Integrated Memory Systems: Interactions and Trade-Offs Between Visual Working Memory and Long-Term Memory

Verena Klar

DPhil Thesis

Department of Experimental Psychology

New College

University of Oxford

Michaelmas Term 2024/25



Abstract

This thesis aimed to extend our understanding of the interactions and dissociations that occur between working memory (WM) and long-term memory (LTM), as well as further defining characteristics of both memory systems from theoretical and empirical perspectives.

Chapter 2 explored the WM compression effect using a continuous response paradigm. Results showed an advantage for the items repeating on every stimulus array in the experiment. However, no improvement in performance was found for non-repeating items. This indicates that the learned item did not free up additional capacity. Even when instructed explicitly, most participants failed to notice the repeating item. This implies that the benefits were likely implicit. Overall, findings suggest some evidence for prioritisation of learned items over non-learned items and highlight the difficulty of creating LTM traces in WM paradigms.

Chapter 3 introduces the Any-back task, a novel paradigm to measure WM and LTM when they both contribute to a task in which discrete object information and continuous color information need to be maintained throughout interference. This enhances our knowledge of WM and LTM by highlighting differential curves as a function of retention intervals and interference. Performance was characterised by a sharp transition that may mark the hand-off of information from WM into LTM. Further, the paradigm enables future studies examining how performance changes when a continuous feature is bound to a discrete object in memory.

In **Chapter 4** we examined the effect of maintaining in WM a stimulus that acted as the state in a probabilistic reinforcement learning (RL) task. Results showed evidence for an enhancement of learning brought about by WM maintenance. However, this effect disappeared when we introduced experimental blocks designed to interfere with WM to disrupt learning. Moreover, no disruptive effect on RL was found. We

interpret that WM can aid RL when it is reliable, but that, when it is unreliable and not necessary for RL, the two systems can be dissociated.

Chapter 5 presents model simulations using a combined Buffer and a Hopfield network to function as WM and LTM, respectively. These simulations demonstrate how the specialised characteristics of these memory systems are balanced against each other in environments varying stochastically in their memory load and intended memory retention intervals. The model provides an implementation as a justification for why the two systems should co-exist.

Finally, **Chapter 6** analyses a macaque dataset where monkeys were required to retain colours selected from either a full or a restricted colour wheel in WM. Results show that the restricted colour space induced a centrality bias that lasted into the following experimental session. It was accompanied by changes at the single-neuron level, notably with the frontal eye fields (FEF) representing colours from the previous limit at higher precision. At a population level, the bias was accompanied by a shift in neural geometry in FEF and prefrontal cortex (PFC). This, too, carried over into the next testing session in FEF, highlighting the area's role for flexible learning in changing contexts.

Statement of Contribution

The research presented in this thesis was conducted collaboratively with colleagues within and outside of my supervisory team.

The work in Chapters 2 and 3 was devised and interpreted together with Masud Husain, with data collection and analysis done by myself.

Chapter 5 was devised and interpreted in collaboration with Andrew Saxe, with input and supervision from Victor Pedrosa. Model implementation was done by me with input from Andrew Saxe.

Chapter 4 was devised collaboratively with Sanjay Manohar and Kengo Shibata, with input from Eda Mizrak. The experiments were implemented by myself, and data collection carried out by Kengo Shibata. Analysis and interpretation were conducted jointly with Kengo Shibata, with guidance from Sanjay Manohar.

Chapter 6 was devised by Timothy Buschman and Motoaki Uchimura; data collection and preprocessing was undertaken by Motoaki Uchimura. I performed the analysis with Timothy Buschman's support.

Acknowledgements

My first thank you goes to my supervisor Masud Husain, for inviting me to join his lab in Oxford as a research assistant, encouraging me to pursue a PhD, being very supportive of my goals and ambitions, offering advice over the years, and unrivalled draft turnaround times. I also thank my secondary supervisor, Andrew Saxe, for his enthusiasm about science and even the smallest of my results, for teaching me modeling techniques, and for giving me the opportunity to integrate into the culture at Gatsby/SWC.

Further, I would like to thank the Economic and Social Sciences Research Council who awarded me a Grand Union DTP AQM scholarship as well as New College who awarded me a 1379 Old Members Scholarship. I am highly thankful for their support which allowed me to complete this research.

My collaborators at Oxford also deserve heartfelt thanks. Sanjay Manohar, for always being there to help when needed, asking the tough questions while maintaining a gentle presence, and for always taking me seriously, even when I was probably wrong. To my collaborator and lab mate, Kengo Shibata, for making collaboration so enjoyable—I looked forward to every one of our meetings—and for being great company outside of work.

Another thank you goes out to Tim Buschman for supporting my grant application to visit his lab at Princeton, providing a fantastic dataset to work on, and teaching me about monkey neural data analysis. Thanks also to Motoaki Uchimura for sharing his valuable data with me and answering all my questions. To the entire Buschman Lab, thank you for the warm welcome to Princeton, with special mentions to Seth, Polina, and Sina for making my time there memorable.

I would also like to thank the members of the Husain lab. My thanks especially to Sofia, for her advice on all things academic and personal, and always being available

despite her busy schedule, and to Panka, for being a great friend outside of work. I would also like to thank Victor for his role as my postdoc supervisor, and to the Saxe Lab in London for being an inspiring and welcoming group.

To finish up the academic mentions, I would like to thank Bernhard Staresina and Rosanne Rademaker for acting as my thesis examiners, for making the viva experience enjoyable and memorable, and providing much helpful feedback in the process.

I am grateful for my mother who has always been supportive of my endless studies and for my friends from various stages of life for the support they have shown me, Lukas, Jess, Ellie, Christian, Jason, Cleo, Nicholas, Kadi, and Robert. I am especially appreciative of Jenny and Wiebke for their lasting friendship. To my amazing housemates and friends at 93MA over the years, Bethany, Clara, and Selina—with special mentions to Caitlin, Jenny, and Anna. I treasure the memories we have made and keep making, from cozy evenings on the couch to pub nights out, 5am May mornings, midnight summer solstice hikes, and three-hour punting sessions. You made Oxford feel like home. To Catharine, for employing me in the best part-time job imaginable and providing a calm and safe oasis to return to. To Nanny and Poppy, and Aunty Joy, thank you for hosting me while I was writing up my thesis.

Finally, my biggest thank you goes to Jamie-Lee, for your silly humour, kindness, and patience with me. For the big and little things during the final stage of the PhD, from taking me to literal paradise to write up, to making me laugh every day, and making sure I stayed hydrated. I was expecting the stress, but not how much fun this period would be, thanks to you.

Thank you all!

Contents

Statement of Contribution **4**

Acknowledgements **5**

1 Introduction **10**
1.1 General Overview 10
1.2 Working Memory 11
1.2.1 Phenomena and function 12
1.2.2 Neural correlates in imaging and electrophysiology 17
1.2.3 Formal models: Spiking neurons, synapses, and normative aspects 22
1.3 Long-term Memory 24
1.3.1 Behavioural effects in declarative memory 24
1.3.2 Neural correlates of semantic memory, episodic memory 26
1.3.3 Formal models: Hippocampal, semantic and RL models 28
1.4 Integrating Visual Working and Long-term Memory 30
1.4.1 Dissociations, re-associations and candidates for integration . . 30
1.4.2 Influence of prior knowledge on working memory 33
1.4.3 Working memory as a platform and bottleneck for long-term memory 42
1.5 Thesis Scope 47

2 Investigating the Working Memory Compression Effect in a Continuous Delayed Estimation Paradigm **49**
2.1 Abstract 49
2.2 Introduction 51
2.3 Experiment 1 55
2.3.1 Methods 55
2.3.2 Results 59
2.4 Experiment 2 61
2.4.1 Methods 62
2.4.2 Results 63
2.5 Experiment 3 64
2.5.1 Methods 65
2.5.2 Results 66
2.6 Discussion 67

3 Characteristics of Working and Long-Term Memory – Evidence from the Any-back Task 72
3.1 Abstract . . . 72
3.2 Introduction . . . 74
3.3 Experiment 1 . . . 78
3.3.1 Methods . . . 78
3.3.2 Results . . . 88
3.4 Experiment 2 . . . 99
3.4.1 Methods . . . 99
3.4.2 Results . . . 99
3.5 Discussion . . . 106

4 Non-Essential Maintenance of State Representations in Working Memory Facilitates Reinforcement Learning 116
4.1 Introduction . . . 117
4.2 General Methods . . . 122
4.2.1 Participants . . . 122
4.2.2 General procedure . . . 123
4.2.3 Stimuli . . . 123
4.3 Experiment 1a . . . 124
4.3.1 Methods . . . 124
4.3.2 Results . . . 128
4.4 Experiment 1b . . . 132
4.4.1 Methods . . . 132
4.4.2 Results . . . 133
4.5 Experiment 2a . . . 134
4.5.1 Methods . . . 135
4.5.2 Results . . . 136
4.6 Experiment 2b . . . 138
4.6.1 Results . . . 139
4.7 Discussion . . . 140

5 When Are Two Systems Better Than One? A Normative Neural Network Account of Short-term and Long-term Memory Stores 147
5.1 Abstract . . . 147
5.2 Introduction . . . 149
5.3 Methods . . . 153
5.3.1 Model architecture . . . 153

5.3.2 Environment . . . 154
5.3.3 Assumptions . . . 160
5.3.4 Simulations . . . 162
5.4 Results . . . 163
5.5 Discussion . . . 171

