (2020). Visually Realistic Rain Modeling Optimization for VR Application. *2020 International Electronics Symposium (IES)*, 688–694. https://doi.org/10.1109/IES50839.2020.9231591

Aguasvivas, J., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., & Duñabeitia, J. A. (2020). How do Spanish speakers read words? Insights from a crowdsourced lexical decision megastudy. *Behavior Research Methods*, *52*(5), 1867–1882. https://doi.org/10.3758/s13428-020-01357-9

Allcoat, D., & von Mühlenen, A. (2018). Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology*, *26*.

Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2009). Cross-modal binding and working memory. *Visual Cognition*, *17*(1–2), 83–102.

Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology*, *13*, 61. https://doi.org/10.2307/270718

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

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

Baddeley, A., Allen, R., & Hitch, G. (2017). The role of the episodic buffer. *Exploring Working Memory: Selected Works of Alan Baddeley. Routledge, London*.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Elsevier.

Baker, D. H., Vilidaitė, G., Lygo, F., Smith, A., Flack, T. R., Gouws, A., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295–314. https://doi.org/10.1037/met0000337

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 340–357. https://doi.org/10.1037/0096-1523.10.3.340

Banks, M. S., Hoffman, D. M., Kim, J., & Wetzstein, G. (2016). 3D displays. *Annual Review of Vision Science*, *2*(1), 397–435. https://doi.org/10.1146/annurev-vision-082114-035800

Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social Cognitive and Affective Neuroscience*, *4*(4), 417–422. https://doi.org/10.1093/scan/nsp053

Berki, B. (2018). Better memory performance for images in MaxWhere 3D VR space than in website. *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 000281–000284.

Bhatarah, P., Ward, G., & Tan, L. (2008). Examining the relationship between free recall and immediate serial recall: The serial nature of recall and the effect of test expectancy. *Memory & Cognition*, *36*, 20–34.

Biafora, M., & Schmidt, T. (2020). Induced dissociations: Opposite time courses of priming and masking induced by custom-made mask-contrast functions. *Attention, Perception, & Psychophysics*, *82*(3), 1333–1354. https://doi.org/10.3758/s13414-019-01822-4

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753), 435. https://doi.org/10.1038/d41586-019-01307-2

Brady, T. F., & Störmer, V. S. (2022). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 942.

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, *113*(27), 7459–7464.

Bramão, I., Inácio, F., Faísca, L., Reis, A., & Petersson, K. M. (2010). The Influence of Color Information on the Recognition of Color Diagnostic and Noncolor Diagnostic Objects. *The Journal of General Psychology*, *138*(1), 49–65. https://doi.org/10.1080/00221309.2010.533718

Brenner, E., & Smeets, J. B. J. (2005). Intercepting moving targets: Why the hand's path depends on the target's velocity. *Proceedings of SPIE*. https://doi.org/10.1117/12.610849

Buchsbaum, B. R. (2016). Working memory and language. In *Neurobiology of language* (pp. 863–875). Elsevier.

Camalier, C. R., Gotler, A., Murthy, A., Thompson, K. G., Logan, G. D., Palmeri, T. J., & Schall, J. D. (2007). Dynamics of saccade target selection: Race model analysis of double step and search step saccade production in human and macaque. *Vision Research*, *47*(16), 2187–2211. https://doi.org/10.1016/j.visres.2007.04.021

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2015). Review and Summary of Research on the Embodied Effects of Expansive (vs. Contractive) Nonverbal Displays. *Psychological Science*, *26*(5), 657–663. https://doi.org/10.1177/0956797614566855

Chen, Y., & Wu, Z. (2023). A review on ergonomics evaluations of virtual reality. *Work*, *74*(3), 831–841. https://doi.org/10.3233/WOR-205232

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/bf03196772

Cowan, N. (2011). The focus of attention as observed in visual working memory tasks: Making sense of competing claims. *Neuropsychologia*, *49*(6), 1401–1406.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Cressman, E. K., Franks, I. M., Enns, J. T., & Chua, R. (2007). On-line control of pointing is modified by unseen visual shapes. *Consciousness and Cognition*, *16*(2), 265–275. https://doi.org/10.1016/j.concog.2006.06.003

David, F. N., & Tukey, J. W. (1977). Exploratory data analysis. *Biometrics*, *33*(4), 768. https://doi.org/10.2307/2529486

Day, B. L., & Lyon, I. N. (2000). Voluntary modification of automatic arm movements evoked by motion of a visual target. *Experimental Brain Research*, *130*(2), 159–168. https://doi.org/10.1007/s002219900218

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431. https://doi.org/10.1016/0001-6918(69)90065-1

Draper, J. V., Kaber, D. B., & Usher, J. M. (1999). Speculations on the value of telepresence. *Cyberpsychology & Behavior*, *2*(4), 349–362. https://doi.org/10.1089/cpb.1999.2.349

Eisenhart, C. (1963). Realistic evaluation of the precision and accuracy of instrument calibration systems. *Journal of Research of the National Bureau of Standards, Section C: Engineering and Instrumentation*, *67C*(2), 161. https://doi.org/10.6028/jres.067c.015

Elliott, D., Helsen, W., & Chua, R. (2001). A century later: Woodworth's (1899) two-component model of goal-directed aiming. *Psychological Bulletin*, *127*(3), 342–357. https://doi.org/10.1037/0033-2909.127.3.342

Ellis, R. J. (2022). Questionable Research Practices, Low Statistical Power, and Other Obstacles to Replicability: Why Preclinical Neuroscience Research Would Benefit from Registered Reports. *Eneuro*, *9*(4), ENEURO.0017-22.2022. https://doi.org/10.1523/ENEURO.0017-22.2022

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception & Psychophysics*, *25*(4), 249–263. https://doi.org/10.3758/BF03198804

Fehrer, E., & Raab, D. (1962). Reaction time to stimuli masked by metacontrast. *Journal of Experimental Psychology*, *63*(2), 143–147. https://doi.org/10.1037/h0040795

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991. https://doi.org/10.3758/s13423-012-0322-y

Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., & Andersen, S. A. W. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: A randomized trial. *Surgical Endoscopy*, *34*(3), 1244–1252. https://doi.org/10.1007/s00464-019-06887-8

Friston, K. J., Price, C. J., Fletcher, P., Moore, C., Frackowiak, R. S. J., & Dolan, R. J. (1996). The Trouble with Cognitive Subtraction. *NeuroImage*, *4*(2), 97–104. https://doi.org/10.1006/nimg.1996.0033

Gathercole, S., & Alloway, T. P. (2008). *Working memory and learning: A practical guide for teachers*. Sage.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112. https://doi.org/10.1007/bf02289823

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 6. https://doi.org/10.1167/3.1.6

Helmholtz, H. V. (1948). On the rate of transmission of the nerve impulse, 1850. In W. Dennis (Ed.), *Readings in the history of psychology.* (pp. 197–198). Appleton-Century-Crofts. https://doi.org/10.1037/11304-024

Hockley, W. E. (2008). Memory Search: A Matter of Time. In *Learning and Memory: A Comprehensive Reference* (pp. 417–444). Elsevier. https://doi.org/10.1016/b978-012370509-9.00157-1

Hoffman, H. G. (1998). Virtual Reality: A new tool for interdisciplinary psychology research. *Cyberpsychology & Behavior*, *1*(2), 195–200. https://doi.org/10.1089/cpb.1998.1.195

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology*, *11*, 721. https://doi.org/10.3389/fpsyg.2020.00721

Hommel, B. (1997). Toward an action-concept model of stimulus-response compatibility. In *Elsevier eBooks*. https://doi.org/10.1016/s0166-4115(97)80041-6

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annu. Rev. Psychol.*, *59*, 193–224.

Jubran, O. F., Rocabado, F., Muntini, L., Duñabeitia, J. A., & Lachmann, T. (2022). Reproducing Classical Priming, Flanker, and Lexical Decision Tasks in VR. *Proceedings of the 33rd European Conference on Cognitive Ergonomics*. https://doi.org/10.1145/3552327.3552362

Jubran, O. F., Wolkersdorfer, M. P., Eyemann, V., Burkard, N., Czernochowski, D., Herrlich, M., Leeuwen, C. van, & Lachmann, T. (Submitted). *Spatiotemporal Survival Analysis: Elucidating Working Memory Estimates in Virtual Reality.* [Submitted].

Juliano, J. M., Schweighofer, N., & Liew, S.-L. (2022). Increased cognitive load in immersive virtual reality during visuomotor adaptation is associated with decreased long-term retention and

context transfer. *Journal of NeuroEngineering and Rehabilitation*, *19*(1), 106. https://doi.org/10.1186/s12984-022-01084-6

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *33*(3), 615–622. https://doi.org/10.1037/0278-7393.33.3.615

Karşılar, H., Simen, P., Papadakis, S., & Balcı, F. (2014). Speed accuracy trade-off under response deadlines. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00248

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358. https://doi.org/10.1037/h0043688

Klotz, W., & Neumann, O. (1999). Motor activation without conscious discrimination in metacontrast masking. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 976–992. https://doi.org/10.1037/0096-1523.25.4.976

Klotz, W., & Wolff, P. (1995). The effect of a masked stimulus on the response to the masking stimulus. *Psychological Research*, *58*(2), 92–101. https://doi.org/10.1007/BF00571098

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility–A model and taxonomy. *Psychological Review*, *97*(2), 253–270. https://doi.org/10.1037/0033-295x.97.2.253

Krokos, E., Plaisant, C., & Varshney, A. (2018). Virtual memory palaces: Immersion aids recall. *Virtual Reality*, *23*(1), 1–15. https://doi.org/10.1007/s10055-018-0346-3

LACHMANN, T., & LEEUWEN, C. V. (2008). Goodness is central: Task invariance of perceptual organization in a dual-task setting. *Japanese Psychological Research*, *50*(4), 193–203. https://doi.org/10.1111/j.1468-5884.2008.00375.x

Lachmann, T., & Van Leeuwen, C. (2004). Negative congruence effects in letter and pseudo-letter recognition: The role of similarity and response conflict. *Cognitive Processing*, *5*(4), 239–248. https://doi.org/10.1007/s10339-004-0032-0

Lamichhane, B., Westbrook, A., Cole, M. W., & Braver, T. S. (2020). Exploring brain-behavior relationships in the N-back task. *NeuroImage*, *212*, 116683. https://doi.org/10.1016/j.neuroimage.2020.116683

Larsen, R. J., Mercer, K. A., & Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion*, *6*(1), 62–72. https://doi.org/10.1037/1528-3542.6.1.62

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Markov, Y. A., Utochkin, I. S., & Brady, T. F. (2021). Real-world objects are not stored in holistic representations in visual working memory. *Journal of Vision*, *21*(3), 18–18.

McLeod, S. (2009). *Short Term Memory| Simply Psychology*.

Meule, A. (2017). Reporting and Interpreting Working Memory Performance in n-back Tasks. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00352

Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology*, *26*(1–3), 3–67. https://doi.org/10.1016/0301-0511(88)90013-0

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. https://doi.org/10.1037/h0031564

Meyer, T., Kim, A. D., Spivey, M., & Yoshimi, J. (2023). Mouse tracking performance: A new approach to analyzing continuous mouse tracking data. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02210-5

Michelon, P., Snyder, A. Z., Buckner, R. L., McAvoy, M., & Zacks, J. M. (2003). Neural correlates of incongruous visual information: An event-related fMRI study. *Neuroimage*, *19*(4), 1612–1626.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Morita, A., & Kambara, T. (2022). Color bizarreness effects in object memory: Evidence from a recall test and eye tracking. *Color Research & Application*, *47*(1), 55–64. https://doi.org/10.1002/col.22697

Murphy, K. R., Myors, B., Murphy, K., & Wolach, A. (2004). *Statistical Power analysis* (2nd ed.). Routledge.

Namazi, H., Babini, M. H., Kuca, K., & Krejcar, O. (2021). Information and memory-based analysis for decoding of the human learning between normal and virtual reality (VR) conditions. *Fractals*, *29*(03), 2150163.

Nicolas, S., & Marchal, A. (1998). Implicit memory, explicit memory and the picture bizarreness effect. *Acta Psychologica*, *99*(1), 43–58. https://doi.org/10.1016/S0377-2217(98)00002-X

Niu, H., Van Leeuwen, C., Hao, J., Wang, G., & Lachmann, T. (2022). Multimodal Natural Human–Computer Interfaces for Computer-Aided Design: A Review paper. *Applied Sciences*, *12*(13), 6510. https://doi.org/10.3390/app12136510

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pachella, R. G. (2021). The Interpretation of Reaction Time in Information-Processing Research 1. In *Routledge eBooks*. https://doi.org/10.4324/9781003176688-2

Panis, S., Schmidt, F., Wolkersdorfer, M. P., & Schmidt, T. (2020). Analyzing response times and other types of Time-to-Event data using event history analysis: A tool for mental chronometry and cognitive psychophysiology. *I-Perception*, *11*(6), 204166952097867. https://doi.org/10.1177/2041669520978673

Panis, S., & Schmidt, T. (2016). What Is Shaping RT and Accuracy Distributions? Active and Selective Response Inhibition Causes the Negative Compatibility Effect. *Journal of Cognitive Neuroscience*, *28*(11), 1651–1671. https://doi.org/10.1162/jocn_a_00998

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 358–377. https://doi.org/10.1037/0096-1523.10.3.358

Paulignan, Y., Jeannerod, M., MacKenzie, C. L., & Marteniuk, R. G. (1991). Selective perturbation of visual input during prehension movements. *Experimental Brain Research*, *83*(3). https://doi.org/10.1007/bf00229827

Posner, M. I. (2005). Timing the Brain: Mental Chronometry as a Tool in Neuroscience. *PLoS Biology*, *3*(2), e51. https://doi.org/10.1371/journal.pbio.0030051

Proctor, R. W., Miles, J. J., & Baroni, G. (2011). Reaction time distribution analysis of spatial correspondence effects. *Psychonomic Bulletin & Review*, *18*(2), 242–266. https://doi.org/10.3758/s13423-011-0053-5

Rajsic, J., & Wilson, D. E. (2014). Asymmetrical access to color and location in visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 1902–1913. https://doi.org/10.3758/s13414-014-0723-2

Ranjith, N. (2012). Serial position curve. *Encyclopedia of the Sciences of Learning*, 3050–3052.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295x.85.2.59

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review*, *88*(6), 552–572. https://doi.org/10.1037/0033-295x.88.6.552

Ratcliff, R., Smith, P. L., Brown, S., & McKoon, G. (2016). Diffusion Decision Model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Redmann, A., FitzPatrick, I., Hellwig, F., & Indefrey, P. (2014). The use of conceptual components in language production: An ERP study. *Frontiers in Psychology*, *5*, 363.

Redmann, A., FitzPatrick, I., & Indefrey, P. (2019). The time course of colour congruency effects in picture naming. *Acta Psychologica*, *196*, 96–108.

Rheem, H., Verma, V., & Becker, D. V. (2018). Use of mouse-tracking method to measure cognitive load. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1982–1986. https://doi.org/10.1177/1541931218621449

Ridderinkhof, K. R. (2002). Micro- and macro-adjustments of task set: Activation and suppression in conflict tasks. *Psychological Research*, *66*(4), 312–323. https://doi.org/10.1007/s00426-002-0104-7

Saint-Aubin, J., & Poirier, M. (2000). Immediate serial recall of words and nonwords: Tests of the retrieval-based hypothesis. *Psychonomic Bulletin & Review*, *7*(2), 332–340.

Scarfe, P., & Glennerster, A. (2019). The science behind virtual reality displays. *Annual Review of Vision Science*, *5*(1), 529–547. https://doi.org/10.1146/annurev-vision-091718-014942

Schmidt, F., Haberkamp, A., & Schmidt, T. (2011). Dos and don'ts in response priming research. *Advances in Cognitive Psychology*, *7*(1), 120–131. https://doi.org/10.2478/v10053-008-0092-2

Schmidt, F., & Schmidt, T. (2010). Feature-based attention to unconscious shapes and colors. *Attention, Perception, & Psychophysics*, *72*(6), 1480–1494. https://doi.org/10.3758/app.72.6.1480

Schmidt, F., Weber, A., & Schmidt, T. (2014). Activation of response force by self-splitting objects: Where are the limits of feedforward Gestalt processing? *Journal of Vision*, *14*(9), 20–20. https://doi.org/10.1167/14.9.20

Schmidt, T. (2002). The finger in flight: Real-Time motor control by visually masked color stimuli. *Psychological Science*, *13*(2), 112–118. https://doi.org/10.1111/1467-9280.00421

Schmidt, T., & Biafora, M. (2024). A theory of visibility measures in the dissociation paradigm. *Psychonomic Bulletin & Review*, *31*(1), 65–88. https://doi.org/10.3758/s13423-023-02332-z

Schmidt, T., Haberkamp, A., Veltkamp, G. M., Weber, A., Seydell-Greenwald, A., & Schmidt, F. (2011). Visual Processing in Rapid-Chase Systems: Image Processing, Attention, and Awareness. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00169

Schmidt, T., Hauch, V., & Schmidt, F. (2015). Mask-triggered thrust reversal in the negative compatibility effect. *Attention, Perception, & Psychophysics*, *77*(7), 2377–2398. https://doi.org/10.3758/s13414-015-0923-4

Schmidt, T., Panis, S., Wolkersdorfer, M. P., & Vorberg, D. (2022). Response inhibition in the Negative Compatibility Effect in the absence of inhibitory stimulus features. *Open Psychology*, *4*(1), 219–230. https://doi.org/10.1515/psych-2022-0012

Schmidt, T., & Schmidt, F. (2018). *An accumulator model for primes and targets with independent response activation rates: Basic equations for average response times* (arXiv:1804.08513). arXiv. http://arxiv.org/abs/1804.08513

Schmidt, T., & Vorberg, D. (2006). Criteria for unconscious cognition: Three types of dissociation. *Perception & Psychophysics*, *68*(3), 489–504. https://doi.org/10.3758/BF03193692

Schoemann, M., Lüken, M., Grage, T., Kieslich, P. J., & Scherbaum, S. (2019). Validating mouse-tracking: How design factors influence action dynamics in intertemporal decision making. *Behavior Research Methods*, *51*(5), 2356–2377. https://doi.org/10.3758/s13428-018-1179-4

Schoemann, M., O'Hora, D., Dale, R., & Scherbaum, S. (2021). Using mouse cursor tracking to investigate online cognition: Preserving methodological ingenuity while moving toward reproducible science. *Psychonomic Bulletin & Review*, *28*(3), 766–787. https://doi.org/10.3758/s13423-020-01851-3

Seli, P., Kane, M. J., Smallwood, J., Schacter, D. L., Maillet, D., Schooler, J. W., & Smilek, D. (2018). Mind-Wandering as a Natural kind: A Family-Resemblances view. *Trends in Cognitive Sciences*, *22*(6), 479–490. https://doi.org/10.1016/j.tics.2018.03.010

Simon, J. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*(1), 174–176. https://doi.org/10.1037/h0027448

Singer, J. D., & Willett, J. B. (2003). Survival analysis. *Handbook of Psychology*, 555–580. https://doi.org/10.1002/0471264385.wei0222

Slater, M., Gonzalez-Liencres, C., Haggard, P., Vinkers, C., Gregory-Clarke, R., Jelley, S., Watson, Z., Breen, G., Schwarz, R., Steptoe, W., Szostak, D., Halan, S., Fox, D., & Silver, J. (2020). The Ethics of Realism in Virtual and Augmented Reality. *Frontiers in Virtual Reality*, *1*, 1. https://doi.org/10.3389/frvir.2020.00001

Slater, M., Lotto, B., Arnold, M., & Sánchez-Vives, M. V. (2009). How we experience immersive virtual environments: The concept of presence and its measurement *. *Anuario de Psicología*, *40*(2), 193–210.

Smith, K., Morrison, S., Henderson, A. M. E., & Erb, C. D. (2022). Moving beyond response times with accessible measures of manual dynamics. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-20579-9

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Steinicke, F., Hinrichs, K., & Ropinski, T. (2005). Virtual reflections and virtual shadows in mixed reality environments. In *Lecture Notes in Computer Science*. https://doi.org/10.1007/11555261{_

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315. https://doi.org/10.1016/0001-6918(69)90055-9

Stoffels, E. J., & Van Der Molen, M. W. (1988). Effects of visual and auditory noise on visual choice reaction time in a continuous-flow paradigm. *Perception & Psychophysics*, *44*(1), 7–14. https://doi.org/10.3758/BF03207468

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202{_

Tanaka, J. W., & Presnell, L. M. (1999). Color diagnosticity in object recognition. *Perception & Psychophysics*, *61*(6), 1140–1153.

Therriault, D. J., Yaxley, R. H., & Zwaan, R. A. (2009). The role of color diagnosticity in object recognition and representation. *Cognitive Processing*, *10*, 335–342.

Van Leeuwen, C., & Lachmann, T. (2004). Negative and positive congruence effects in letters and shapes. *Perception & Psychophysics*, *66*(6), 908–925. https://doi.org/10.3758/BF03194984

Vath, N., & Schmidt, T. (2007). Tracing sequential waves of rapid visuomotor activation in lateralized readiness potentials. *Neuroscience*, *145*(1), 197–208. https://doi.org/10.1016/j.neuroscience.2006.11.044

Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., & Schwarzbach, J. (2003). Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences*, *100*(10), 6275–6280. https://doi.org/10.1073/pnas.0931489100

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159. https://doi.org/10.1016/j.jml.2007.04.006

Weis, T., Theobald, S., Schmitt, A., Van Leeuwen, C., & Lachmann, T. (2018). There's a SNARC in the size congruity task. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01978

Wheeler, M. E., & Treisman, A. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*(1), 48–64. https://doi.org/10.1037/0096-3445.131.1.48

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

Wilson, C. J., & Soranzo, A. (2015). The Use of Virtual Reality in Psychology: A Case study in Visual Perception. *Computational and Mathematical Methods in Medicine*, *2015*, 1–7. https://doi.org/10.1155/2015/151702

Wolkersdorfer, M. P., Panis, S., & Schmidt, T. (2020). Temporal dynamics of sequential motor activation in a dual-prime paradigm: Insights from conditional accuracy and hazard functions. *Attention, Perception, & Psychophysics*, *82*(5), 2581–2602. https://doi.org/10.3758/s13414-020-02010-5

Woodworth, R. S. (1899). THE ACCURACY OF VOLUNTARY MOVEMENT. *The Journal of Nervous and Mental Disease*, *26*(12), 743–752. https://doi.org/10.1097/00005053-189912000-00005

Appendix

Part 1: N-back ANOVA results without trimming

Classical correct RT vs VR Arrival correct times

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	91.506[a]	2.000[a]	45.753[a]	29.611[a]	< .001[a]	0.457	0.622
	Greenhouse-Geisser	91.506	1.136	80.557	29.611	< .001	0.457	0.622
Residuals	None	55.626	36.000	1.545				
	Greenhouse-Geisser	55.626	20.446	2.721				
Device	None	0.331	1.000	0.331	0.791	0.385	0.002	0.042
Residuals	None	7.520	18.000	0.418				
TrialType	None	6.662	1.000	6.662	16.164	< .001	0.033	0.473
Residuals	None	7.418	18.000	0.412				
N ✻ Device	None	4.309[a]	2.000[a]	2.155[a]	6.789[a]	0.003[a]	0.022	0.274
	Greenhouse-Geisser	4.309	1.475	2.921	6.789	0.008	0.022	0.274
Residuals	None	11.426	36.000	0.317				
	Greenhouse-Geisser	11.426	26.558	0.430				
N ✻ TrialType	None	1.902	2.000	0.951	6.033	0.005	0.009	0.251
	Greenhouse-Geisser	1.902	1.546	1.231	6.033	0.011	0.009	0.251
Residuals	None	5.675	36.000	0.158				

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
	Greenhouse-Geisser	5.675	27.824	0.204				
Device ✻ TrialType	None	0.451	1.000	0.451	3.348	0.084	0.002	0.157
Residuals	None	2.425	18.000	0.135				
N ✻ Device ✻ TrialType	None	0.031	2.000	0.015	0.110	0.896	1.532×10^{-4}	0.006
	Greenhouse-Geisser	0.031	1.661	0.018	0.110	0.861	1.532×10^{-4}	0.006
Residuals	None	5.025	36.000	0.140				
	Greenhouse-Geisser	5.025	29.889	0.168				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	106.279	18	5.904		

Note. Type III Sum of Squares

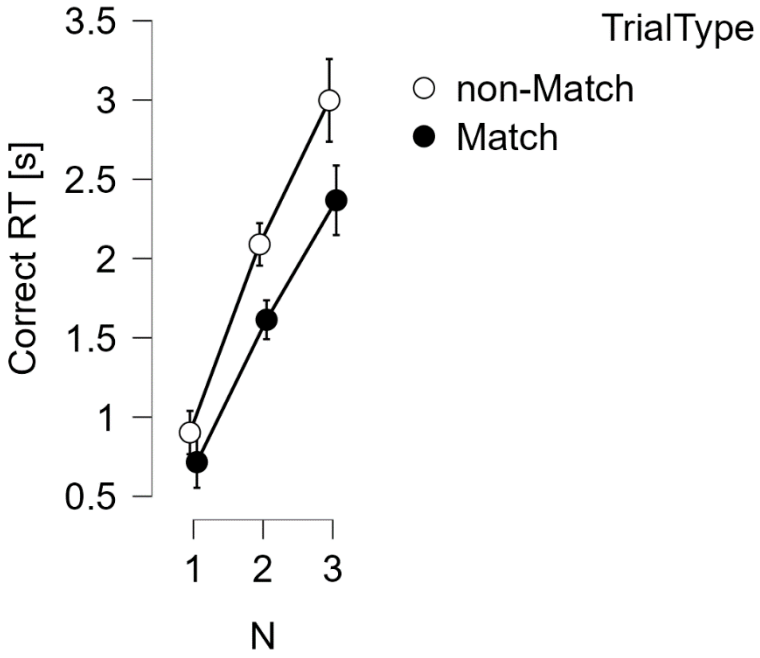
Descriptives

Descriptives

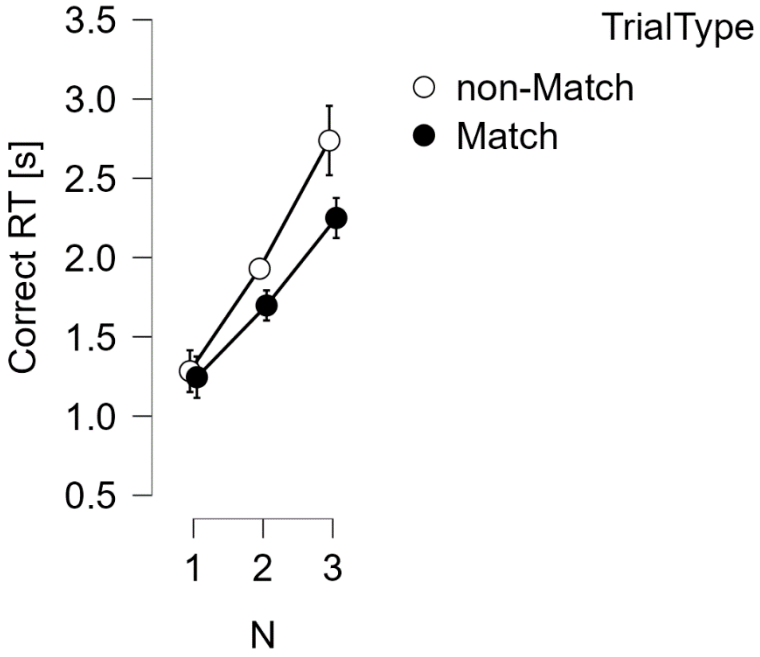
N	Device	TrialType	N	Mean	SD	SE	Coefficient of variation
1	Classical	non-Match	19	0.903	0.203	0.047	0.225
		Match	19	0.716	0.086	0.020	0.121
	VR	non-Match	19	1.283	0.262	0.060	0.204
		Match	19	1.245	0.274	0.063	0.220
2	Classical	non-Match	19	2.089	1.124	0.258	0.538
		Match	19	1.614	0.684	0.157	0.424
	VR	non-Match	19	1.930	0.715	0.164	0.370
		Match	19	1.698	0.561	0.129	0.331
3	Classical	non-Match	19	2.998	1.742	0.400	0.581
		Match	19	2.368	1.536	0.352	0.649
	VR	non-Match	19	2.738	1.484	0.340	0.542
		Match	19	2.250	0.918	0.211	0.408

Descriptives plots

Device: Classical



Device: VR



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.239	24.310	2	< .001	0.568	0.581	0.500
N ✻ Device	0.644	7.468	2	0.024	0.738	0.788	0.500
N ✻ TrialType	0.706	5.915	2	0.052	0.773	0.832	0.500
N ✻ Device ✻ TrialType	0.796	3.888	2	0.143	0.830	0.904	0.500