6 Changes in Neural Representational Geometry Reflect Lasting Traces of Induced Centrality Bias in Macaque Working Memory 177
6.1 Abstract . . . 177
6.2 Introduction . . . 179
6.3 Methods . . . 184
6.3.1 Subjects . . . 184
6.3.2 Experimental design . . . 185
6.3.3 Recordings and signal processing . . . 187
6.3.4 Analysis . . . 188
6.4 Results . . . 193
6.4.1 Behavioural data . . . 193
6.4.2 Neural data . . . 199
6.5 Discussion . . . 208

7 Discussion 213
7.1 Summary of Results . . . 213
7.2 Implications and Future Directions . . . 217
7.3 Conclusion . . . 222

References 224

Appendix 284
7.4 Chapter 3 . . . 284
7.4.1 Tables . . . 284
7.5 Chapter 4 . . . 292
7.5.1 Tables . . . 292
7.6 Chapter 5 . . . 294
7.6.1 Figures . . . 294
7.7 Chapter 6 . . . 299
7.7.1 Figures . . . 299
7.7.2 Tables . . . 304

1 Introduction

1.1 General Overview

Research has long sought to grasp the nature of interactions, as well as commonalities and distinctions between working memory (WM) and long-term memory (LTM). WM is a resource-limited system that can be used to store a few pieces of information for a few seconds at a time in short-term memory (STM), while also allowing manipulation of this information or using it to guide action (Baddeley & Hitch, 1974). LTM, in contrast, is a system with virtually unlimited capacity, designed to store information for periods ranging from days to years (Tulving, 1986, 2002). Both rely on encoding, maintenance, and retrieval mechanisms, and on partially shared, partially dissociated brain mechanisms. What drives their specialisations, and how do they interact to allow for complementary use?

Intuitively, one might simply characterise LTM as a scaffold for WM, providing additional information, and WM as a gateway into LTM, acting as a bottleneck for filtering unnecessary information, akin to attention (Gazzaley & Nobre, 2012). There are, however, many facets to investigating their interactions. For example, what determines the successful transformation of information held in WM into an LTM representation? Is any information lost in the process? If so, what information? Or, under what circumstances does WM rely on prior representations in LTM? How does this change the representational format and the precision of the information in WM? Is the recruitment of these resources always beneficial or does information interfere between the two sys-

tems? Following initial dissociations, there is a long tradition in cognitive psychology of studying these systems separately (Shallice & Warrington, 1970). While research efforts in this area have recently increased, many of these questions remain unanswered.

The studies presented in this thesis have been conducted along three main axes that also guide the structure of this Introduction. The different Chapters investigate (1) the importance of a specialisation in memory systems and the characteristics of using them in conjunction, (2) the prerequisites for forming LTM traces that allow for more efficient storage of information in WM and the resulting representations, and (3) the role of active maintenance of information in reinforcement learning (RL).

The following Introduction provides a review of existing literature. This includes a brief outline of important debates and findings for both visual WM and LTM. There is a summary on the literature to date on the dissociation of the two systems and direct comparisons, on the effects of pre-existing knowledge in LTM on WM, and finally on the role of WM maintenance on long-term recall. The aims of this thesis are described at the end of this Introduction.

1.2 Working Memory

Visual WM is the ability to hold visual information in a state of heightened accessibility and manipulability, and serves to separate information from perception over short time spans. The following Chapter introduces long-studied properties of WM, including the severe capacity limitation, flexibility of representation formats, and feature binding.

The terms STM and WM have many definitions and are sometimes distinguished on the basis of timescales, with WM covering a shorter timespan and STM covering an intermediate timespan. Alternatively, they are defined on the basis of including manipulation of information (WM) or only storage of information (STM) (Cowan, 2017). For the purposes of this thesis, these terms are interchangeably used to mean the active maintenance of information.

1.2.1 Phenomena and function

Working memory resolution and representational formats An important question regarding the cognitive psychology of WM has been whether WM stores discrete (Zhang & Luck, 2008) or continuous representations (Ma et al., 2014). Discrete models assume that there are fixed capacity slots, where each slot maintains the same amount of high-resolution information (Zhang & Luck, 2008). Continuous models, on the other hand, argue that WM is a resource that is shared between individual items in memory, with their precision varying based on the resources allocated (Bays et al., 2009; Ma et al., 2014). Since the cumulative WM resource is fixed, characteristic set size effects emerge: typically, as the number of items to be remembered increases, memory performance, encoding rate, and maintenance rate all decrease (Gorgoraptis et al., 2011). Under this resource allocation view, items can be prioritised depending on the task. This yields an extremely high item precision if the task requires it. Items are prioritised based on the task, such as cues indicating their importance or the overall probability of the item being recalled when probed (Myers et al., 2018; Yoo et al.,

2018). WM can also reallocate resources over time; for example, items that are no longer relevant can be deprioritised, whereas items of current or imminent relevance can be prioritised (Gorgoraptis et al., 2011; Udale et al., 2022). Prioritised items are held in a state of heightened precision, which enables manipulation of the stored information, and is marked by a separate neural state. This has been dubbed the focus of attention. Items outside of the focus of attention are maintained at a lower precision or are forgotten (Lewis-Peacock et al., 2012; Myers et al., 2017; Zokaei, Ning, et al., 2014). There are advantages and drawbacks to either account. The resource allocation view allows for graded representations which can measure prioritisation of information in WM, as shown by many studies. However, it is less clear what leads to forgetting, as differences in precision are continuous, a phenomenon that slot-based models can readily explain. While the debate is not settled, in recent years, the evidence has increasingly favoured hybrid or resource-based models to capture the variability observed in WM performance (Hojjati et al., 2024).

The question of whether WM stores objects versus features is another contested area. For example, Vogel and Machizawa (2004) found that WM can store feature conjunctions and whole objects just as well as features (Vogel & Machizawa, 2004). Under a resource-based model, resource can be flexibly allocated to either specific features or objects in WM, depending on task context (Niklaus et al., 2017). Integrating features of objects, however, is costly: although adding features does not impact the number of objects that can be stored, it severely decreases the precision with which they are stored (Fougnie et al., 2010). Moreover, storing information in an object-based manner

does not necessarily provide any unique benefits. Despite performance benefits when presented as objects, features can still be represented as features as evidenced by independent errors (Fougnie et al., 2013). This result has led to the view that combining features to objects works through a general chunking mechanism rather than integrating representations.

Memory limitations and forgetting The question of how units are represented in WM is closely tied to understanding its limitations. While it is widely accepted that WM has a limited capacity, the exact nature of what imposes this limit remains unclear.

Initially, STM was believed to be able to store seven ± two items (Miller, 1956). The consensus for WM has long been that the storage capacity is around four items (Cowan, 2001). However, there has been a debate whether this limit is fixed, and whether it is best described by the number of items regardless of the number of features (Awh et al., 2007; Fukuda et al., 2010; Luck & Vogel, 1997) or by the number of items as well as their feature complexity (Alvarez & Cavanagh, 2004; Hardman & Cowan, 2015). From a resource allocation perspective, capacity limits are not fixed, since memory resources can be shifted between items flexibly according to task priority (Bays, Gorgoraptis, et al., 2011; Bays & Husain, 2008; Myers et al., 2017). Taking encoding costs into account, resources can strategically be shifted to store more items at a lower precision or fewer items at higher precision to minimise behavioural error (Fougnie et al., 2016; Van den Berg & Ma, 2018). In this case, limitations are therefore best understood with regard to precision of maintained information rather than the

number of objects. Another, not mutually exclusive explanation for capacity limits is that they arise through external influences, such as interference from other objects (Oberauer et al., 2016), or internal restrictions such as noise in neural populations (Quirk et al., 2020; van den Berg et al., 2012). Together, these perspectives suggest that WM is limited through multiple sources shaped by a dynamic interplay between internal resource constraints, task complexity, and external interference.

Feature binding In WM, the process of integrating features into an object or remembering an object in a specific location is called binding (Baddeley et al., 2011). Bindings between elements have to be maintained in memory as well as the features, with impaired binding leading to a so-called swap error, where a feature is associated with the wrong object or position. Such swap errors can occur independently for each feature associated with an object (Bays, Wu, & Husain, 2011). Bindings are more prone to interference from other items than simple features (Allen et al., 2006; Wheeler & Treisman, 2002), and storing multiple features within a single object demands more capacity than a single feature (Palmer et al., 2015). These findings show that binding is a resource-intensive process in WM, particularly vulnerable to interference and capacity limitations.

Structured representations Recent research has begun to move away from representing single objects and towards including representation of relations between objects (Bays et al., 2024; Brady et al., 2011). Individual items may interact with one another: similar items may enhance overall memory accuracy through attraction biases (Jiang et al., 2016), they may support discrimination by being repelled from

each other (Bae & Luck, 2017), or they may induce serial dependence based on their context, supporting integration of information over time (Fischer et al., 2020). Going beyond dyadic interactions, when items are grouped to formations, e.g. according to Gestalt cues, this leads to improved WM performance and specifically improved filtering performance and WM capacity (Allon et al., 2019; Peterson & Berryhill, 2013).

Moreover, items can be represented in hierarchical groups with individual items being remembered as part of a so-called ensemble representation (Brady et al., 2011). This entails statistical information about a stimulus set being remembered alongside individual items. While each individual item is biased towards the mean of its sample distribution, the summed accuracy for all items is improved by the ensemble statistic acting as an anchor (Brady & Alvarez, 2011). The narrower the range within the ensemble representation, the more participants tend to use it to determine individual features (Utochkin & Brady, 2020). This effect is explained well by summed activation in feature space as shown by a recent computational model (Robinson and Brady, 2023, see also Chapter 6).

Such studies highlight that WM captures the complexity of entire stimulus sets by encoding not only individual items, but also their relational structure. One interpretation of this is that remembering summary statistics is one way of WM overcoming its capacity limitation.

1.2.2 Neural correlates in imaging and electrophysiology

There are many methods used to investigate neural correlates of cognition. The following sections address some key findings from functional magnetic resonance imaging (fMRI) and electroencephalograms (EEG) as well as from single-cell recordings from intracranial EEG (iEEG) or non-human primate (NHP) studies, all of which have contributed to our understanding of visual WM.

Functional neuroanatomy Many brain areas have been related to WM activities. While the prefrontal cortex (PFC) and parietal cortex are regarded as the focal areas for WM, representations underlying the delay activity in particular have been found to be far more widely distributed throughout the brain, with stimulus-specific information being decodable from a variety of brain areas (Christophel et al., 2012, 2017, 2018; Haxby et al., 2000).

The PFC has long been implicated in control and monitoring processes, and executive functions (Miller & Cohen, 2001). It was first tied to WM due to the discovery of persistent firing in macaque PFC during WM delays (Funahashi, 2006; Fuster & Alexander, 1971). It is regarded as especially important for WM due to its neurons exhibiting mixed selectivity: they can be tuned to several task-relevant parameters at the population level, even when not decodable at the single-cell level (Rigotti et al., 2013). This allows the PFC to be flexible in the type of information it represents. As a result, it has gained a role as a top-down "controller" (D'Esposito & Postle, 2015), recruiting other brain areas usually involved in non-memory processes. For example, it

has been proposed to underlie the transfer of sensory information into WM (Noudoost et al., 2021), the filtering of irrelevant information (McNab & Klingberg, 2008), and to attentional modulation and the integration of converging information from visual dorsal and ventral streams (Takahashi et al., 2013; Voytek & Knight, 2010). Altogether, PFC activity seems to support hallmark features of WM such as flexibility, selectivity, and the capacity to integrate and transform task-relevant information.

Activity in the parietal cortex during WM delay intervals has been suggested as a neural correlate for successful recall (Pessoa et al., 2002) and for capacity, as activity increases with WM load (Todd & Marois, 2004). While both parietal and prefrontal areas respond to increased load, parietal activity appears to peak at two to three objects and is not recruited beyond capacity limits (Linden et al., 2003). Within the intraparietal sulcus, the inferior part may support fixed representations at capacity, while the superior part tracks object number (Xu & Chun, 2006). The parietal cortex is also implicated in object binding, with activity correlating with colour-shape bindings and temporal order positions (Öztekin et al., 2009; Parra et al., 2014). Contrasted with the flexible role of the PFC, the parietal cortex seems to play a key role in representing WM content, particularly in supporting object number, binding, and recall precision.

Initially regarded as a hallmark of episodic LTM, the medial temporal lobe (MTL) has been shown to support several functions of WM as well. At a basic neural level, it has been suggested that persistently active neurons in the MTL support maintenance of WM information (Kamiński et al., 2017), similar to persistent firing in PFC. Patients with MTL lesions, initially thought to only be impaired in their LTM

performance, are also impaired in their visual WM performance, especially at higher WM loads (Ezzyat & Olson, 2008; Jeneson & Squire, 2012; Jeneson et al., 2012). Moreover, these patients have uniquely been found to be impaired in their capacity to bind features to locations (Liang et al., 2016; Olson, 2006). Based on these studies, some researchers have suggested that the hippocampus in particular is crucial for relational binding of high resolution information (Yonelinas, 2013). Converging evidence for this has been found in healthy subjects for colour-location binding (Borders et al., 2021) and object-location as well as object-to-object binding (Piekema et al., 2010), though not all studies agree (Allen et al., 2014). Other studies suggest that the MTL is activated when LTM is involved in a WM (Ranganath, 2006). For example, maintaining images in WM activates PFC was activated, however, retrieving the same information from LTM during WM delay activates the hippocampus (Ranganath et al., 2004). These studies highlight an important role for the MTL and particularly the hippocampus in WM and challenge the traditional dissociation between WM and LTM.

Contrary to flexible representations in PFC, visual sensory areas, as early as V1, have been found to represent stimulus-specific information during WM delay activity in NHPs and humans (Christophel et al., 2012; Harrison & Tong, 2009; Pasternak & Greenlee, 2005; Peters et al., 2015; Riggall & Postle, 2012; Serences et al., 2009). This activity seems to be task-dependent, since it is not present when no visual information has to be maintained (Lee et al., 2013) and can be diminished by attention, such as moving an item outside of the focus of attention (Bettencourt & Xu, 2016; Christophel et al., 2018; Zokaei, Manohar, et al., 2014), although visual areas are able

to maintain sensory and mnemonic activity at the same time, exhibiting resistance to disruption (Rademaker et al., 2019). This evidence has led some researchers to consider the possibility that sensory areas are recruited by PFC to support maintenance of stimulus-specific information. However, successful decoding of stimulus-specific information from parietal and prefrontal areas has called this into question and suggested that the relationship is more flexible and different areas are involved according to task demands (Ester et al., 2015; Lee & Baker, 2016; Serences, 2016; Xu, 2017).

Event-related potentials and oscillations Electrophysiological correlates of WM tasks have also been found in the form of event-related potentials (ERPs) and oscillations.

One ERP, the contralateral delay activity (CDA), arising contralaterally 300–900ms after stimulus onset, has been shown to reflect WM capacity (Luria et al., 2016). CDA amplitude is thought to reflect the number of items maintained in WM and serves as a good measure of both individual capacity and trial-by-trial performance (Adam et al., 2018; Vogel & Machizawa, 2004). Some suggest it tracks the object number rather than the resolution with which objects are stored (Gao et al., 2011; Ikkai et al., 2010), although one study showed that it may also track flexible resource allocation for continuous features (Salahub et al., 2019). The CDA modulation effect of capacity, however, has been replicated by many studies with concerted replication efforts ongoing (Roy & Faubert, 2023; Strzelczyk et al., 2023). It thus seems a reliable marker of WM engagement when justified by the experimental paradigm.

Rhythmic brain activity, or neural oscillations, has been linked to WM control, with different frequency bands supporting distinct functions (Benchenane et al., 2011; Roux & Uhlhaas, 2014). For instance, alpha power suppression is associated with inhibition and filtering (de Vries et al., 2020; Sauseng et al., 2009) or item number (Chen et al., 2022; Fukuda et al., 2015), whereas gamma oscillations are tied to maintenance and item separation, and theta may encode item order (Axmacher et al., 2010; Lundqvist et al., 2016; Roux & Uhlhaas, 2014). Oscillations also support communication between brain regions via phase coupling (phase alignment) or synchronisation (phase and amplitude alignment) (Roux & Uhlhaas, 2014). In NHPs, for example, it has been found that synchronisation of activity between frontal and parietal areas was important for maintenance of contents (Salazar et al., 2012). Alpha-gamma coupling has been linked to feature selection (Davoudi et al., 2021), and theta to WM "replay" during maintenance (Fuentemilla et al., 2010). These electrophysiological markers like ERPs and oscillations offer real-time insights into the neural correlates of WM, and can help to dissociate cognitive mechanisms.

Cellular mechanisms: Persistent, silent, and population activity At a cellular level, WM is thought to rely on persistent firing of neurons in the PFC and other brain regions during the delay periods (Constantinidis et al., 2018; Fuster & Alexander, 1971; Goldman-Rakic, 1995). More recently, it was hypothesised that working memories can also be stored in synapses, through a process called short-term potentiation. This was dubbed activity-silent WM since it is characterised by decodable neural activity during encoding and retrieval of information, but not during the delay (Masse et al.,

2020; Stokes, 2015). The information can be recovered using a visual impulse or a cueing participants to shift information into the focus of attention (Sprague et al., 2016; Wolff et al., 2017), suggesting a form of passive maintenance rather than a recruitment of other brain areas.

Although single-cell PFC activity underlying WM is dynamic and experiences drift, the population code remains steady (Spaak et al., 2017; Stroud et al., 2024; Wolff et al., 2020), explaining how WM can appear as a consistent memory experience. Population codes and neural geometry can be used to investigate how neural activity transforms throughout WM encoding, maintenance, and retrieval (Stokes et al., 2013) and how it relates to behaviour (Kriegeskorte & Wei, 2021; Lundqvist et al., 2018). For example, how task-relevant information like sequence position has been shown to be stored in separate population subspaces (Xie et al., 2022), and how WM and attention may share neural mechanisms as identified by similar subspaces (Panichello and Buschman, 2021, see also Chapter 6).

Together, these findings demonstrate that WM relies on a heterogeneous set of mechanisms, namely persistent activity, silent synaptic mechanisms, and stable population codes, to maintain and transform information over time.

1.2.3 Formal models: Spiking neurons, synapses, and normative aspects

Many models have been devised to fit to empirical data (van den Berg et al., 2014), as well as simulate WM function at various levels of abstraction (Marr, 2010) to explain behavioural and neural findings and offer new theoretical perspectives.

At the neural level, models explain persistent firing through recurrent excitation and inhibition, forming "bump" attractors that gradually drift over time due to noise (Compte, 2000; Durstewitz et al., 2000; Zipser et al., 1993). This drift fundamentally limits WM precision and capacity for binding (Bays, 2014, 2015; Burak & Fiete, 2012; Matthey et al., 2015; Schneegans & Bays, 2019), although these limited resources are maximised by WM using an optimal coding scheme (Koyluoglu et al., 2017).

Activity-silent WM was inspired by models showing that information can be stored via synaptic weights rather than persistent firing (Mongillo et al., 2008), with Hebbian plasticity offering a plausible mechanism to explain selectivity properties in PFC (Fiebig & Lansner, 2017; Lindsay et al., 2017). Recent work suggests persistent activity is required for manipulation, while short-term plasticity may suffice for brief maintenance (Masse et al., 2019).