Classical correct RT vs VR Halfway correct times

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	79.900 [a]	2.000 [a]	39.950 [a]	31.724 [a]	< .001 [a]	0.421	0.638
	Greenhouse-Geisser	79.900	1.145	69.757	31.724	< .001	0.421	0.638
Residuals	None	45.334	36.000	1.259				
	Greenhouse-Geisser	45.334	20.617	2.199				
Device	None	4.310	1.000	4.310	6.019	0.025	0.023	0.251
Residuals	None	12.890	18.000	0.716				
TrialType	None	5.475	1.000	5.475	15.588	< .001	0.029	0.464
Residuals	None	6.322	18.000	0.351				
N ✻ Device	None	7.047 [a]	2.000 [a]	3.523 [a]	8.936 [a]	< .001 [a]	0.037	0.332
	Greenhouse-Geisser	7.047	1.353	5.208	8.936	0.003	0.037	0.332
Residuals	None	14.194	36.000	0.394				
	Greenhouse-Geisser	14.194	24.355	0.583				
N ✻ TrialType	None	1.487	2.000	0.743	5.031	0.012	0.008	0.218
	Greenhouse-Geisser	1.487	1.645	0.904	5.031	0.018	0.008	0.218
Residuals	None	5.319	36.000	0.148				
	Greenhouse-Geisser	5.319	29.615	0.180				

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Device ✻ TrialType	None	0.833	1.000	0.833	6.305	0.022	0.004	0.259
Residuals	None	2.379	18.000	0.132				
N ✻ Device ✻ TrialType	None	0.056	2.000	0.028	0.233	0.794	2.940×10^{-4}	0.013
	Greenhouse-Geisser	0.056	1.548	0.036	0.233	0.737	2.940×10^{-4}	0.013
Residuals	None	4.320	36.000	0.120				
	Greenhouse-Geisser	4.320	27.868	0.155				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	86.460	18	4.803		

Note. Type III Sum of Squares

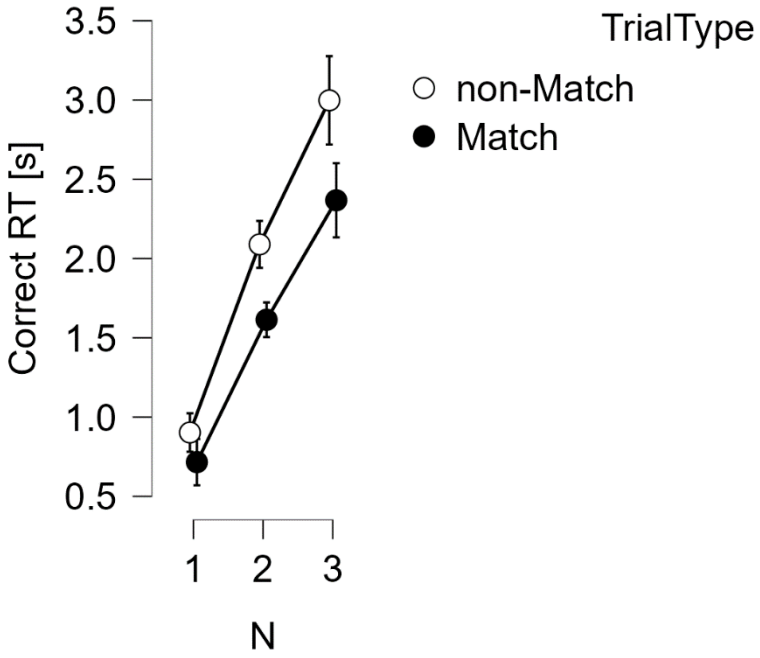
Descriptives

Descriptives

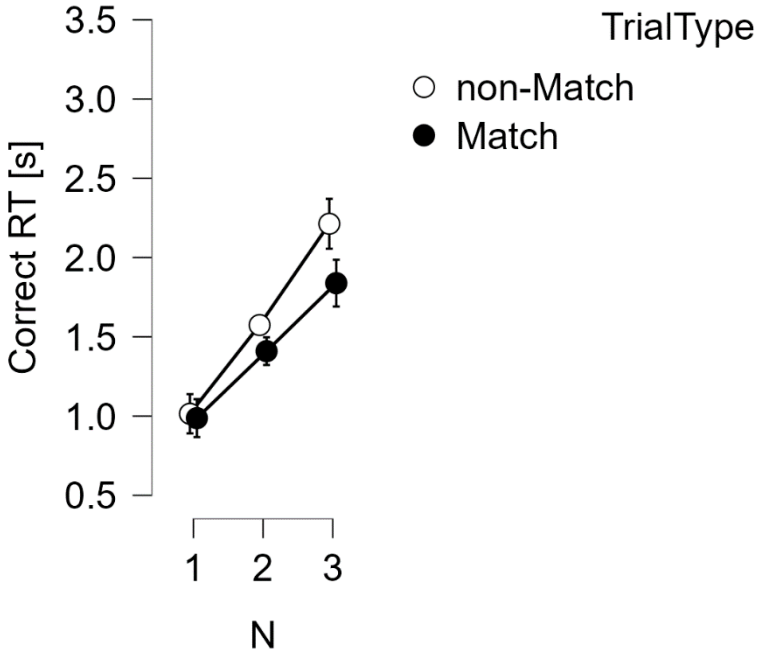
N	Device	TrialType	N	Mean	SD	SE	Coefficient of variation
1	Classical	non-Match	19	0.903	0.203	0.047	0.225
		Match	19	0.716	0.086	0.020	0.121
	VR	non-Match	19	1.015	0.176	0.040	0.173
		Match	19	0.987	0.244	0.056	0.248
2	Classical	non-Match	19	2.089	1.124	0.258	0.538
		Match	19	1.614	0.684	0.157	0.424
	VR	non-Match	19	1.574	0.609	0.140	0.387
		Match	19	1.409	0.457	0.105	0.324
3	Classical	non-Match	19	2.998	1.742	0.400	0.581
		Match	19	2.368	1.536	0.352	0.649
	VR	non-Match	19	2.213	1.140	0.261	0.515
		Match	19	1.838	0.837	0.192	0.455

Descriptives plots

Device: Classical



Device: VR



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.254	23.305	2	< .001	0.573	0.586	0.500
N ✻ Device	0.522	11.056	2	0.004	0.677	0.712	0.500
N ✻ TrialType	0.784	4.128	2	0.127	0.823	0.895	0.500
N ✻ Device ✻ TrialType	0.708	5.866	2	0.053	0.774	0.833	0.500

Classical Accuracy vs VR Arrival Accuracy

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.789 [a]	2.000 [a]	0.395 [a]	37.592 [a]	< .001 [a]	0.330	0.676
	Greenhouse-Geisser	0.789	1.173	0.672	37.592	< .001	0.330	0.676
Residuals	None	0.378	36.000	0.010				
	Greenhouse-Geisser	0.378	21.121	0.018				
Device	None	0.013	1.000	0.013	2.050	0.169	0.005	0.102
Residuals	None	0.114	18.000	0.006				
TrialType	None	0.249	1.000	0.249	22.169	< .001	0.104	0.552
Residuals	None	0.202	18.000	0.011				
N ✻ Device	None	0.003 [a]	2.000 [a]	0.002 [a]	0.401 [a]	0.673 [a]	0.001	0.022
	Greenhouse-Geisser	0.003	1.393	0.002	0.401	0.600	0.001	0.022
Residuals	None	0.135	36.000	0.004				
	Greenhouse-Geisser	0.135	25.079	0.005				
N ✻ TrialType	None	0.141 [a]	2.000 [a]	0.071 [a]	12.199 [a]	< .001 [a]	0.059	0.404
	Greenhouse-Geisser	0.141	1.121	0.126	12.199	0.002	0.059	0.404
Residuals	None	0.209	36.000	0.006				
	Greenhouse-Geisser	0.209	20.181	0.010				
Device ✻ TrialType	None	0.008	1.000	0.008	2.495	0.132	0.003	0.122

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Residuals	None	0.059	18.000	0.003				
N ✻ Device ✻ TrialType	None	0.003	2.000	0.001	0.509	0.605	0.001	0.028
	Greenhouse-Geisser	0.003	1.619	0.002	0.509	0.568	0.001	0.028
Residuals	None	0.090	36.000	0.003				
	Greenhouse-Geisser	0.090	29.149	0.003				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.444	18	0.025		

Note. Type III Sum of Squares

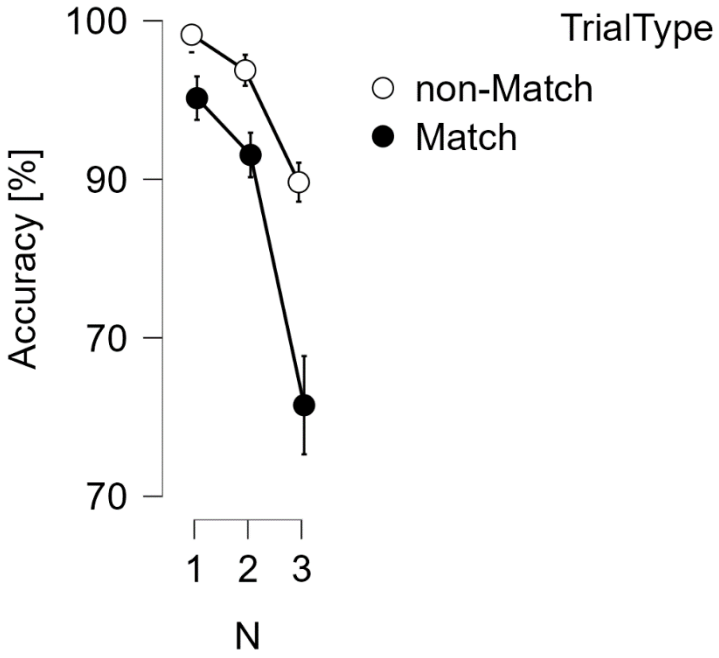
Descriptives

Descriptives

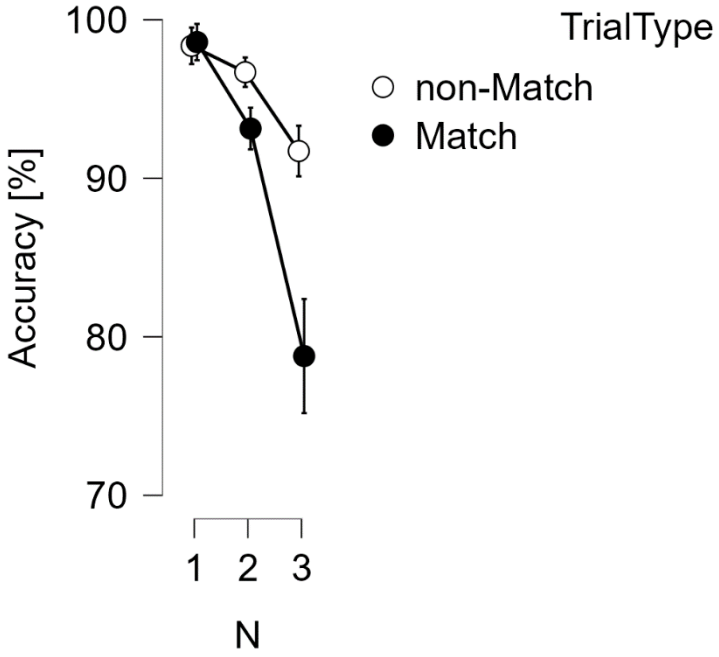
N	Device	TrialType	N	Mean	SD	SE	Coefficient of variation
1	Classical	non-Match	19	0.991	0.013	0.003	0.013
		Match	19	0.951	0.045	0.010	0.047
	VR	non-Match	19	0.984	0.023	0.005	0.023
		Match	19	0.986	0.026	0.006	0.026
2	Classical	non-Match	19	0.969	0.029	0.007	0.029
		Match	19	0.915	0.077	0.018	0.084
	VR	non-Match	19	0.967	0.045	0.010	0.047
		Match	19	0.931	0.072	0.017	0.078
3	Classical	non-Match	19	0.898	0.067	0.015	0.075
		Match	19	0.758	0.162	0.037	0.214
	VR	non-Match	19	0.917	0.083	0.019	0.090
		Match	19	0.788	0.188	0.043	0.239

Descriptives plots

Device: Classical



Device: VR



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.296	20.720	2	< .001	0.587	0.603	0.500
N ✻ Device	0.565	9.719	2	0.008	0.697	0.737	0.500
N ✻ TrialType	0.216	26.038	2	< .001	0.561	0.572	0.500
N ✻ Device ✻ TrialType	0.765	4.555	2	0.103	0.810	0.878	0.500

Classical Accuracy vs VR Halfway Accuracy

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.867 [a]	2.000 [a]	0.434 [a]	43.628 [a]	< .001 [a]	0.347	0.708
	Greenhouse-Geisser	0.867	1.202	0.721	43.628	< .001	0.347	0.708
Residuals	None	0.358	36.000	0.010				
	Greenhouse-Geisser	0.358	21.641	0.017				
Device	None	5.014×10^{-4}	1.000	5.014×10^{-4}	0.079	0.781	2.006×10^{-4}	0.004
Residuals	None	0.114	18.000	0.006				
TrialType	None	0.192	1.000	0.192	15.473	< .001	0.077	0.462
Residuals	None	0.223	18.000	0.012				
N ✻ Device	None	0.002 [a]	2.000 [a]	7.996×10^{-4} [a]	0.185 [a]	0.832 [a]	6.398×10^{-4}	0.010
	Greenhouse-Geisser	0.002	1.340	0.001	0.185	0.743	6.398×10^{-4}	0.010
Residuals	None	0.156	36.000	0.004				
	Greenhouse-Geisser	0.156	24.121	0.006				
N ✻ TrialType	None	0.092 [a]	2.000 [a]	0.046 [a]	6.585 [a]	0.004 [a]	0.037	0.268
	Greenhouse-Geisser	0.092	1.125	0.082	6.585	0.016	0.037	0.268
Residuals	None	0.252	36.000	0.007				
	Greenhouse-Geisser	0.252	20.241	0.012				

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Device ✻ TrialType	None	0.023	1.000	0.023	4.380	0.051	0.009	0.196
Residuals	None	0.094	18.000	0.005				
N ✻ Device ✻ TrialType	None	0.003 [a]	2.000 [a]	0.002 [a]	0.485 [a]	0.620 [a]	0.001	0.026
	Greenhouse-Geisser	0.003	1.442	0.002	0.485	0.560	0.001	0.026
Residuals	None	0.123	36.000	0.003				
	Greenhouse-Geisser	0.123	25.954	0.005				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.414	18	0.023		