More complex neural models link smaller components to simulate interactions between brain regions. For instance, bidirectional connections between neuron layers representing sensory stores and a pool of neurons representing PFC can model mixed selectivity and set size effects (Bouchacourt & Buschman, 2019) and show how sharing resources between items can lead to interference and swap errors (Schneegans & Bays, 2017; Swan & Wyble, 2014). In such networks, incorporating activity-silent WM via Hebbian synapses helps reduce interference (Manohar et al., 2019), and network models have also captured WM gating via the basal ganglia or short-term stores (Hazy et al., 2006, 2007; Lindsey & Litwin-Kumar, 2022; O'Reilly & Frank, 2006).

Another class of models, probabilistic models, are either concerned with describing behavioural data or take a more normative stance offering a rational theory of WM. For example, Orban et al. (2008) demonstrated that Bayesian learning can form chunks to guide maintenance and use WM capacity more efficiently. Similarly, probabilistic models have been successful in modelling the maintenance of higher-order estimates of stimulus sets, so-called ensemble representations (Brady & Tenenbaum, 2013). Several other studies analyse WM as an optimisation problem, relying on Information Theory viewing memory as an information processing channel that humans rationally use (Sims, 2018; Sims et al., 2012; Taylor and Bays, 2018; Van den Berg and Ma, 2018.

While a lot of WM findings have been explained by computational modelling, there are few studies that model different systems in an integrated way. In one example, Hedayati et al. (2022) combined a deep neural network with a binding pool model that stores different resources in a shared pool of neurons to model. This allowed prior knowledge and latent representations to be combined with the feature binding of novel information in WM. This seems a promising avenue for bridging the gap between WM and LTM (see also Chapter 5).

1.3 Long-term Memory

1.3.1 Behavioural effects in declarative memory

Long-term memories we are able to consciously recall and articulate are part of the declarative memory system which has two main branches: episodic, or event, memory, and semantic, or fact, memory (Baddeley et al., 2001; Squire, 2004; Tulving,

1993). The distinction between episodic LTM and WM stems from patient studies with amnesia who exhibited deficits in LTM with intact STM (Baddeley & Warrington, 1970). Early accounts argue for a distinction based on whether an experience was personal or not (Tulving, 1986, 1993). In the following, I briefly summarise research outlining the main behavioural and neural characteristics.

In everyday life, episodic or LTM answers the questions of "what", "where", and "when". It is detailed knowledge about a particular event, situated in time and space (Eichenbaum, 2013). One of the most astonishing features of LTM is its extremely large capacity. People can recognise an extraordinarily large number of images (10,000 pictures) after viewing them once (Standing, 1973), and distinguish detailed pictures from novel items, even when they share visual and semantic properties (Brady et al., 2008). LTM function is supported by two distinct recall processes termed recollection and familiarity (Yonelinas, 2001), respectively referring to detailed contextual information accompanied by source memory and a sense of knowing without being able to specify the context. They have been distinguished by different neural correlates (Curran, 2000; Yonelinas et al., 2005). Memories are considered to be robust, requiring retrieval and reconsolidation to be altered (Alberini & LeDoux, 2013), although their vividness fades over time (Cooper et al., 2019). Episodic memories are adaptive, preserving the gist of a memory over the details at the cost of distortion (Nagy et al., 2020; Schacter et al., 2011).

Semantic memory describes the memory of concepts, containing general features of categories of objects or abstract entities, organised in a hierarchical structure of cat-

egories (Ashby & Maddox, 2005; Collins & Quillian, 1969; Squire & Wixted, 2011; Warrington & Shallice, 1984). In contrast to episodic memory, this system stores conceptual knowledge in an abstract, decontextualised code. A lot of this evidence comes from patients with the semantic variant primary progressive aphasia (semantic dementia), who show very specific loss of their category knowledge (Warrington, 1975). In this progressive disease, patients first forget the finer-grained distinctions for subcategories, before forgetting the larger grained distinctions (Hodges et al., 1995; Lambon Ralph et al., 2017).

1.3.2 Neural correlates of semantic memory, episodic memory

The molecular mechanism underlying memory and learning function at a fundamental level is thought to be long-term potentiation. This is the strengthening of synapses between neurons based on activity transmission between them (Kandel et al., 2014).

The region most crucial for LTM is the MTL: patients with lesions in the MTL, show distinct impairments in memory (Squire et al., 2004). In episodic memory encoding, the hippocampus in particular is implicated in the encoding of items and their relationships (Davachi & Wagner, 2002) and binding items to their context (Hannula & Ranganath, 2008). Sparse coding has been suggested as a hallmark feature of the MTL, with only a few neurons responding strongly to a few specific items (Quiroga et al., 2008; Quiroga, 2012; Wixted et al., 2018). This sparsity can help with pattern separation during memory retrieval, as it allows activity for different items to be

distinct and easier to differentiate (Lohnas et al., 2018; Rolls, 2013; Urgolites et al., 2022). Another line of research suggests that the neural patterns in MTL at encoding are reinstated during retrieval (Danker & Anderson, 2010; Morcom, 2014; Rissman & Wagner, 2012; Xue, 2018).

In order to encode complex events there is evidence that the hippocampus interacts with the PFC and other neocortical areas during encoding to control which information should be selected for encoding and retrieval (Backus et al., 2016; Cooper & Ritchey, 2019; Parker et al., 2002; Simons & Spiers, 2003; Summerfield et al., 2006; Wagner et al., 2005). Specifically, during memory recollection and familiarity are supported by the anterior medial PFC and the anterior and dorsolateral PFC, respectively (Yonelinas et al., 2005), and retrieval success is tracked by the parietal "old/new" ERP (Rugg, 1995).

In contrast to the MTL, the anterior temporal lobe (ATL) has been regarded as pivotal for semantic memory, due to the severe impairments of category knowledge in semantic dementia patients, whose major neuropathological and neuroimaging marker is atrophy in the ATL (Mummery et al., 2000). Converging evidence surfaced from fMRI studies that show activations for semantic memory tasks in the ATL (Binney et al., 2010; Visser et al., 2009) as well as TMS disruption of ATL activity increasing semantic difficulties in healthy individuals (Pobric et al., 2007).

It has been suggested that the ATL area functions as an amodal hub for semantic representations, integrating modality-specific information from sensory brain regions into a coherent concept (Binder & Desai, 2011; Lambon Ralph et al., 2010; Lambon

Ralph et al., 2017; Patterson et al., 2007). This idea is corroborated by a decoding study that found that using brain activity from the ATL to train machine learning models on activity induced by one sensory modality could also decode activity from another modality successfully (Fairhall & Caramazza, 2013). While this body of work firmly places importance on the ATL for the storage of concepts, there is also evidence that it interacts with the PFC when cognitive control is required, for example when making decisions about semantic material (Hoffman et al., 2015).

In summary, LTM enables storage of large amounts of information over long periods of time, and marked by diverse functions, supported by distinct mechanisms and neural correlates, including the MTL and ATL and interactions with PFC and parietal areas. Whether semantic or episodic memory, this contrasts with the flexible, capacity-limited nature and frontoparietal neural correlates of WM.

1.3.3 Formal models: Hippocampal, semantic and RL models

Numerous models of episodic memory emphasise the hippocampus, explaining its role in both rapid encoding and incremental learning (Gluck et al., 2003; Norman et al., 2008; Schapiro et al., 2017; Sučević & Schapiro, 2023). As in WM, the hippocampus is often modelled as an attractor network, inspired by Hopfield networks that support pattern completion (Bruce et al., 1987; Rolls, 2010, see also Chapter 5). Many computational models build on complementary learning systems theory, which proposes that the hippocampus encodes episodic memories that are later consolidated

into neocortical structures (Helfer & Shultz, 2020; McClelland et al., 1995; Schapiro et al., 2017; Squire, 1992; Sun et al., 2023), bridging episodic and semantic memory.

Many semantic memory models are network-based, linking sensory inputs to an amodal hub or using backpropagation in distributed systems to map concepts to features (Kumar, 2020; McClelland & Rogers, 2003; Rogers et al., 2004). Recent linear deep learning models even mirror semantic development in children (Saxe et al., 2019). However, such models struggle with abstract concepts lacking sensory ties. In contrast, co-occurrence models capture abstract meaning from text (Landauer & Dumais, 1997), but lack grounding in perception. A recent connectionist model bridges this gap, showing that concepts can acquire sensory features indirectly through associations with perceptually grounded concepts (Hoffman et al., 2018).

Beyond associative learning, many models draw on RL, rooted in human reward learning. A pivotal neural correlate is the reward prediction error (RPE). This midbrain dopamine activity is caused by discrepancies between the expected and received reward which guide learning through linking outcomes with preceding states (Schultz et al., 1997). RL formalises this process computationally, with agents learning to associate states of the environment and actions based on reward feedback (Sutton & Barto, 1998). This has been effective in explaining human behaviour in different tasks, and constitutes another form of long-term knowledge.

1.4 Integrating Visual Working and Long-term Memory

The question of whether WM and LTM are distinct has sparked considerable debate. While previous sections described their functions in isolation, in this section I will highlight the reasons for traditional and modern distinctions, drawing on concepts from the previous sections and employing direct comparisons. I will then review evidence for the two systems interacting.

1.4.1 Dissociations, re-associations and candidates for integration

Despite the classical Atkinson-Shiffrin model positing an interdependence between the two (Shiffrin & Atkinson, 1969), WM and LTM have historically been studied separately. An early dissociation arose due to a study of a neurological patient with a left temporoparietal lesion who had impaired STM performance but intact LTM performance, leading researchers to assume that STM was not necessary for LTM (Shallice & Warrington, 1970). However, in recent years several investigations have shed doubt on this dissociation, not least due to the MTL presenting an overlap in neural correlates of both WM and LTM (Jonides et al., 2008; Ranganath & Blumenfeld, 2005). The interactions between these two systems have therefore come into sharper focus (Rhodes & Cowan, 2019; Schurgin, 2018).

The classical Embedded Process Theory proposed that WM is simply activated LTM (Cowan, 2001). This has since incorporated the focus of attention as part of WM, which in turn is subsumed within LTM as a whole (Cowan, 1999; Hakim et al., 2020). In a similar vein, one recent study asked whether activity-silent WM might be considered

LTM (Beukers et al., 2021), hypothesising that if it were, bringing it back into the focus of attention would leave long-term traces long after the WM task. Generally, theories which assume WM to be part of LTM have struggled to explain the active mechanism of WM maintenance, in particular of novel information.

More recently, a number of studies have called into question the distinction that WM is a store for extremely precise memory, reporting that visual LTM has the same level of fidelity (Brady et al., 2013), although another study showed higher precision for WM than LTM, with precision deteriorating with increasing time (Biderman et al., 2019). LTM may produce the same precision as WM if given enough repetitions (Miner et al., 2020), however, is constrained by the initial encoding precision in WM, typically resulting in lower-precision LTM representations (Forsberg, Adams, & Cowan, 2021). For example, holding four items in WM is possible at a higher precision than encoding four items into LTM at once (Forsberg, Adams, & Cowan, 2021). Together, these findings indicate that visual WM still holds a special role for representing *novel* high-precision information within a limited capacity store.

WM and LTM have been contrasted in the realm of RL, too. RL tasks require agents to link states, actions, and rewards, and are sufficiently solved by WM at set sizes within WM capacity limit. At set sizes below WM capacity, RPEs are blunted, resulting in less long-term recall (Collins, Ciullo, et al., 2017a). At higher set sizes however, when WM capacity is exceeded, the tasks are solved with RL, marked by higher RPEs and greater long-term retention (Collins, 2018). These studies established a competitive relationship between WM and RL where, rather than interact, they trade

off and complement one another. This, however, neglects the possibility that they might rely on one other in certain circumstances (see Chapter 4).

In contrast to these renewed dissociations, overlapping mechanisms between WM and LTM have also been pointed out. For example, similar costs were associated with gating perceptual information into LTM and WM when switching between the two, suggesting a shared mechanism for selecting information to load into WM (Verschooren et al., 2021). Another study compared fMRI signals of visuo-spatial information in matched WM and LTM tasks, reporting sensory-like representations in occipital and parietal cortex, and abstract representations in lateral parietal cortex in both tasks—consistent with shared mechanisms. Notably, parietal cortex activity distinguished between WM and LTM tasks, whereas occipital regions did not (Vo et al., 2022).

Overall, differences in neural mechanisms and behavioural signatures, and their isolation under specific conditions, challenge the idea that WM and LTM are the same system. However, the involvement of certain brain regions in both functions demonstrates their overlap. As described earlier, the PFC plays a role in both active WM maintenance and LTM encoding and involvement of the MTL and hippocampus in episodic memory and WM binding, indicating shared neural mechanisms between systems. The following sections adopt the view that WM and LTM are specialised, complementary systems that interact.

1.4.2 Influence of prior knowledge on working memory

In this section, I will review literature on the involvement of LTM in WM focusing on three different mechanisms and their potential neural underpinnings: (1) the retrieval of information from LTM to aid the WM of an item or the allocation of resource towards specific items, (2) the involvement of LTM due to stimulus familiarity or meaningfulness of the WM memorandum itself, and (3) the alteration of the WM representation due to grouping or chunking.

Compensatory recruitment of long-term memory during working memory tasks A number of studies on verbal and visual WM have found that LTM can be recruited as an additional resource when WM capacity is exceeded. In particular, a number of studies mixed novel items with previously studied associative memories.

For example, in an investigation on verbal WM, participants initially learned word pairs (Bartsch & Shepherdson, 2022). Crucially, the addition of novel pairs, but not learned pairs, affected WM performance (Bartsch & Shepherdson, 2022). Extending this effect to visual WM, another study found that only when WM capacity was exceeded, subsequent items suffered from proactive interference, a hallmark of LTM (Bartsch et al., 2024). Following up on this, another investigation suggested a specific contribution of episodic LTM to the binding of word-picture pairs in WM. When WM capacity was exceeded at set sizes larger than three, swap errors increased and performance was marked by proactive interference, but was immune to distractors, as would be expected in WM (Bartsch & Oberauer, 2023). Overall, these studies suggest when

items are retrieved into WM from LTM, they do not take up space in WM, suggesting LTM may be recruited as an additional resource to aid WM processes.

Importantly, these LTM recruitment effects seem to be of a supporting rather than interfering nature, thus still suggesting segregation of WM and LTM as sources of feature information. When using WM arrays of items that matched or mismatched previously associated colours (Oberauer et al., 2017), WM performance was better in the matched condition than in the novel and mismatched conditions which were on par. In addition, there was no proactive interference, suggesting that this is indeed a WM rather than an LTM effect (Oberauer et al., 2017). This suggests a supportive role of LTM rather than a trade-off, which would predict that information from LTM would interfere in the mismatched condition. However, since WM has to store at least a link to the information in LTM in order to retrieve relevant information, it remains unclear how this segregation might be achieved.

A contrasting finding comes from a study with detailed contexts: when words were linked to scenes that either matched or mismatched the context in a subsequent WM task, fMRI data showed episodic memory intrusion when the original context was retrieved. Instead of the new item, the previously associated item was recalled, leading to swap errors (Hoskin et al., 2019). This suggests that the interference-less recruitment of LTM might be limited to simple stimuli.

The recruitment of LTM during WM may depend on strategic allocation of WM resources. When participants learned colour-location associations before a spatial WM task requiring report of all items, they showed greater overall precision but tended to

report novel items before familiar ones, while they performed worse when forced to report novel items last (Bruning & Lewis-Peacock, 2020). This suggests participants strategically prioritised unknown items, knowing they could "off-load" recall of familiar ones to LTM.

Little is known about the brain regions supporting this type of LTM recruitment, although some older imaging studies give some insight. Ranganath et al. (2004) used category-selective regions—fusiform face area (FFA) and parahippocampal place area (PPA) to examine WM for faces, houses, and face-house pairs. In a delayed match-to-sample (DMTS) task, activity in category-selective regions reflected the type of information maintained. When required to retrieve previously studied pairs, increased activity in inferior temporal cortex and hippocampus suggested LTM support during WM maintenance (Ranganath et al., 2004). Applying multivariate pattern analysis in a similar task, Lewis-Peacock and Postle (2008) extended this finding showing that LTM can support WM through temporary activation. Participants completed a delayed paired-associate WM task, in which they judged whether a probe matched the previously studied associate of the item held in WM. Rather than the encoded WM item, delay-period activity reflected its associate, possibly in preparation for matching (Lewis-Peacock & Postle, 2008).

These studies suggest that brain regions usually implicated in LTM can be recruited during the WM delay period. Conceptually this is related to the sensory-recruitment hypothesis, except rather than sensory areas the tasks recruited brain regions in inferior temporal cortex (FFA, PPA). While in the summarised studies engage-

ment of LTM was required by the task, it suggests that similar mechanisms might be strategically employed when information from LTM is non-essential but beneficial.

Effects of pre-existing and lab-acquired familiarity Studies reviewed in this category examine cases in which the stimulus itself is recognisable or familiar. They use one of three approaches: (1) comparing unrecognisable objects to recognisable objects or object categories humans are "experts" in, such as faces and scenes, (2) contrasting items participants are already familiar with versus unfamiliar ones, such as native versus foreign alphabets, and (3) inducing familiarity through training in the lab.

Historically, WM tasks have used abstract stimuli to create "process-pure" tasks isolating WM from other cognitive systems. Even at the basic feature level, familiarity or category effects can also be found. Specifically, colours that belonged to a prototypical colour category (i.e. "blue") were remembered more frequently, despite all target colours being equally likely (Bae et al., 2015). WM can represent this information as mixture of categorical and continuous representations simultaneously (Hardman et al., 2017, see also Chapters 3 and 6).

To boost ecological validity, researchers have started exploring WM for real-world versus abstract objects and scaling abstract WM models to real-world contexts (Bates et al., 2024; Orhan & Jacobs, 2014; Orhan et al., 2014). Some have found that capacity for real-world objects is greater than for abstract features, and feature memory improves when remembered as part of a real-world object (Brady & Störmer, 2020a; Chung et al., 2023), most likely due to semantic information rather than visual

complexity (Torres et al., 2023). Some studies also suggest that real-world objects benefit from additional time and sequential presentation, allowing for "deeper" encoding (Brady & Störmer, 2020b; Brady et al., 2016), although this has failed to replicate (Li et al., 2020).

Moreover, research has shown a memory benefit for faces, a domain of human expertise (Curby & Gauthier, 2007). For instance, Asp et al. (2019) reported higher performance and increased neural delay activity for meaningful stimuli compared to ambiguous Mooney or scrambled faces. Whether participants perceived the stimulus as a face was the key factor, suggesting recruitment of additional neural capacity (Asp et al., 2019).