Note. Type III Sum of Squares

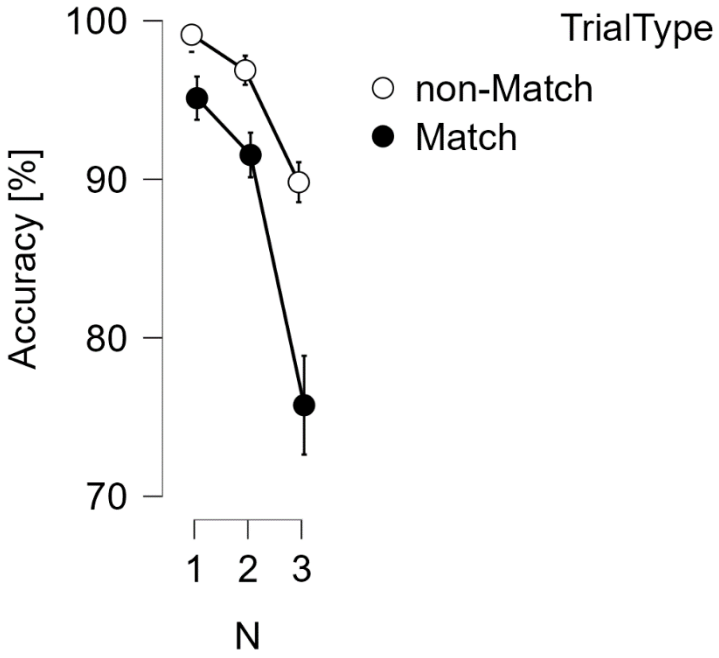
Descriptives

Descriptives

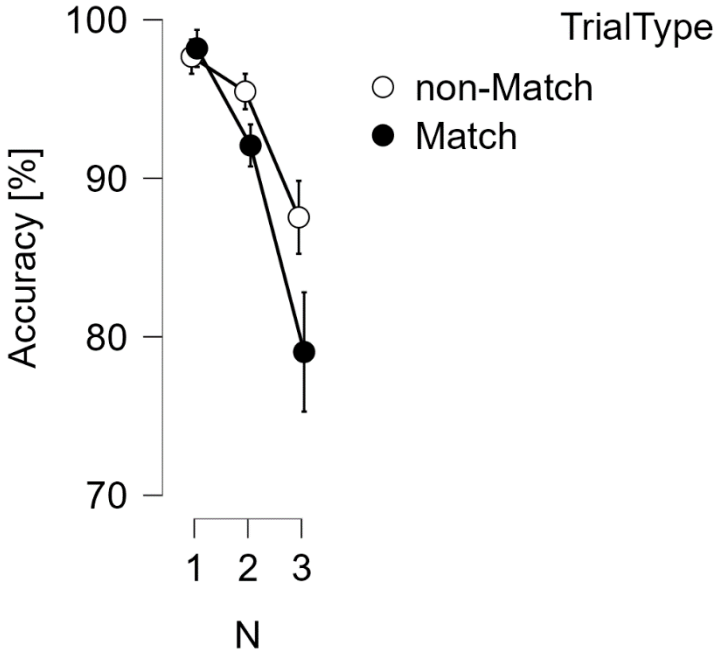
N	Device	TrialType	N	Mean	SD	SE	Coefficient of variation
1	Classical	non-Match	19	0.991	0.013	0.003	0.013
		Match	19	0.951	0.045	0.010	0.047
	VR	non-Match	19	0.977	0.025	0.006	0.026
		Match	19	0.982	0.026	0.006	0.026
2	Classical	non-Match	19	0.969	0.029	0.007	0.029
		Match	19	0.915	0.077	0.018	0.084
	VR	non-Match	19	0.955	0.050	0.012	0.053
		Match	19	0.921	0.072	0.016	0.078
3	Classical	non-Match	19	0.898	0.067	0.015	0.075
		Match	19	0.758	0.162	0.037	0.214
	VR	non-Match	19	0.875	0.102	0.023	0.116
		Match	19	0.790	0.193	0.044	0.244

Descriptives plots

Device: Classical



Device: VR



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.337	18.515	2	< .001	0.601	0.620	0.500
N ✻ Device	0.508	11.529	2	0.003	0.670	0.704	0.500
N ✻ TrialType	0.221	25.627	2	< .001	0.562	0.574	0.500
N ✻ Device ✻ TrialType	0.613	8.321	2	0.016	0.721	0.767	0.500

Correct Classical RT

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	66.977 [a]	2.000 [a]	33.488 [a]	26.346 [a]	< .001 [a]	0.514	0.594
	Greenhouse-Geisser	66.977	1.188	56.383	26.346	< .001	0.514	0.594
Residuals	None	45.759	36.000	1.271				
	Greenhouse-Geisser	45.759	21.382	2.140				
TrialType	None	5.290	1.000	5.290	16.618	< .001	0.041	0.480
Residuals	None	5.730	18.000	0.318				
N ✻ TrialType	None	0.963	2.000	0.481	3.156	0.055	0.007	0.149
	Greenhouse-Geisser	0.963	1.817	0.530	3.156	0.060	0.007	0.149
Residuals	None	5.491	36.000	0.153				
	Greenhouse-Geisser	5.491	32.702	0.168				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	72.173	18	4.010		

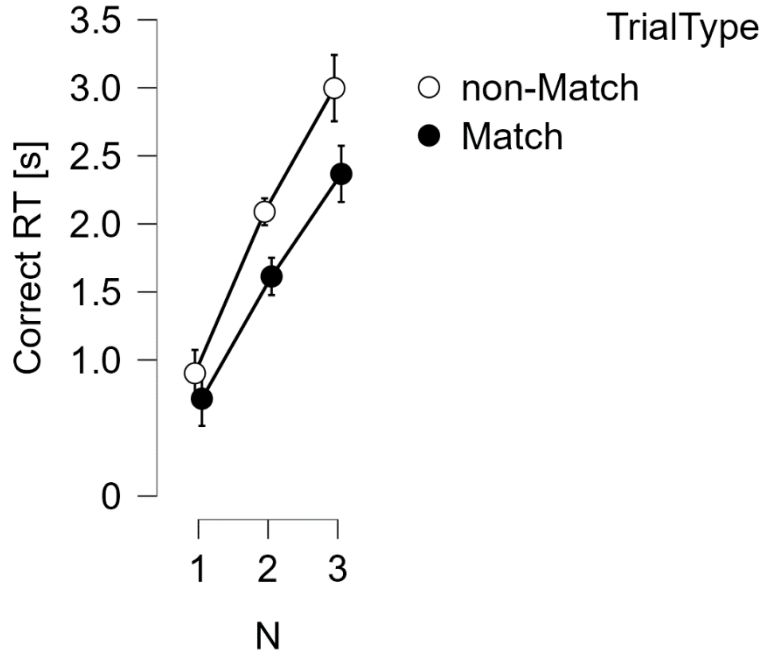
Note. Type III Sum of Squares

Descriptives

Descriptives

N	TrialType	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.903	0.203	0.047	0.225
	Match	19	0.716	0.086	0.020	0.121
2	non-Match	19	2.089	1.124	0.258	0.538
	Match	19	1.614	0.684	0.157	0.424
3	non-Match	19	2.998	1.742	0.400	0.581
	Match	19	2.368	1.536	0.352	0.649

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.316	19.565	2	< .001	0.594	0.612	0.500
N ✻ TrialType	0.899	1.807	2	0.405	0.908	1.000	0.500

Correct VR Arrival times

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	28.838[a]	2.000[a]	14.419[a]	24.380[a]	< .001[a]	0.463	0.575
	Greenhouse-Geisser	28.838	1.212	23.802	24.380	< .001	0.463	0.575
Residuals	None	21.292	36.000	0.591				
	Greenhouse-Geisser	21.292	21.809	0.976				
TrialType	None	1.823	1.000	1.823	7.977	0.011	0.029	0.307
Residuals	None	4.113	18.000	0.229				
N ✻ TrialType	None	0.970[a]	2.000[a]	0.485[a]	3.352[a]	0.046[a]	0.016	0.157
	Greenhouse-Geisser	0.970	1.137	0.854	3.352	0.077	0.016	0.157
Residuals	None	5.210	36.000	0.145				
	Greenhouse-Geisser	5.210	20.458	0.255				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	41.626	18	2.313		

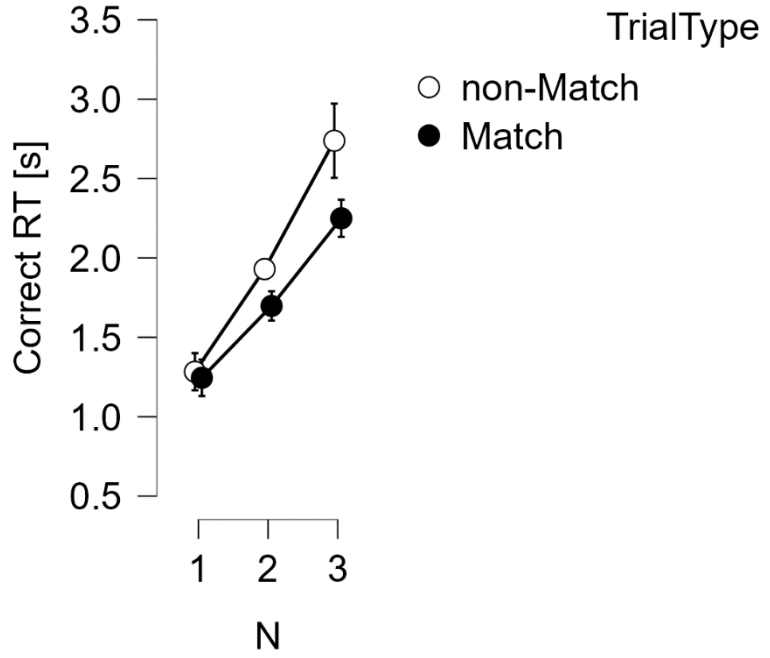
Note. Type III Sum of Squares

Descriptives

Descriptives

N	TrialType	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	1.283	0.262	0.060	0.204
	Match	19	1.245	0.274	0.063	0.220
2	non-Match	19	1.930	0.715	0.164	0.370
	Match	19	1.698	0.561	0.129	0.331
3	non-Match	19	2.738	1.484	0.340	0.542
	Match	19	2.250	0.918	0.211	0.408

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.349	17.881	2	< .001	0.606	0.626	0.500
N ✻ TrialType	0.240	24.239	2	< .001	0.568	0.581	0.500

Correct VR Halfway times

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	19.970 [a]	2.000 [a]	9.985 [a]	26.107 [a]	< .001 [a]	0.470	0.592
	Greenhouse-Geisser	19.970	1.223	16.334	26.107	< .001	0.470	0.592
Residuals	None	13.769	36.000	0.382				
	Greenhouse-Geisser	13.769	22.006	0.626				
TrialType	None	1.018	1.000	1.018	6.168	0.023	0.024	0.255
Residuals	None	2.971	18.000	0.165				
N ✻ TrialType	None	0.580 [a]	2.000 [a]	0.290 [a]	2.516 [a]	0.095 [a]	0.014	0.123
	Greenhouse-Geisser	0.580	1.155	0.502	2.516	0.124	0.014	0.123
Residuals	None	4.149	36.000	0.115				
	Greenhouse-Geisser	4.149	20.792	0.200				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	27.178	18	1.510		

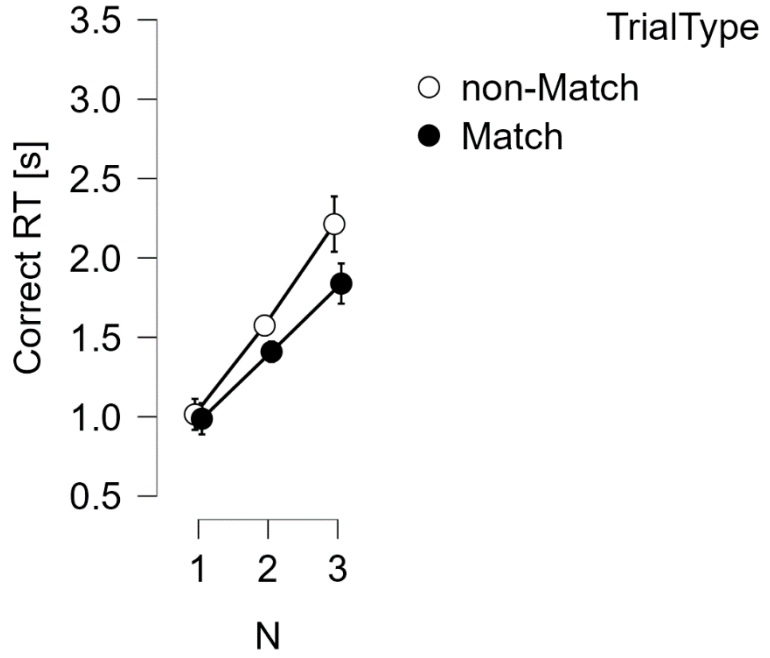
Note. Type III Sum of Squares

Descriptives

Descriptives

N	TrialType	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	1.015	0.176	0.040	0.173
	Match	19	0.987	0.244	0.056	0.248
2	non-Match	19	1.574	0.609	0.140	0.387
	Match	19	1.409	0.457	0.105	0.324
3	non-Match	19	2.213	1.140	0.261	0.515
	Match	19	1.838	0.837	0.192	0.455

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.364	17.175	2	< .001	0.611	0.633	0.500
N ✻ TrialType	0.269	22.349	2	< .001	0.578	0.592	0.500

Accuracy Classical

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.436 [a]	2.000 [a]	0.218 [a]	35.211 [a]	< .001 [a]	0.380	0.662
	Greenhouse-Geisser	0.436	1.348	0.324	35.211	< .001	0.380	0.662
Residuals	None	0.223	36.000	0.006				
	Greenhouse-Geisser	0.223	24.259	0.009				
TrialType	None	0.174	1.000	0.174	20.578	< .001	0.151	0.533
Residuals	None	0.152	18.000	0.008				
N ✻ TrialType	None	0.057 [a]	2.000 [a]	0.028 [a]	9.677 [a]	< .001 [a]	0.049	0.350
	Greenhouse-Geisser	0.057	1.408	0.040	9.677	0.002	0.049	0.350
Residuals	None	0.106	36.000	0.003				
	Greenhouse-Geisser	0.106	25.339	0.004				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.235	18	0.013		

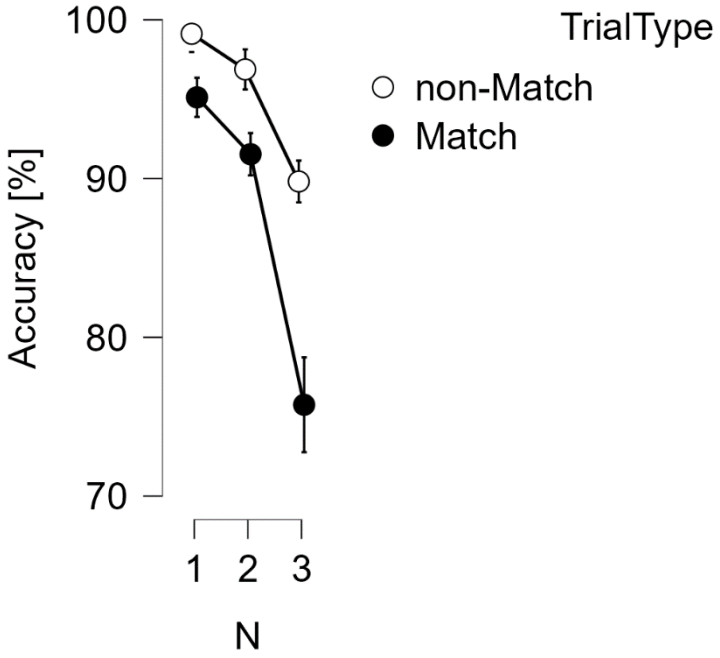
Note. Type III Sum of Squares

Descriptives

Descriptives

N	TrialType	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.991	0.013	0.003	0.013
	Match	19	0.951	0.045	0.010	0.047
2	non-Match	19	0.969	0.029	0.007	0.029
	Match	19	0.915	0.077	0.018	0.084
3	non-Match	19	0.898	0.067	0.015	0.075
	Match	19	0.758	0.162	0.037	0.214