A substantial line of evidence comes also from studies that show that WM performance is increased when remembering familiar stimuli within a category. For faces, WM capacity is greater for famous compared to unfamiliar faces, but inverting the face negates this effect, suggesting that low-level features are not the cause of this familiarity effect (Jackson & Raymond, 2008). This echoes findings from real-world objects. This finding extends to stimuli other than faces: participants familiar with specific Pokémon stimuli them remembered more items in a WM change detection task than those unfamiliar (Xie & Zhang, 2018). EEG data from the same task showed that CDA amplitude tracked familiarity at shorter, but not longer, encoding times, suggesting that familiarity led to a larger number of maintained items (Xie & Zhang, 2018).

Familiarity can also aid access to features if the entire object is retrievable. In one study using the US flag, familiarity improved performance only when detecting

the whole flag, not its features, suggesting a hierarchy between object and features (Huang, 2011). Similarly, training participants on abstract shapes improved change detection performance for trained versus novel stimuli, but only when training focused on discrimination (Blalock, 2015), implying this is a task-based, non-generalising effect.

One cognitive mechanism that might underlie these effects is a higher encoding rate. When comparing English and Chinese alphabets as familiar and unfamiliar items respectively in a change detection task, this seems to be due to a higher encoding rate as well as a higher capacity group for familiar versus unfamiliar stimuli rather than their complexity (Ngiam, Khaw, et al., 2019).

In terms of neural correlates, learning has been shown to occur in the PFC in many NHP studies (Miller & Constantinidis, 2024). For instance, in a spatial DMTS task, PFC showed greater plasticity evidenced by increased firing rates during the WM delay, enhanced colour selectivity, and greater stimulus information (Riley et al., 2018), suggesting an attunement of PFC to task demands. A reanalysis of the same dataset revealed that intrinsic timescales (from biophysical properties) increased, while task-induced "seasonal" timescales decreased from posterior to anterior subregions across dorsal and ventral PFC (Trepka et al., 2023). Additionally, the proportion of neurons selectively exhibiting these timescales also increased (Trepka et al., 2023).

Similar plasticity in has been shown to be arise in human PFC neurons as a result of stimulus-specific training. In one study, three participants completed 16 sessions of a DMTS task involving repeated sequences of patterns, resulting in clear performance improvements (Miller et al., 2022). fMRI revealed that with training item-specific

representations emerged over time in voxel activity patterns across PFC. This suggests that not only can LTM be recruited during WM tasks, but WM-related areas can also adapt to the task (see also Chapter 6). However, it remains unclear what determines the use of additional resources versus modification of WM representations.

One interpretation of these effects is that training improves neural efficiency. When training humans on Chinese characters and artificial patterns, the training enhanced performance only for trained items (Zimmer et al., 2012). Neural activity in superior intraparietal sulcus and middle occipital cortex generally increased as a function of set size. It plateaued after two items in untrained items, but only after three for trained items. In line with parietal cortex being sensitive to set size increments, the authors interpreted this as a sign of neural capacity efficiency for trained items (Zimmer et al., 2012).

In addition to familiarity, LTM may be recruited in the form of semantic knowledge. Participants better remembered already chunked object pairs when they were semantically related (O'Donnell et al., 2018). Performance further improved when items interacted in a functionally congruent way or followed real-world spatial regularities, e.g. a watering can pointing at a plant (Kaiser et al., 2015; O'Donnell et al., 2018). This may be supported by concurrent visual and semantic neural representations in temporal areas. In a DMTS task using iEEG, Liu et al. (2020) showed that WM can maintain semantic representations at the same time as holding visual ones. These developed at time points during encoding, with the earlier period corresponding

to visual representations and the later period to semantic representations (Liu et al., 2020).

In summary, the described findings demonstrate that recognisability, whether through expertise, prior familiarity, or induced learning, may enhance WM performance through more efficient encoding and a richer, semantically structured representation. This may be driven by plastic changes in PFC and engagement of parietal cortex.

Efficient representations through chunking and compression Chunking describes the effect of storing an association of features or objects leading to capacity savings in WM. Chunks can be formed on the basis of information from LTM, such as the example of "FBI" being easier to remember than a random three-letter combination, but also by grouping novel combinations of items, such as chunking phone numbers into multi-digit groups (Thalmann et al., 2019). Assessing the chunkability of stimuli is important for correct assessment of WM capacity as it increases the amount of items that can be stored (Mathy & Feldman, 2012).

As opposed to the recruitment effects, chunking has been thought to rely on the re-coding of information into a more capacity-efficient representation within WM, a so-called memory "compression" effect. Brady et al. (2009) examined how statistical regularities help compress memory information. Participants viewed arrays of four colour conjunctions. Simple covariance between colours improved performance over time, however, when the pattern was removed performance returned to baseline. Crucially, the benefit applied to all items, not just repeating ones, showing that information was compressed, and as a result freed up space for more items (Brady et al., 2009).

The effect has been replicated and found to rely on explicit memory of the association (Ngiam, Brissenden, & Awh, 2019). As previously described, LTM can support WM without interference, meaning that the compression effect does not necessarily rely on re-coding, but may instead result from building familiar information in LTM, recruited during the WM task (see also Chapter 2).

Compression effects extend to rules in any stimulus dimension, not just colour. Chekaf et al. (2016a) used stimuli conducive to compression in one dimension, such as large white and black squares classifiable by size. Recall was better in the rule condition, with capacity around three items in unstructured lists but increasing linearly with added structure. Mathy et al. (2018) used the same stimulus set in a simple (simple maintenance) and complex (interference through task-unrelated processing) WM span tasks, finding that both tasks benefited from compressibility, suggesting chunking is resistant to interference (Mathy et al., 2018).

A major discussion in chunking research is whether compressed or chunked representations are "content-free" or contain visual details of their components. In a study similar to Brady et al. (2009), Huang and Awh (2018) examined colour conjunctions and found greater capacity in patterned versus controlled blocks, with effects increasing over time. However, longer reaction times and reduced performance under response deadlines suggested that feature information was not stored in WM but decoded from another store, such as LTM (Huang & Awh, 2018). This suggests that the WM representations may have acted as abstract links. In contrast, Allen et al. (2021) showed that chunks are not content-free, relying instead on the capacity within WM to represent

items hierarchically associated with detailed content. This was evident when learning two associated features with one varying dimension, as LTM chunks did not interfere but enhanced memory detail (Allen et al., 2021).

It is still unclear to what extent the compression effect in WM is similar to the familiarity effect, and whether it neurally corresponds to representational shifts in PFC (Miller et al., 2022). Moreover, it is unknown how it relates to additional recruitment from LTM. One possibility is that a hierarchical representation in WM allows for incorporation of the corresponding LTM chunk only when it is similar enough, not when it would interfere. This would be similar to a summed activation proposed to underlie ensemble representations (Robinson & Brady, 2023) and consistent with only helpful recruitment from LTM.

1.4.3 Working memory as a platform and bottleneck for long-term memory

WM contributions to LTM encoding The early evidence for WM having an influence on LTM came from Hebb's repetition effect: in list learning effects, simply repeating the list increases the span of items one can remember (Hebb, 1961). In the Atkinson and Shiffrin (1971) multi-store memory model, STM was a necessary store for information to move on to LTM. Modern studies have found several effects of holding information in WM and its effect on LTM.

As summarised in the previous sections, statistical regularities in WM tasks can be learned and over time lead to a performance benefit. A secondary finding from these

studies provides insights into the role of WM in learning. Overall, it seems that the learned traces are very fickle and only form under certain circumstances, at least when multiple items are involved.

An early investigation of this showed that when visuospatial arrays were sometimes repeating, a learning effect only emerged when a specific item changed on every trial or the same target location was probed (Olson et al., 2005). This suggests encoding alone is not sufficient for learning. Initially, Colzato et al. (2006) had found faster reaction times for repeated bindings than for non-repeated bindings, indicating some retention. However, there was no trace of the repeated bindings in a subsequent memory test, leading to the conclusion that temporary memory and learning are distinct mechanisms (Colzato et al., 2006).

On the contrary, Logie et al. (2009) found that when repeating the same arrays of colour-shape conjunctions participants did notice this repetition of bindings, but only when they occurred on every trial and only for bound objects. Repetition of single features however relied on cued recall and was weaker. Equally, interleaving repeated arrays with novel arrays erased any trace with no learning following (Logie et al., 2009). In a follow-up study, Shimi and Logie (2019) showed that performance was immediately above chance for repeated arrays, but further improvement only occurred in the group who later reported being aware of the repetition, similar to Ngiam (2023). They found that asking participants to reconstruct the array led to more robust learning (Shimi & Logie, 2019).

However, is it not only binding within objects as in Logie et al. (2009) that helps form LTM representation. Chunking repeated sequences of two out of six stimuli regardless of their position within a list also helped participants remember. However, this was only the case as long as they were able to recognise the repeated structure. This again points to awareness as being a key driving factor (Musfeld, Dutli, et al., 2024; Ngiam, Khaw, et al., 2019).

Musfeld, Souza, and Oberauer (2024) demonstrated another case of retrieval practice affecting learning in WM. They found that in a change detection paradigm, participants were only able to learn repeated arrays under high retrieval demands, i.e. when they were asked to recall all items on every single trial, but not when recalling one item or using a recognition test. This full-report procedure is extremely effective, yielding improvements in WM performance after a single repetition, including when WM is stretched beyond capacity at six to eight items (Adam et al., 2023).

Taken together, these findings suggest that items must be complex enough to prompt awareness and strong learning effects, unless repetitions are explicitly drawn attention to. Items may also need to be sufficiently distinct, which can be achieved through adding feature dimensions. While retrieval practice in WM plays a key role, it is unclear whether it aids pattern recognition on subsequent trials or if retrieval itself strengthens memory traces (see also Chapter 2). Another explanation for why learning is difficult is that WM capacity limits the ability to encode items in LTM. Yet, this does not rule out an influence of WM on LTM formation. In fact, maintenance appears critical, as early and late WM stages affect LTM differently. For example, early task-

irrelevant interference and early probing in a DMTS task impaired later LTM retention more than late interference (Bergmann et al., 2013).

In terms of neural correlates, both PFC and MTL support the role of WM in LTM formation. Dorsolateral PFC and hippocampal activity during the delay period predicted successful LTM encoding, which is disrupted by early WM interference (Ranganath et al., 2005). Further evidence for dlPFC involvement came from comparisons between conditions in which participants had to maintain versus actively manipulate information, showing that manipulation was more strongly linked to LTM formation than retention alone (Blumenfeld & Ranganath, 2006). Additionally, when participants maintained both the number of letters in a word and its meaning in a DMTS task, activation in the parahippocampal cortex during WM predicted subsequent LTM (Axmacher et al., 2008). The parahippocampal cortex has also been implicated in WM maintenance of novel information (Hasselmo & Stern, 2006), suggesting a role in selecting information from WM for the formation of new LTM traces.

Altogether, these findings demonstrate that encoding of information into WM does not automatically qualify it for a transformation into LTM traces. Rather, strong environmental structure combined with cognitive engagement such as awareness, and retrieval practice, and active WM maintenance characterise the conditions under which information is successfully transferred to LTM.

WM as a bottleneck for LTM encoding While the previous section establishes the contribution of WM to LTM encoding, here I summarise studies that show that this connection is also characterised by limits akin to the ones in WM capacity.

WM limitations seem to directly reduce the amount of information stored in LTM. Forsberg, Adams, and Cowan (2021) found that higher WM set sizes led to lower LTM accuracy in a DMTS task, in a way mirroring WM set size effects in LTM. In other studies, LTM also shows lower baseline accuracy (Forsberg et al., 2022, 2023). Crucially, instructions did not reveal to participants that they would be tested on the WM items later, suggesting that this was an incidental effect. An open question is whether WM would also constrain LTM encoding if participants focussed on it intentionally. Subsequent investigations found that this effect is present at both detailed and gist levels of memory (Greene et al., 2024).

In a similar fashion to encoding, retrieval from LTM is limited by the capacity of WM. Fukuda and Woodman (2017) found that WM acts as a buffer for LTM retrieval. Using coloured stimulus grids, they compared encoding of novel arrays with retrieval of previously learned ones and found identical capacity patterns for retrieval and encoding tasks, with alpha power plateauing after three encoded or retrieved items (Fukuda & Woodman, 2017). These findings suggest that WM provides a platform for LTM retrieval, which is therefore limited by the same capacity limit as WM. In contrast, Liu et al. (2022) showed that LTM retrieval can bypass WM: even when WM was engaged beyond its capacity limit, participants were able to retrieve from LTM. EEG showed alpha suppression peaked with WM load but increased further when LTM items were also retrieved (Liu et al., 2022). This suggests retrieval can extend beyond WM limits. One explanation is that Fukuda and Woodman (2017) assumed a WM capacity of four, but prior familiarity with LTM items in the task may have boosted this limit.

Together, these findings show that while WM plays a crucial role in both encoding into and retrieval from LTM, its capacity limitations shape LTM representations, but do not categorically restrain retrieval from LTM.

1.5 Thesis Scope

Despite the large body of research discussed above, many questions remain about the mechanisms underlying the interplay between WM and LTM. How does the brain utilise prior information for encoding and storage in WM, and how does this change the representational format of items, both at a behavioural and neural level? What role does WM play in the formation of long-term memories?

In my thesis I set out to contribute to each of the outlined dimensions, effects of WM on LTM and vice versa, as well as to further our understanding of the characteristics of these memory systems from normative and empirical perspectives. It contributes to the introduced themes in the following ways.

Chapter 2 focuses on the compression effect, trying to recreate it in a continuous delayed-estimation paradigm. It relates to the debate about binding and object-based representations in asking whether, if exposed to the same feature conjunction, this would lead to permanent binding and an object-based representation and entail decreased resource consumption. Its findings also reiterate the difficulty of creating LTM traces in common WM paradigms.

In **Chapter 3**, we devised a novel paradigm to investigate the representation underlying binding a continuous feature to a unique real-world object over short and

medium intervals. The Chapter contributes to our understanding of the differences in characteristics between WM and LTM, and on using real world objects in WM tasks. It also contributes to the research on feature-object binding, investigating how performance develops when a continuous features is bound to a unique object over time, and when it might be "handed off" from WM to LTM.

Chapter 4 explores effect of maintenance of information in WM on learning using an RL learning task combined with a visual WM paradigm. This ties into researching the effect of active storage of WM on learning.

In **Chapter 5**, I run simulations using a WM buffer in conjunction with a Hopfield network as proxies for WM and LTM, respectively, to show how properties of these modules trade-off in high- and low-load memory environments. This study provides an implementation to the fundamental question of why WM and LTM exist in their current states.

Finally, in **Chapter 6**, I analyse a macaque dataset where the monkeys had to maintain colours drawn from either the full colour reel versus a restricted colour wheel. The focus of the Chapter is on whether performance benefits and behavioural biases following the restriction of colour space are detectable in the next testing session, i.e. whether they involve long-term traces. It demonstrates how that forms of learning in WM exist that persist beyong the normal duration of WM. It also contributes to the question about categorical representations as monkeys learn a long-term representation of the colour space while solving a WM task

2 Investigating the Working Memory Compression Effect in a Continuous Delayed Estimation Paradigm

2.1 Abstract

Prior research has found an interaction between statistical learning and working memory (WM) such that repeated conjunctions take up less space in WM, thereby freeing up space for novel items - the so-called compression effect. However, it is unclear how the underlying representation differs between novel and compressed items and how much space is saved. Here, we examined whether repeatedly encountered items ("learned items") would exhibit enhanced performance and reduce the cognitive load for other items in a delayed estimation task. Across three experiments, participants remembered features of visual stimuli (colour and orientation), with one item repeating consistently in patterned blocks. We hypothesised that space savings could occur via a more object-like representation, evidenced by increased response error correlations for object features.

Results revealed a consistent benefit for the learned item, particularly in terms of orientation accuracy, but no evidence of improved performance for non-learned items, suggesting that the learned item did not free up additional WM capacity. Surprisingly, most participants failed to notice the repeating item, even under explicit instruction in Experiment 3. While the small group of "learners" on average showed better performance than "non-learners", some non-learners performed at the level of learners. This suggests that the observed benefits may involve implicit mechanisms.

The lack of explicit learning happening complicates drawing conclusions on compression effects in WM. Instead, the minimal explicit awareness of repetition raises questions about the conditions necessary for statistical learning to influence WM. The findings contribute to prior research showing the difficulty of creating long-term traces within WM.

2.2 Introduction

The world around us is largely structured, with objects and their features following statistical patterns. Our cognitive abilities allow us to make use of this in order to achieve adaptive behaviour. For example, being stung by a bee should ideally lead to avoiding bees, but also generalising this behaviour to similar insects like wasps. In support of this, perceptual discrimination and generalisation, as well as working memory (WM), have been found to be attuned to the world around us (Sims, 2016, 2018).

WM is highly flexible and can be used to actively hold old and new information in the mind, but it is highly limited (Cowan, 2001). It has been suggested that structuring our representations in WM according to learned information also allows for more efficient use of capacity. For example, more familiar visual objects can be encoded at a higher rate and maintained in WM with higher accuracy (Ngiam, Khaw, et al., 2019). Provided there is enough processing time, there is a benefit of holding real-world objects in memory compared to novel items (Brady & Störmer, 2020b; Brady et al., 2016) and there is evidence from EEG decoding that WM holds semantic-like as well as perceptual representations (Liu et al., 2020).

What mechanisms are used to store already known items in WM to maximise capacity? Some have suggested a process similar to the sensory recruitment hypothesis (Adam et al., 2022; Xu, 2017), which posits that WM draws on other brain regions to represent sensory features. In this case, WM would utilise existing knowledge in long-term memory (LTM) to encode information in WM (Norris et al., 2019). Another

line of research has suggested that frequently seen combinations of features achieve this benefit by allowing for more efficient encoding (Brady et al., 2008).

These information-theoretical approaches have claimed that the performance benefit of familiar stimuli is due to "compression", i.e. assigning more frequent items a shorter code in memory (Norris et al., 2019). For example, Brady et al. (2009) presented participants with a visual WM task in which they had to remember two-colour items and report a colour cued by location. Markedly, they introduced statistical regularities in the form of covariance of colours, leading some colours to be paired more often than others. In the course of the experiment, participants in this condition were able to store more items in WM than in a control condition where colours were paired at random. The authors interpreted that rendering items compressible by co-occurrence enabled participants to use a shorter code for them, allowing for more efficient storage (Brady et al., 2009). This effect has since been replicated and found to rely on the explicit knowledge of feature conjunctions (Ngiam, Brissenden, & Awh, 2019). However, this contrasts with Olson (2006), who found that even when participants were able to identify repeated arrays at the end of the task, this did not benefit their change detection performance throughout.

The compression account also predicts that compressed items should lead to longer response times, since information has to be decoded at retrieval. Huang and Awh (2018) found that chunked information was impaired when restricting participants to short response deadlines. This suggests that additional information is present, but takes time to access. This is challenged by the finding that chunks from LTM do

not interfere with incoming WM, but rather combined, suggesting that they are in an accessible state (Allen et al., 2021).