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.516	11.248	2	0.004	0.674	0.709	0.500
N ✻ TrialType	0.579	9.281	2	0.010	0.704	0.746	0.500

Accuracy VR Arrival

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.356 [a]	2.000 [a]	0.178 [a]	22.088 [a]	< .001 [a]	0.318	0.551
	Greenhouse-Geisser	0.356	1.275	0.279	22.088	< .001	0.318	0.551
Residuals	None	0.290	36.000	0.008				
	Greenhouse-Geisser	0.290	22.943	0.013				
TrialType	None	0.084	1.000	0.084	13.815	0.002	0.075	0.434
Residuals	None	0.109	18.000	0.006				
N ✻ TrialType	None	0.087 [a]	2.000 [a]	0.044 [a]	8.131 [a]	0.001 [a]	0.078	0.311
	Greenhouse-Geisser	0.087	1.188	0.074	8.131	0.007	0.078	0.311
Residuals	None	0.193	36.000	0.005				
	Greenhouse-Geisser	0.193	21.381	0.009				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.323	18	0.018		

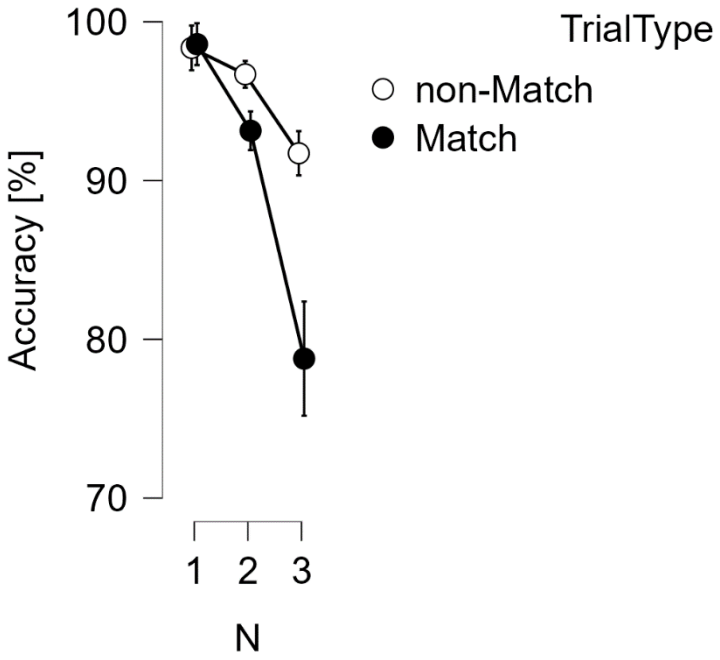
Note. Type III Sum of Squares

Descriptives

Descriptives

N	TrialType	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.984	0.023	0.005	0.023
	Match	19	0.986	0.026	0.006	0.026
2	non-Match	19	0.967	0.045	0.010	0.047
	Match	19	0.931	0.072	0.017	0.078
3	non-Match	19	0.917	0.083	0.019	0.090
	Match	19	0.788	0.188	0.043	0.239

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.431	14.313	2	< .001	0.637	0.664	0.500
N ✻ TrialType	0.316	19.571	2	< .001	0.594	0.612	0.500

Accuracy VR Halfway

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.433 [a]	2.000 [a]	0.216 [a]	26.786 [a]	< .001 [a]	0.349	0.598
	Greenhouse-Geisser	0.433	1.341	0.323	26.786	< .001	0.349	0.598
Residuals	None	0.291	36.000	0.008				
	Greenhouse-Geisser	0.291	24.132	0.012				
TrialType	None	0.041	1.000	0.041	4.462	0.049	0.033	0.199
Residuals	None	0.165	18.000	0.009				
N ✻ TrialType	None	0.039 [a]	2.000 [a]	0.019 [a]	2.590 [a]	0.089 [a]	0.031	0.126
	Greenhouse-Geisser	0.039	1.168	0.033	2.590	0.119	0.031	0.126
Residuals	None	0.270	36.000	0.007				
	Greenhouse-Geisser	0.270	21.018	0.013				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.293	18	0.016		

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p

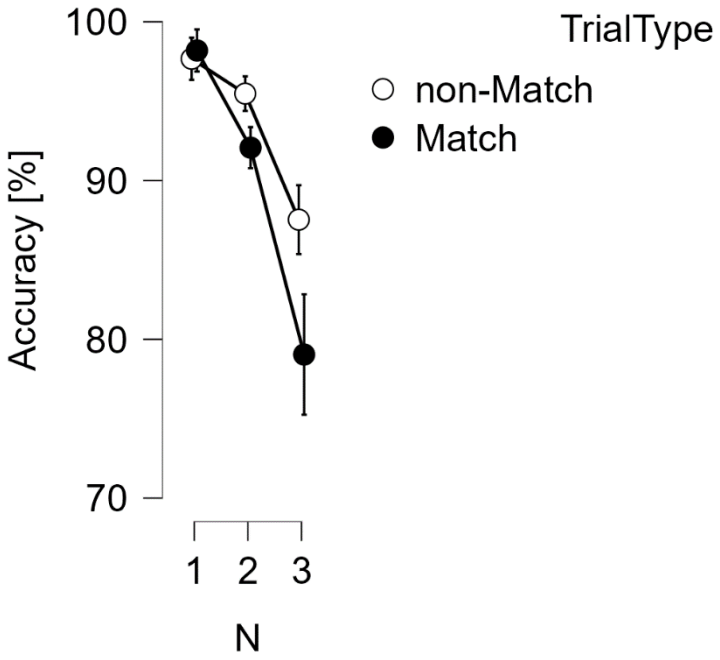
Note. Type III Sum of Squares

Descriptives

Descriptives

	TrialType	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.977	0.025	0.006	0.026
	Match	19	0.982	0.026	0.006	0.026
2	non-Match	19	0.955	0.050	0.012	0.053
	Match	19	0.921	0.072	0.016	0.078
3	non-Match	19	0.875	0.102	0.023	0.116
	Match	19	0.790	0.193	0.044	0.244

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.508	11.508	2	0.003	0.670	0.704	0.500
N ✻ TrialType	0.287	21.209	2	< .001	0.584	0.600	0.500

Part 2: N-back ANOVA results with trimming

Trimming Procedure

The data was trimmed as follows:

We started with 8664 total number of trials before any trimming, excluding task irrelevant trials (Task irrelevant trails refers to trials with no correct response, which are the first trial in N1, the first two trials in N2, and the first 3 trials for N3).

The following process was done for each condition (Device/Trial Type/ N) separately.

1) Outlier detection and trimming using interquartile ranges (IQR), outliers were defined as trials that are outside the range of (Q1 -1.5*IQR - Q3+1.5*IQR)
2) Standard deviation (SD) trimming, outliers were defined as trials outside the range of mean + 2* SD.

Table 1. Trimming procedure

Trimming Type	Condition	Total # of Trials	# of Task relevant Trials*	# Remaining	# Removed	% Remaining	% Removed
3-way-ANOVAs							
Outlier Rejection	Classical vs Arrival	9120	8664	8125	539	93.77	6.22
Outlier Rejection	Classical vs Halfway	9120	8664	8119	545	93.7	6.29
Standard Deviation Trimming	Classical vs Arrival	9120	8125	7725	400	95.07	4.92
Standard Deviation Trimming	Classical vs Halfway	9120	8119	7721	398	95.09	4.9
2-way-ANOVAs							
Outlier Rejection	Classical	4560	4332	4091	241	94.43	5.56
Outlier Rejection	VR Arrival	4560	4332	4034	298	93.12	6.87
Outlier Rejection	VR Halfway	4560	4332	4028	304	92.98	7.01

Standard Deviation Trimming	Classical	4560	4091	3884	207	94.94	5.05
Standard Deviation Trimming	VR Arrival	4560	4034	3841	193	95.21	4.78
Standard Deviation Trimming	VR Halfway	4560	4028	3837	191	95.25	4.74
Error trimming	Classical	4560	3884	3643	241	93.79	6.2
Error trimming	VR Arrival	4560	3841	3543	298	92.24	7.75
Error trimming	VR Halfway	4560	3837	3533	304	92.07	7.92
*Task relevant means: excluding first trial in N1, First 2 trials in N2 first 3 trials in N3, as well excluding data trimmed from outlier rejection in the Standard Deviation Trimming							

Classical correct RT vs VR Arrival correct times

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	64.754 [a]	2.000 [a]	32.377 [a]	26.446 [a]	< .001 [a]	0.434	0.595
	Greenhouse-Geisser	64.754	1.135	57.039	26.446	< .001	0.434	0.595
Residuals	None	44.073	36.000	1.224				
	Greenhouse-Geisser	44.073	20.435	2.157				
Device	None	0.351	1.000	0.351	0.996	0.332	0.002	0.052
Residuals	None	6.342	18.000	0.352				
Trial Type	None	4.443	1.000	4.443	12.692	0.002	0.030	0.414
Residuals	None	6.301	18.000	0.350				
N ✻ Device	None	4.087 [a]	2.000 [a]	2.044 [a]	8.267 [a]	0.001 [a]	0.027	0.315
	Greenhouse-Geisser	4.087	1.323	3.090	8.267	0.005	0.027	0.315
Residuals	None	8.900	36.000	0.247				
	Greenhouse-Geisser	8.900	23.809	0.374				
N ✻ Trial Type	None	1.286 [a]	2.000 [a]	0.643 [a]	6.066 [a]	0.005 [a]	0.009	0.252
	Greenhouse-Geisser	1.286	1.432	0.898	6.066	0.013	0.009	0.252
Residuals	None	3.815	36.000	0.106				
	Greenhouse-Geisser	3.815	25.776	0.148				

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Device ✻ Trial Type	None	0.193	1.000	0.193	2.478	0.133	0.001	0.121
Residuals	None	1.404	18.000	0.078				
N ✻ Device ✻ Trial Type	None	0.207	2.000	0.103	1.196	0.314	0.001	0.062
	Greenhouse-Geisser	0.207	1.543	0.134	1.196	0.307	0.001	0.062
Residuals	None	3.113	36.000	0.086				
	Greenhouse-Geisser	3.113	27.766	0.112				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	78.166	18	4.343		

Note. Type III Sum of Squares

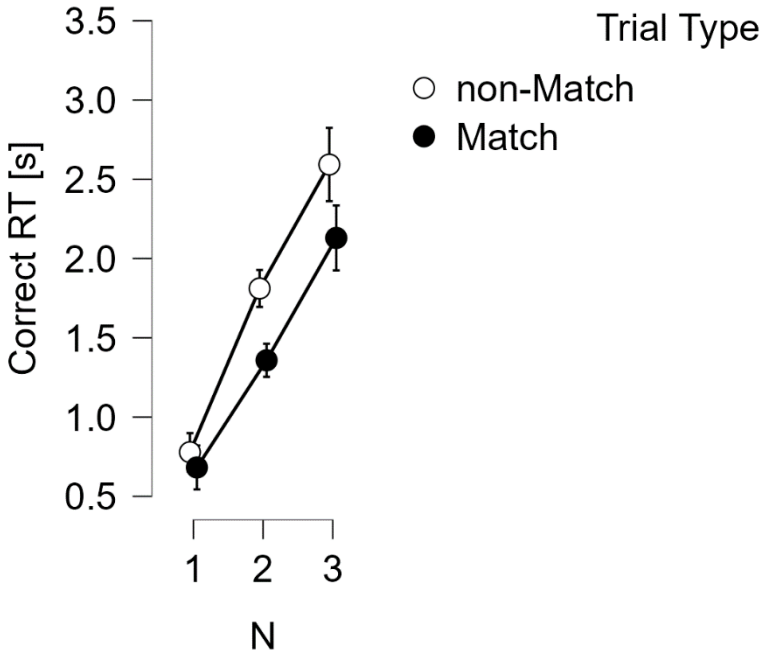
Descriptives

Descriptives

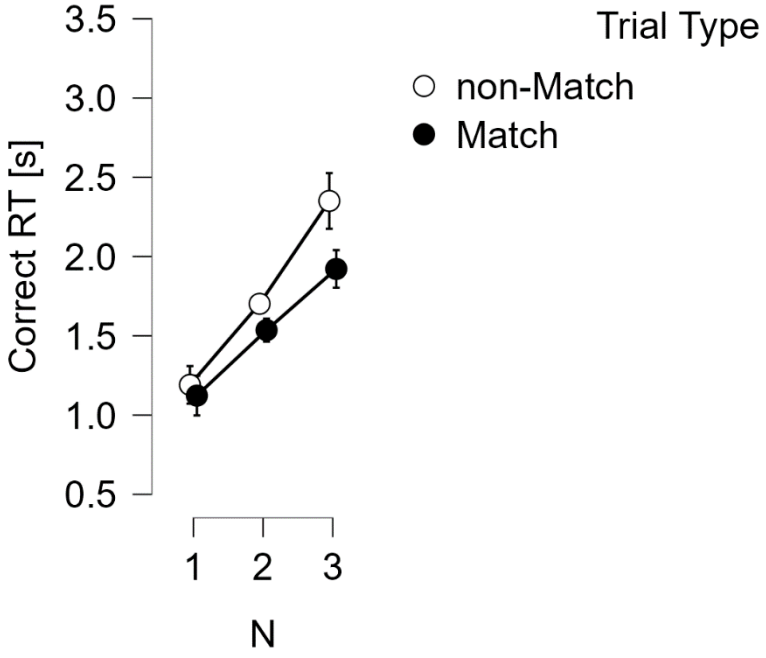
N	Device	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	Classical	non-Match	19	0.779	0.139	0.032	0.179
		Match	19	0.682	0.075	0.017	0.110
	VR	non-Match	19	1.191	0.242	0.056	0.204
		Match	19	1.123	0.201	0.046	0.179
2	Classical	non-Match	19	1.811	0.951	0.218	0.525
		Match	19	1.358	0.586	0.134	0.432
	VR	non-Match	19	1.702	0.644	0.148	0.379
		Match	19	1.536	0.466	0.107	0.303
3	Classical	non-Match	19	2.593	1.518	0.348	0.585
		Match	19	2.130	1.389	0.319	0.652
	VR	non-Match	19	2.351	1.229	0.282	0.523
		Match	19	1.922	0.837	0.192	0.436

Descriptives plots

Device: Classical



Device: VR



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.238	24.382	2	< .001	0.568	0.580	0.500
N ✻ Device	0.488	12.197	2	0.002	0.661	0.694	0.500
N ✻ Trial Type	0.603	8.590	2	0.014	0.716	0.761	0.500
N ✻ Device ✻ Trial Type	0.703	5.980	2	0.050	0.771	0.830	0.500

Classical correct RT vs VR Halfway correct times

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	55.426 [a]	2.000 [a]	27.713 [a]	28.524 [a]	< .001 [a]	0.386	0.613
	Greenhouse-Geisser	55.426	1.146	48.357	28.524	< .001	0.386	0.613
Residuals	None	34.976	36.000	0.972				
	Greenhouse-Geisser	34.976	20.631	1.695				
Device	None	3.776	1.000	3.776	5.505	0.031	0.026	0.234
Residuals	None	12.346	18.000	0.686				
Trial Type	None	3.613	1.000	3.613	11.163	0.004	0.025	0.383
Residuals	None	5.825	18.000	0.324				
N ✻ Device	None	6.842 [a]	2.000 [a]	3.421 [a]	11.009 [a]	< .001 [a]	0.048	0.380
	Greenhouse-Geisser	6.842	1.215	5.631	11.009	0.002	0.048	0.380
Residuals	None	11.187	36.000	0.311				
	Greenhouse-Geisser	11.187	21.871	0.512				
N ✻ Trial Type	None	0.880	2.000	0.440	4.402	0.019	0.006	0.197
	Greenhouse-Geisser	0.880	1.701	0.517	4.402	0.026	0.006	0.197
Residuals	None	3.597	36.000	0.100				
	Greenhouse-Geisser	3.597	30.614	0.118				

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Device ✻ Trial Type	None	0.419	1.000	0.419	4.451	0.049	0.003	0.198
Residuals	None	1.692	18.000	0.094				
N ✻ Device ✻ Trial Type	None	0.151	2.000	0.076	0.950	0.396	0.001	0.050
	Greenhouse-Geisser	0.151	1.551	0.098	0.950	0.378	0.001	0.050
Residuals	None	2.867	36.000	0.080				
	Greenhouse-Geisser	2.867	27.918	0.103				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	60.936	18	3.385		

Note. Type III Sum of Squares

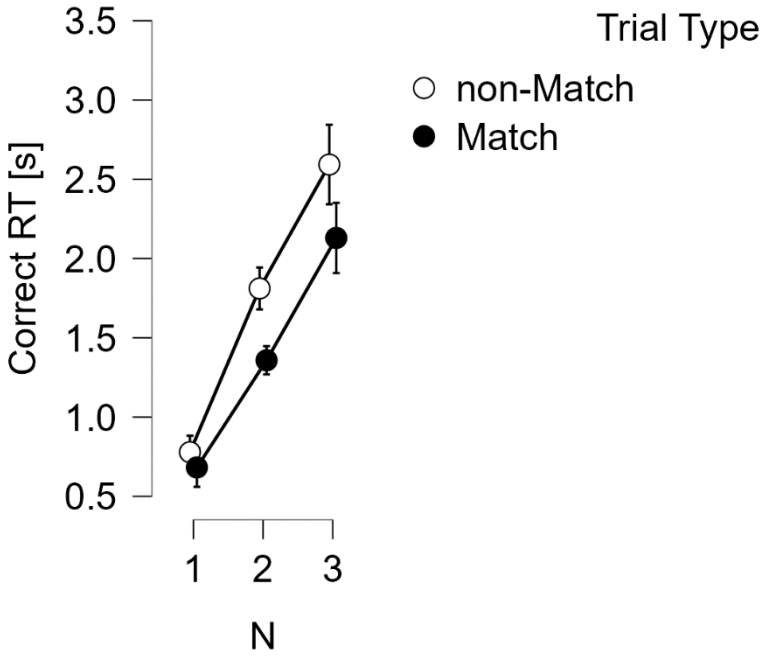
Descriptives

Descriptives

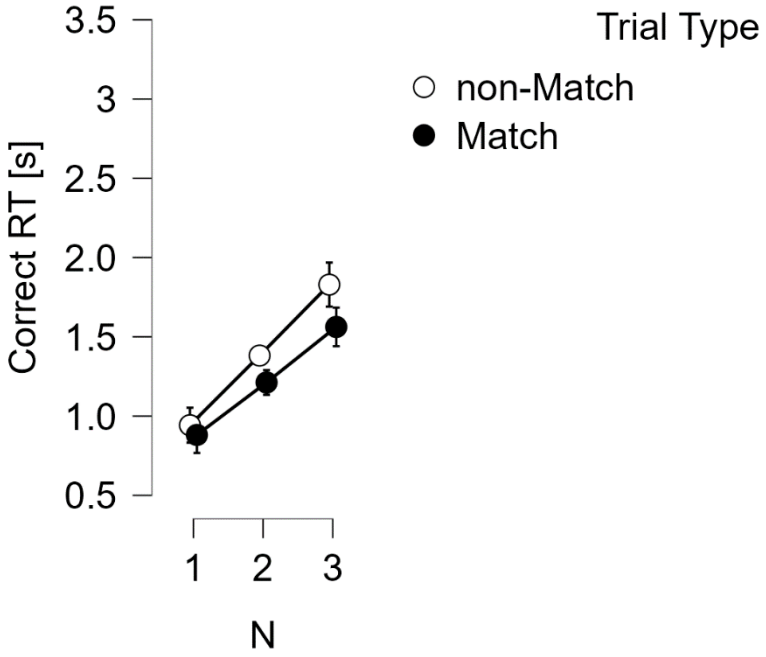
N	Device	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	Classical	non-Match	19	0.779	0.139	0.032	0.179
		Match	19	0.682	0.075	0.017	0.110
	VR	non-Match	19	0.943	0.131	0.030	0.139
		Match	19	0.881	0.130	0.030	0.147
2	Classical	non-Match	19	1.811	0.951	0.218	0.525
		Match	19	1.358	0.586	0.134	0.432
	VR	non-Match	19	1.382	0.554	0.127	0.401
		Match	19	1.212	0.362	0.083	0.298
3	Classical	non-Match	19	2.593	1.518	0.348	0.585
		Match	19	2.130	1.389	0.319	0.652
	VR	non-Match	19	1.829	0.971	0.223	0.531
		Match	19	1.562	0.701	0.161	0.448

Descriptives plots

Device: Classical



Device: VR



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.255	23.227	2	< .001	0.573	0.587	0.500
N ✻ Device	0.354	17.655	2	< .001	0.608	0.628	0.500
N ✻ Trial Type	0.824	3.290	2	0.193	0.850	0.930	0.500
N ✻ Device ✻ Trial Type	0.711	5.810	2	0.055	0.776	0.835	0.500

Classical Accuracy vs VR Arrival Accuracy

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.748 [a]	2.000 [a]	0.374 [a]	35.973 [a]	< .001 [a]	0.328	0.666
	Greenhouse-Geisser	0.748	1.189	0.629	35.973	< .001	0.328	0.666
Residuals	None	0.374	36.000	0.010				
	Greenhouse-Geisser	0.374	21.395	0.017				
Device	None	0.008	1.000	0.008	1.471	0.241	0.004	0.076
Residuals	None	0.102	18.000	0.006				
Trial Type	None	0.180	1.000	0.180	17.096	< .001	0.079	0.487
Residuals	None	0.190	18.000	0.011				
N ✻ Device	None	2.412×10^{-4} [a]	2.000 [a]	1.206×10^{-4} [a]	0.035 [a]	0.965 [a]	1.060×10^{-4}	0.002
	Greenhouse-Geisser	2.412×10^{-4}	1.323	1.823×10^{-4}	0.035	0.909	1.060×10^{-4}	0.002
Residuals	None	0.123	36.000	0.003				
	Greenhouse-Geisser	0.123	23.820	0.005				
N ✻ Trial Type	None	0.139 [a]	2.000 [a]	0.070 [a]	10.574 [a]	< .001 [a]	0.061	0.370
	Greenhouse-Geisser	0.139	1.127	0.124	10.574	0.003	0.061	0.370
Residuals	None	0.237	36.000	0.007				
	Greenhouse-Geisser	0.237	20.290	0.012				

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
Device ✻ Trial Type	None	8.966×10^{-4}	1.000	8.966×10^{-4}	0.410	0.530	3.939×10^{-4}	0.022
Residuals	None	0.039	18.000	0.002				
N ✻ Device ✻ Trial Type	None	0.005 [a]	2.000 [a]	0.003 [a]	0.751 [a]	0.479 [a]	0.002	0.040
	Greenhouse-Geisser	0.005	1.253	0.004	0.751	0.424	0.002	0.040
Residuals	None	0.128	36.000	0.004				
	Greenhouse-Geisser	0.128	22.554	0.006				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.403	18	0.022		

Note. Type III Sum of Squares

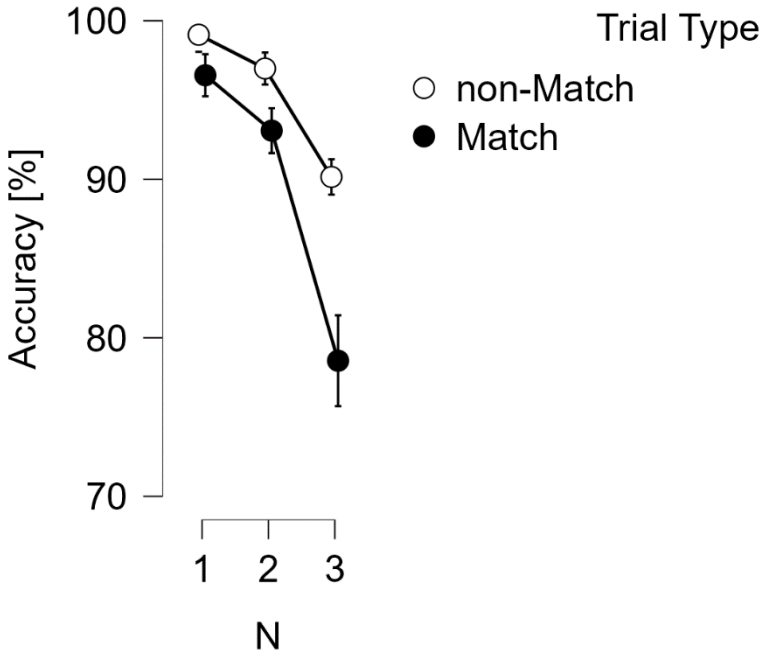
Descriptives

Descriptives

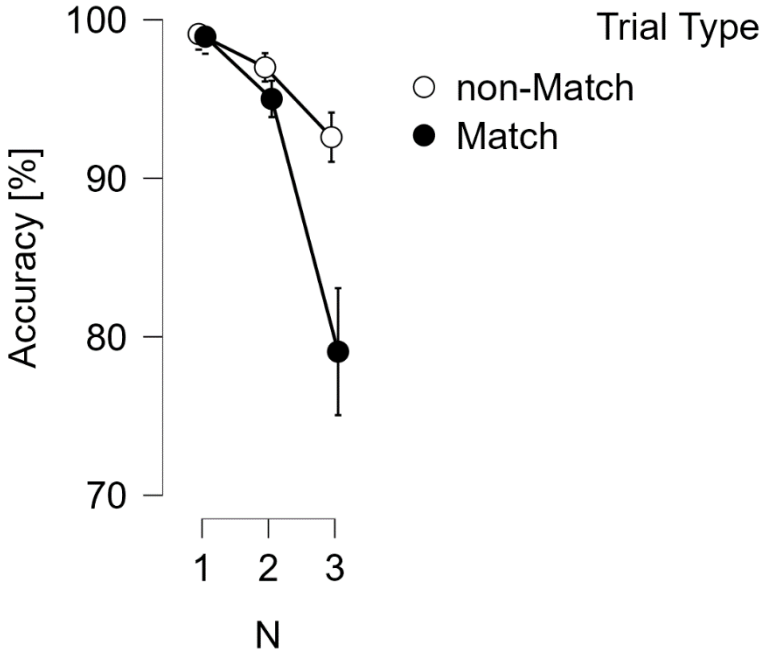
N	Device	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	Classical	non-Match	19	0.991	0.013	0.003	0.013
		Match	19	0.966	0.041	0.009	0.043
	VR	non-Match	19	0.991	0.018	0.004	0.018
		Match	19	0.989	0.026	0.006	0.026
2	Classical	non-Match	19	0.970	0.029	0.007	0.030
		Match	19	0.931	0.077	0.018	0.083
	VR	non-Match	19	0.970	0.044	0.010	0.045
		Match	19	0.950	0.059	0.014	0.062
3	Classical	non-Match	19	0.901	0.063	0.014	0.069
		Match	19	0.786	0.148	0.034	0.188
	VR	non-Match	19	0.926	0.076	0.017	0.082
		Match	19	0.791	0.205	0.047	0.260

Descriptives plots

Device: Classical



Device: VR



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.317	19.510	2	< .001	0.594	0.612	0.500
N ✻ Device	0.489	12.173	2	0.002	0.662	0.694	0.500
N ✻ Trial Type	0.226	25.304	2	< .001	0.564	0.575	0.500
N ✻ Device ✻ Trial Type	0.404	15.414	2	< .001	0.627	0.651	0.500

Classical Accuracy vs VR Halfway Accuracy

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.802[a]	2.000[a]	0.401[a]	39.945[a]	< .001[a]	0.341	0.689
	Greenhouse-Geisser	0.802	1.246	0.644	39.945	< .001	0.341	0.689
Residuals	None	0.361	36.000	0.010				
	Greenhouse-Geisser	0.361	22.421	0.016				
Device	None	1.022×10^{-4}	1.000	1.022×10^{-4}	0.018	0.896	4.340×10^{-5}	9.816×10^{-4}
Residuals	None	0.104	18.000	0.006				
Trial Type	None	0.127	1.000	0.127	10.527	0.004	0.054	0.369
Residuals	None	0.217	18.000	0.012				
N ✻ Device	None	7.197×10^{-4}[a]	2.000[a]	3.599×10^{-4}[a]	0.089[a]	0.915[a]	3.058×10^{-4}	0.005
	Greenhouse-Geisser	7.197×10^{-4}	1.355	5.312×10^{-4}	0.089	0.841	3.058×10^{-4}	0.005
Residuals	None	0.146	36.000	0.004				
	Greenhouse-Geisser	0.146	24.388	0.006				
N ✻ Trial Type	None	0.077[a]	2.000[a]	0.039[a]	5.098[a]	0.011[a]	0.033	0.221

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
	Greenhouse-Geisser	0.077	1.183	0.065	5.098	0.029	0.033	0.221
Residuals	None	0.272	36.000	0.008				
	Greenhouse-Geisser	0.272	21.291	0.013				
Device ✻ Trial Type	None	0.010	1.000	0.010	2.727	0.116	0.004	0.132
Residuals	None	0.064	18.000	0.004				
N ✻ Device ✻ Trial Type	None	7.510×10^{-4} [a]	2.000 [a]	3.755×10^{-4} [a]	0.079 [a]	0.924 [a]	3.191×10^{-4}	0.004
	Greenhouse-Geisser	7.510×10^{-4}	1.197	6.272×10^{-4}	0.079	0.826	3.191×10^{-4}	0.004
Residuals	None	0.172	36.000	0.005				
	Greenhouse-Geisser	0.172	21.554	0.008				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.360	18	0.020		

Note. Type III Sum of Squares

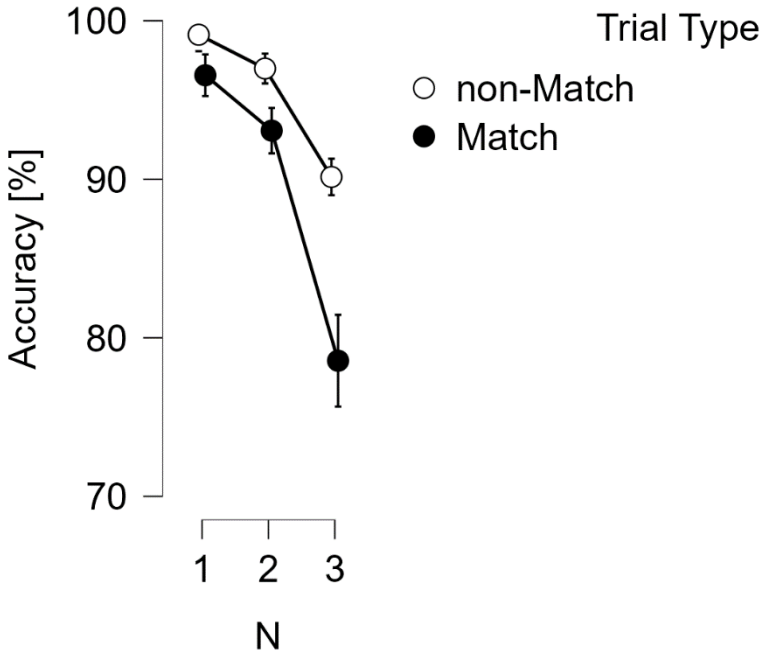
Descriptives

Descriptives

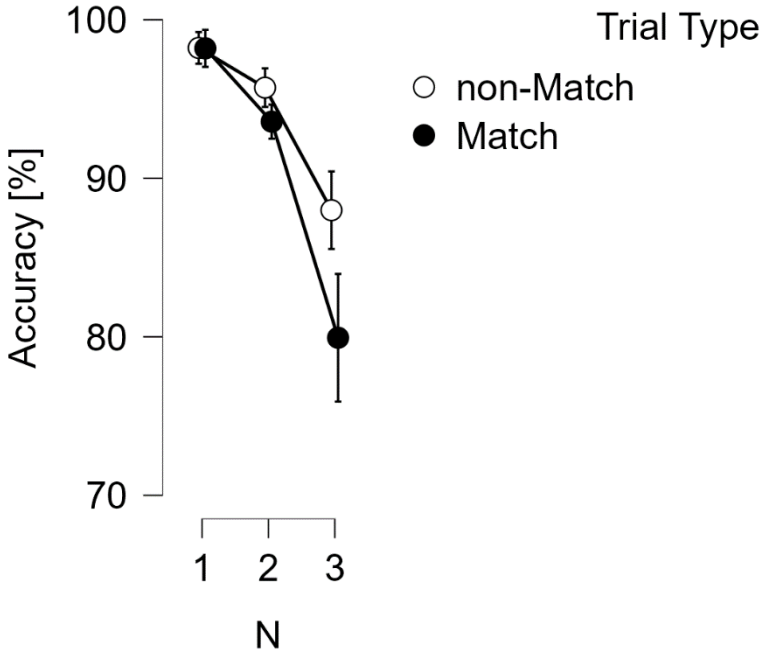
N	Device	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	Classical	non-Match	19	0.991	0.013	0.003	0.013
		Match	19	0.966	0.041	0.009	0.043
	VR	non-Match	19	0.982	0.024	0.006	0.024
		Match	19	0.982	0.026	0.006	0.026
2	Classical	non-Match	19	0.970	0.029	0.007	0.030
		Match	19	0.931	0.077	0.018	0.083
	VR	non-Match	19	0.957	0.052	0.012	0.054
		Match	19	0.936	0.057	0.013	0.061
3	Classical	non-Match	19	0.901	0.063	0.014	0.069
		Match	19	0.786	0.148	0.034	0.188
	VR	non-Match	19	0.880	0.107	0.025	0.122
		Match	19	0.799	0.203	0.047	0.254

Descriptives plots

Device: Classical



Device: VR



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.394	15.817	2	< .001	0.623	0.647	0.500
N ✻ Device	0.524	10.991	2	0.004	0.677	0.713	0.500
N ✻ Trial Type	0.309	19.955	2	< .001	0.591	0.609	0.500
N ✻ Device ✻ Trial Type	0.330	18.858	2	< .001	0.599	0.618	0.500

Correct Classical RT

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	50.587 [a]	2.000 [a]	25.294 [a]	24.989 [a]	< .001 [a]	0.508	0.581
	Greenhouse-Geisser	50.587	1.178	42.939	24.989	< .001	0.508	0.581
Residuals	None	36.439	36.000	1.012				
	Greenhouse-Geisser	36.439	21.206	1.718				
Trial Type	None	3.245	1.000	3.245	13.134	0.002	0.033	0.422
Residuals	None	4.447	18.000	0.247				
N ✻ Trial Type	None	0.830	2.000	0.415	3.680	0.035	0.008	0.170
	Greenhouse-Geisser	0.830	1.956	0.424	3.680	0.036	0.008	0.170
Residuals	None	4.060	36.000	0.113				
	Greenhouse-Geisser	4.060	35.215	0.115				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	54.180	18	3.010		

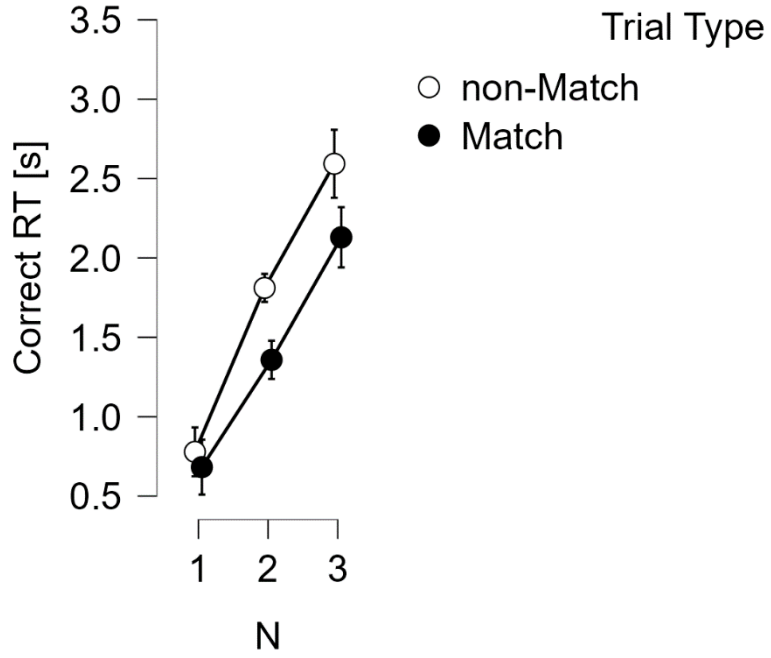
Note. Type III Sum of Squares

Descriptives

Descriptives

N	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.779	0.139	0.032	0.179
	Match	19	0.682	0.075	0.017	0.110
2	non-Match	19	1.811	0.951	0.218	0.525
	Match	19	1.358	0.586	0.134	0.432
3	non-Match	19	2.593	1.518	0.348	0.585
	Match	19	2.130	1.389	0.319	0.652

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.302	20.334	2	< .001	0.589	0.606	0.500
N ✻ Trial Type	0.978	0.383	2	0.826	0.978	1.000	0.500

Correct VR Arrival times

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	18.254 [a]	2.000 [a]	9.127 [a]	19.873 [a]	< .001 [a]	0.425	0.525
	Greenhouse-Geisser	18.254	1.170	15.596	19.873	< .001	0.425	0.525
Residuals	None	16.534	36.000	0.459				
	Greenhouse-Geisser	16.534	21.068	0.785				
Trial Type	None	1.391	1.000	1.391	7.686	0.013	0.032	0.299
Residuals	None	3.259	18.000	0.181				
N ✻ Trial Type	None	0.662 [a]	2.000 [a]	0.331 [a]	4.157 [a]	0.024 [a]	0.015	0.188
	Greenhouse-Geisser	0.662	1.233	0.537	4.157	0.046	0.015	0.188
Residuals	None	2.868	36.000	0.080				
	Greenhouse-Geisser	2.868	22.201	0.129				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	30.328	18	1.685		

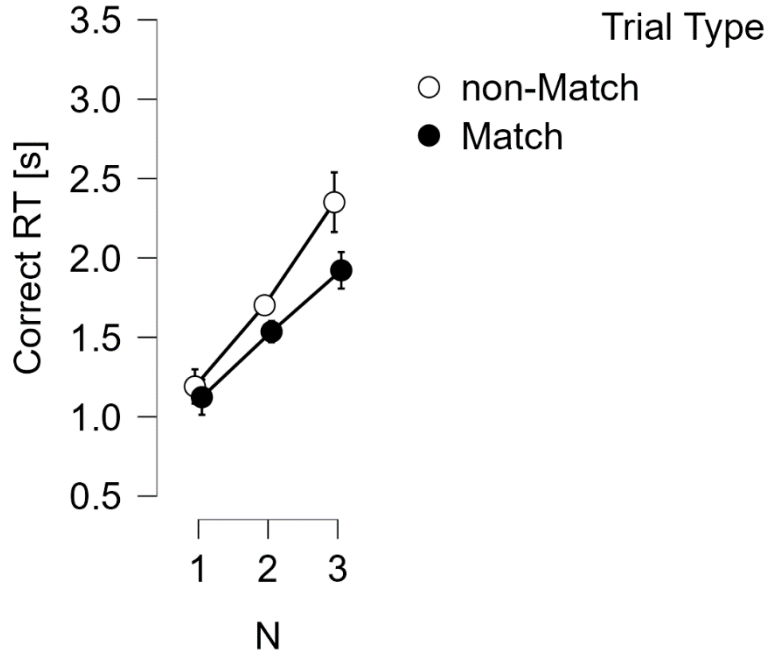
Note. Type III Sum of Squares

Descriptives

Descriptives

N	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	1.191	0.242	0.056	0.204
	Match	19	1.123	0.201	0.046	0.179
2	non-Match	19	1.702	0.644	0.148	0.379
	Match	19	1.536	0.466	0.107	0.303
3	non-Match	19	2.351	1.229	0.282	0.523
	Match	19	1.922	0.837	0.192	0.436

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.291	20.972	2	< .001	0.585	0.601	0.500
N ✻ Trial Type	0.378	16.518	2	< .001	0.617	0.639	0.500

Correct VR Halfway times

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	11.681 [a]	2.000 [a]	5.841 [a]	21.623 [a]	< .001 [a]	0.419	0.546
	Greenhouse-Geisser	11.681	1.140	10.243	21.623	< .001	0.419	0.546
Residuals	None	9.724	36.000	0.270				
	Greenhouse-Geisser	9.724	20.527	0.474				
Trial Type	None	0.786	1.000	0.786	4.608	0.046	0.028	0.204
Residuals	None	3.070	18.000	0.171				
N ✻ Trial Type	None	0.201 [a]	2.000 [a]	0.100 [a]	1.504 [a]	0.236 [a]	0.007	0.077
	Greenhouse-Geisser	0.201	1.216	0.165	1.504	0.238	0.007	0.077
Residuals	None	2.404	36.000	0.067				
	Greenhouse-Geisser	2.404	21.896	0.110				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	19.102	18	1.061		

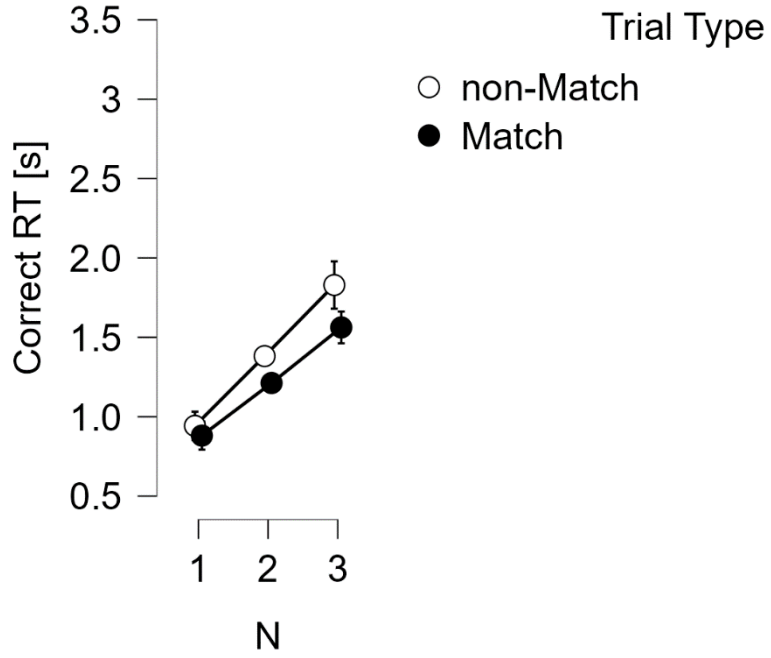
Note. Type III Sum of Squares

Descriptives

Descriptives

N	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.943	0.131	0.030	0.139
	Match	19	0.881	0.130	0.030	0.147
2	non-Match	19	1.382	0.554	0.127	0.401
	Match	19	1.212	0.362	0.083	0.298
3	non-Match	19	1.829	0.971	0.223	0.531
	Match	19	1.562	0.701	0.161	0.448

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.246	23.827	2	< .001	0.570	0.583	0.500
N ✻ Trial Type	0.356	17.564	2	< .001	0.608	0.629	0.500

Accuracy Classical

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.385 [a]	2.000 [a]	0.192 [a]	35.240 [a]	< .001 [a]	0.399	0.662
	Greenhouse-Geisser	0.385	1.382	0.279	35.240	< .001	0.399	0.662
Residuals	None	0.197	36.000	0.005				
	Greenhouse-Geisser	0.197	24.877	0.008				
Trial Type	None	0.103	1.000	0.103	15.494	< .001	0.107	0.463
Residuals	None	0.120	18.000	0.007				
N ✻ Trial Type	None	0.045	2.000	0.023	7.046	0.003	0.047	0.281
	Greenhouse-Geisser	0.045	1.578	0.029	7.046	0.006	0.047	0.281
Residuals	None	0.115	36.000	0.003				
	Greenhouse-Geisser	0.115	28.408	0.004				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.187	18	0.010		

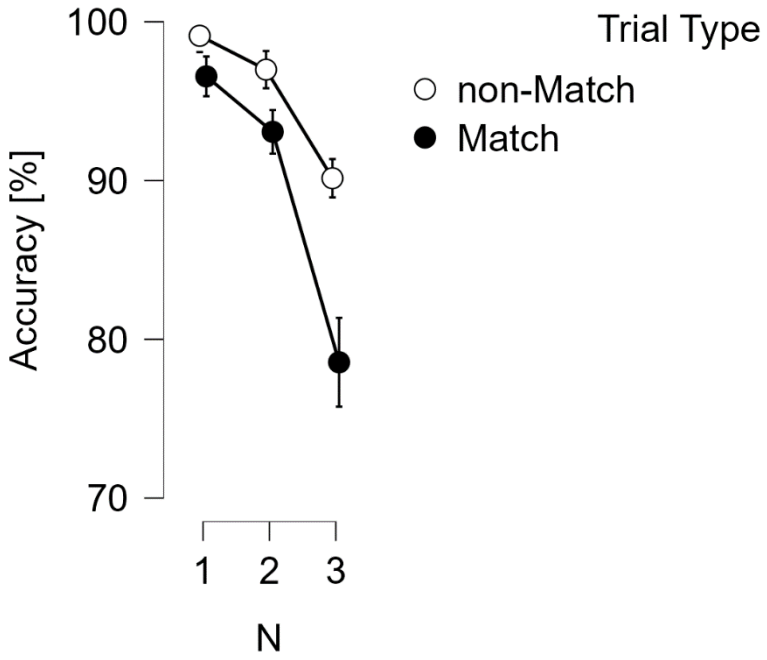
Note. Type III Sum of Squares

Descriptives

Descriptives

N	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.991	0.013	0.003	0.013
	Match	19	0.966	0.041	0.009	0.043
2	non-Match	19	0.970	0.029	0.007	0.030
	Match	19	0.931	0.077	0.018	0.083
3	non-Match	19	0.901	0.063	0.014	0.069
	Match	19	0.786	0.148	0.034	0.188

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.553	10.074	2	0.006	0.691	0.730	0.500
N ✻ Trial Type	0.733	5.286	2	0.071	0.789	0.852	0.500

Accuracy VR Arrival

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.363 [a]	2.000 [a]	0.182 [a]	21.757 [a]	< .001 [a]	0.302	0.547
	Greenhouse-Geisser	0.363	1.227	0.296	21.757	< .001	0.302	0.547
Residuals	None	0.300	36.000	0.008				
	Greenhouse-Geisser	0.300	22.088	0.014				
Trial Type	None	0.078	1.000	0.078	12.844	0.002	0.065	0.416
Residuals	None	0.109	18.000	0.006				
N ✻ Trial Type	None	0.100 [a]	2.000 [a]	0.050 [a]	7.164 [a]	0.002 [a]	0.083	0.285
	Greenhouse-Geisser	0.100	1.071	0.093	7.164	0.013	0.083	0.285
Residuals	None	0.251	36.000	0.007				
	Greenhouse-Geisser	0.251	19.278	0.013				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.318	18	0.018		

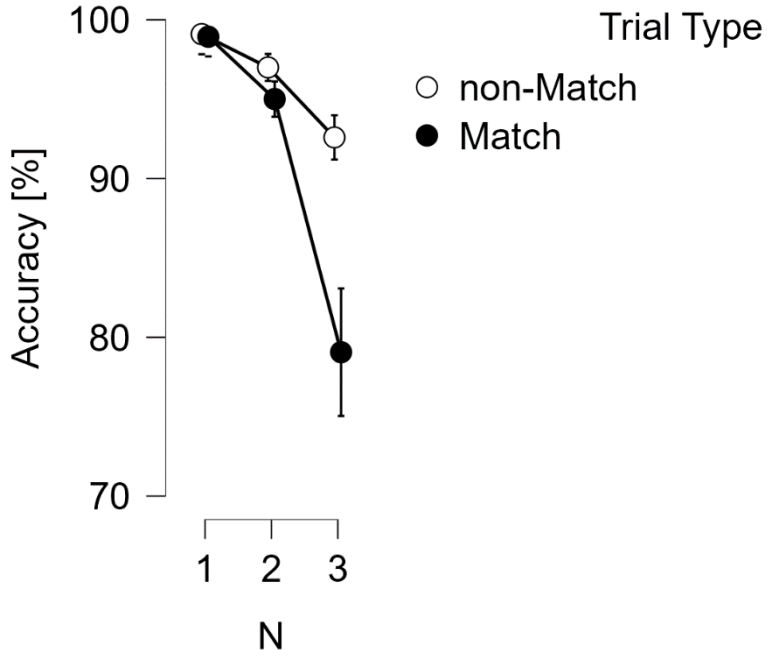
Note. Type III Sum of Squares

Descriptives

Descriptives

N	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.991	0.018	0.004	0.018
	Match	19	0.989	0.026	0.006	0.026
2	non-Match	19	0.970	0.044	0.010	0.045
	Match	19	0.950	0.059	0.014	0.062
3	non-Match	19	0.926	0.076	0.017	0.082
	Match	19	0.791	0.205	0.047	0.260

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.370	16.896	2	< .001	0.614	0.635	0.500
N ✻ Trial Type	0.133	34.344	2	< .001	0.536	0.542	0.500

Accuracy VR Halfway

Within Subjects Effects

	Sphericity correction	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p
N	None	0.418 [a]	2.000 [a]	0.209 [a]	24.212 [a]	< .001 [a]	0.325	0.574
	Greenhouse-Geisser	0.418	1.308	0.319	24.212	< .001	0.325	0.574
Residuals	None	0.311	36.000	0.009				
	Greenhouse-Geisser	0.311	23.544	0.013				
Trial Type	None	0.033	1.000	0.033	3.697	0.070	0.026	0.170
Residuals	None	0.161	18.000	0.009				
N ✻ Trial Type	None	0.033 [a]	2.000 [a]	0.016 [a]	1.796 [a]	0.181 [a]	0.026	0.091
	Greenhouse-Geisser	0.033	1.117	0.029	1.796	0.196	0.026	0.091
Residuals	None	0.329	36.000	0.009				
	Greenhouse-Geisser	0.329	20.106	0.016				

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < .05$).

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Residuals	0.277	18	0.015		

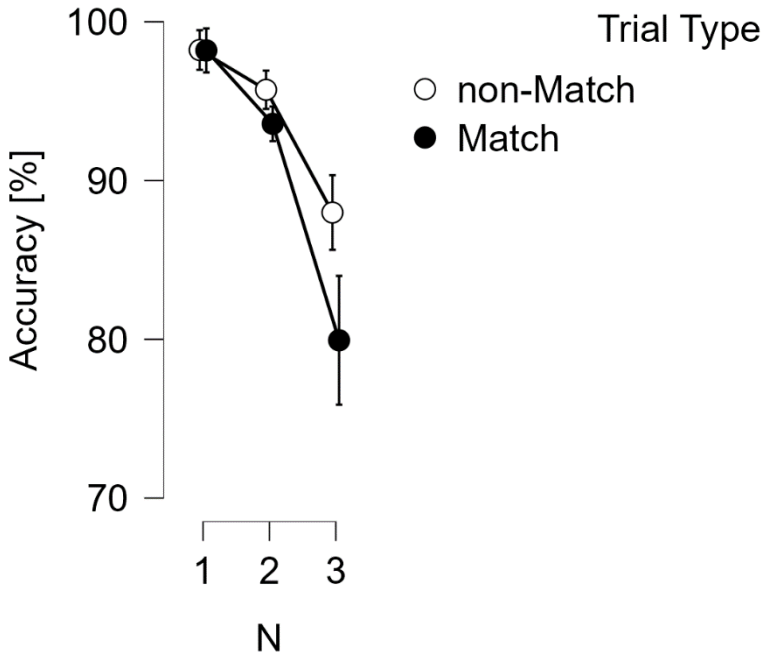
Note. Type III Sum of Squares

Descriptives

Descriptives

N	Trial Type	N	Mean	SD	SE	Coefficient of variation
1	non-Match	19	0.982	0.024	0.006	0.024
	Match	19	0.982	0.026	0.006	0.026
2	non-Match	19	0.957	0.052	0.012	0.054
	Match	19	0.936	0.057	0.013	0.061
3	non-Match	19	0.880	0.107	0.025	0.122
	Match	19	0.799	0.203	0.047	0.254

Descriptives plots



Assumption Checks

Test of Sphericity

	Mauchly's W	Approx. X^2	df	p-value	Greenhouse-Geisser ε	Huynh-Feldt ε	Lower Bound ε
N	0.471	12.801	2	0.002	0.654	0.685	0.500
N ✻ Trial Type	0.209	26.575	2	< .001	0.558	0.569	0.500

Part 3: Color vs Location ANOVA results

Repeated Measures ANOVA

Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
Device	None	0.271	1.000	0.271	74.972	< .001	0.066
Residuals	None	0.025	7.000	0.004			
Number of Items	None	1.617	1.000	1.617	77.251	< .001	0.397
Residuals	None	0.147	7.000	0.021			
Location vs Color	None	1.684	1.000	1.684	131.235	< .001	0.413
Residuals	None	0.090	7.000	0.013			
Device ✻ Number of Items	None	0.006	1.000	0.006	1.982	0.202	0.002
Residuals	None	0.023	7.000	0.003			
Device ✻ Location vs Color	None	0.034	1.000	0.034	15.977	0.005	0.008
Residuals	None	0.015	7.000	0.002			
Number of Items ✻ Location vs Color	None	0.120	1.000	0.120	25.329	0.002	0.029
Residuals	None	0.033	7.000	0.005			
Device ✻ Number of Items ✻ Location vs Color	None	5.109×10^{-6}	1.000	5.109×10^{-6}	0.003	0.958	1.253×10^{-6}
Residuals	None	0.012	7.000	0.002			

Within Subjects Effects

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2

Note. Sphericity corrections not available for factors with 2 levels.

Note. Type III Sum of Squares

Between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p
Residuals	0.104	7	0.015		

Note. Type III Sum of Squares

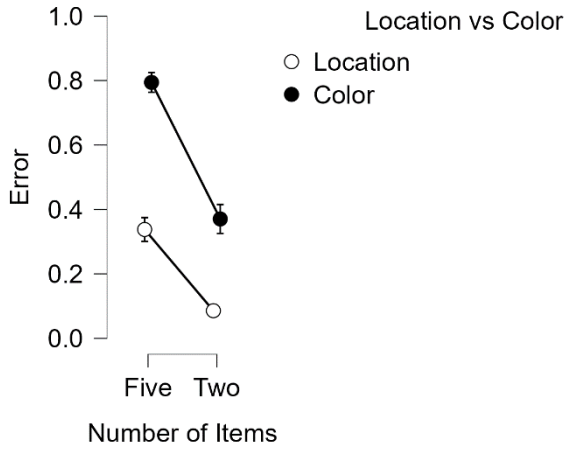
Descriptives

Descriptives

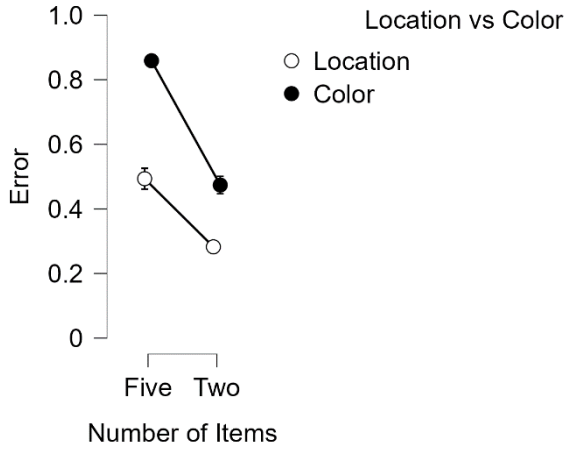
Device	Number of Items	Location vs Color	N	Mean	SD	SE	Coefficient of variation
PC	Five	Location	8	0.338	0.127	0.045	0.375
		Color	8	0.794	0.114	0.040	0.144
	Two	Location	8	0.086	0.021	0.007	0.244
		Color	8	0.370	0.117	0.041	0.316
VR	Five	Location	8	0.493	0.110	0.039	0.223
		Color	8	0.859	0.063	0.022	0.073
	Two	Location	8	0.283	0.016	0.006	0.058
		Color	8	0.474	0.066	0.023	0.139

Descriptives plots

Device: PC



Device: VR



Post Hoc Tests

Post Hoc Comparisons - Number of Items ✻ Location vs Color

		Mean Difference	SE	t	p_{holm}
Five, Location	Two, Location	0.231	0.040	5.781	< .001
	Five, Color	-0.411	0.033	-12.405	< .001
	Two, Color	-0.007	0.046	-0.142	0.890
Two, Location	Five, Color	-0.642	0.046	-13.983	< .001
	Two, Color	-0.238	0.033	-7.185	< .001
Five, Color	Two, Color	0.404	0.040	10.098	< .001

Note. P-value adjusted for comparing a family of 6

Note. Results are averaged over the levels of: Device

Post Hoc Comparisons - Device ✻ Location vs Color

		Mean Difference	SE	t	p_{holm}
PC, Location	VR, Location	-0.176	0.019	-9.301	< .001
	PC, Color	-0.371	0.031	-12.117	< .001
	VR, Color	-0.455	0.032	-14.178	< .001
VR, Location	PC, Color	-0.194	0.032	-6.062	< .001
	VR, Color	-0.278	0.031	-9.100	< .001
PC, Color	VR, Color	-0.084	0.019	-4.431	< .001

Note. P-value adjusted for comparing a family of 6

Note. Results are averaged over the levels of: Number of Items

“Boran loves you.

Yukata!”

My wife, Boran Al Amro 2024

Declaration of authorship

Hiermit versichere ich,

– dass ich die vorgelegte Arbeit selbst angefertigt und alle benutzten Hilfsmittel in der Arbeit angegeben habe,

– dass ich diese Dissertation nicht schon als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht, und

– dass weder die gleiche noch eine andere Abhandlung der Dissertation bei einer anderen Universität oder einem anderen Fachbereich der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau veröffentlicht wurde.

31.03.2024, Kaiserslautern

Omar Jubran

Curriculum Vitae

Name: Omar Jubran

Anschrift: Erwin-Schrödinger-Straße 57, 67653 Kaiserslautern

E-Mail: ojubran@rptu.de

Seit 2018	Promotion (Dr. phil.) RPTU Kaiserslautern (formerly TU Kaiserslautern)
2015-2018	Master-Studium (M.Sc.) in Cognitive Science Technische Universität Kaiserslautern
2009-2014	Bachelor of science in biomedical engineering German Jordanian University (GJU)