Another aspect to this line of research is how WM contributes to learning LTM traces. Using change detection tasks, LTM traces are fragile and accumulate slowly. In order to learn an association, it seems that the same feature-feature conjunction or feature-location conjunction has to be present on each trial. For example, several early studies investigating this showed arrays of squares where half of the configurations repeated (Olson & Jiang, 2004), but this did not lead to any performance improvement. Similarly, repeating one feature of an object while the other varies does not produce a learning effect (Logie et al., 2009). It seems that retrieval practice plays an important role in facilitating learning. For example, Shimi and Logie (2019) reported slow learning in a change detection task and rapid learning when arrays had to be reconstructed. Similarly, the effect can equally be found when all items are tested on each trial (Musfeld, Souza, & Oberauer, 2024). These results point towards either enhanced awareness of the repetition through retrieval or learning itself being prompted by retrieval.

In summary, a beneficial effect of prior information on WM performance has been established, while the resulting representational format remains unclear. It is unlikely that the brain uses a binary code to represent information like a computer. For example, visual WM has been shown to make use of both continuous and categorical representations. With increasing set size, more items are stored in a categorical fashion than in a continuous manner (Hardman et al., 2017). This suggests that representational formats in WM – and perhaps also in compression – changes according to task

demands. Another open question is the development of a compression code. Compression increases linearly with compressibility in verbal WM (Chekaf et al., 2016b), but this has not been examined in visual WM.

If compression entails different codes, representations for learned associations should be different. Several known properties of WM could play a role in this. For example, while the format of WM is debated (Luck & Vogel, 1997; Wheeler & Treisman, 2002), it is known that it can flexibly represent either features or bound objects. Generally, storing more features per item decreases performance (Palmer et al., 2015), however, this cost is reduced when features are combined into the same object part (Fougnie et al., 2013; Xu, 2002). One mechanism to compression could thus be that representing frequently appearing conjunctions as objects takes up less WM capacity. Learning to reliably combine feature associations into objects could therefore play a role in the compression effect.

The current study therefore adapted the original compression task (Brady et al., 2009) to a continuous delayed estimation paradigm. Instead of using colour-colour conjunctions, this paradigm used colour-orientation conjunctions and enables investigation of whether stimuli are encoded object or feature-based and whether this changes biases that are present in WM tasks (Bae, 2021). Participants learn one colour-orientation association in the form of an arrow through encountering it on every trial in an array of varying number of objects.

In replication of Brady et al. (2009), we expected learning prior information about items to result in better performance. If there is a true compression effect, there

will be a benefit for non-learned items when presented with learned items. If an item can be off-loaded completely, the performance should be equal to the lower set size, i.e. adding a learned item should make performance at set size four the same as at set size three. If a learned item still requires a pointer, however, stored information will exceed capacity at higher set sizes and prior information should have no effect. In addition, these effects will emerge over time, if they are due to participants learning long-term associations between features and in line with previous research this improvement should depend on a LTM trace as measured by correct response of the learned features after the experiment. If learned items are stored in a more object-based representation, potentially referred to by a pointer, their error correlation should be higher than if they were stored solely in WM.

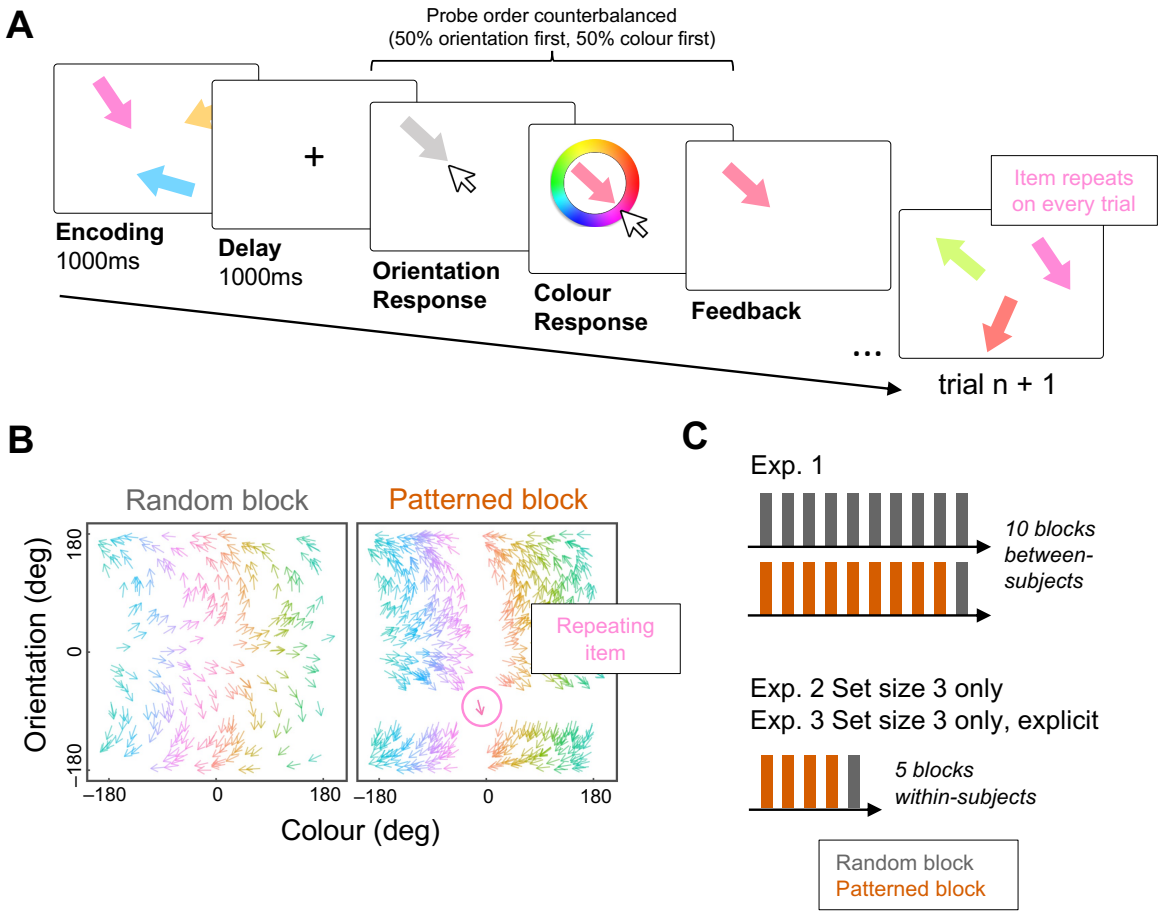
2.3 Experiment 1

2.3.1 Methods

Participants & Materials 38 participants were recruited from the Oxford community via the Sona Participant Management System (sona-systems.com). They participated in the experiment in person at the Department for Experimental Psychology. An additional 12 participants were recruited on Prolific (prolific.com) and completed the experiment online. All participants were fluent English speakers and were compensated for their participation. Five participants were excluded due to chance-level responses, leading to a final sample of N = 45 (33 female, M_{age} = 25.73, SD_{age} = 5.24). All participants had normal or corrected-to-normal vision and three of them

were left-handed. All experiments were conducted with the approval of the Research Ethics Committee and all participants gave informed consent.

Figure 2.1
Delayed estimation task with statistical learning



(A) Participants were presented with three to five items and, after a 1 s delay, reproduced two features (colour and orientation) of one of the arrows cued by location. They were shown feedback at the end of the trial. **(B)** There were two types of experimental blocks. In random blocks, colours and orientations were picked pseudo-randomly for each trial. In patterned blocks, one item, chosen individually for each participant, consistently maintained the same colour-orientation conjunction but appeared in different locations. The remaining items were selected pseudo-randomly. Feature values were constrained to be 30 degrees apart. **(C)** Succession of block types for each experiment. Orange indicates a patterned block, grey a random block.

Procedure The offline version of the task was presented using MATLAB (The MathWorks Inc., 2001) and Psychtoolbox 3 (Kleiner, 2007). The online version was

programmed using PsychoPy (Peirce et al., 2019) and hosted on Pavlovia (pavlovia.org). Participants were instructed to sit 60cm, or an arm's length, away from the monitor (1980×1080 pixels offline, variable sizes online).

The task was a combined statistical learning task and delayed estimation WM task (task sequence shown in Fig. 2.1A). On each trial, after a 1 s ITI, an array of three to five arrows was presented for 1 s. Participants were instructed to remember the colour, orientation, and location of each arrow. After a delay of 1 s, one of the arrows was probed by location and participants used the mouse to reproduce colour by selecting from a colour wheel and orientation by moving the mouse and confirming with a click. They were instructed to focus on accuracy over reaction time and to provide their best guess in case they could not remember. At the end of each trial, participants were shown feedback in the form of the correct arrow. The experiment consisted of 10 blocks of 36 trials each. After the main experiment, participants were asked whether they noticed a specific arrow repeating and to reproduce the colour and orientation "arrow they saw most frequently" three times.

Stimuli and design The experiment manipulated the structure in the stimulus feature space. Participants were either allocated to the "Patterned" condition, which contained structure (repetition of an item) in the stimuli presented, or the "Random" condition (Fig. 2.1B). The Patterned condition contained nine Patterned blocks where one item repeated on every trial, making its features (colour and orientation) predictable. It was chosen to be only one item to assess the effect of potentially being able to save a whole item. This "Learned" item was selected for each participant by

sampling feature (colour and orientation values) from the circular normal (von Mises) distribution. Its location was randomised across trials. The "Learned" or the "Non-learned" item were probed at random ("Probe Type"). The last block contained only random stimuli. In random blocks, features (colour, orientation, location) were chosen randomly from the circular normal (von Mises) distribution. To avoid interference between stimuli, colour and orientation values were constrained to be at least 30 degrees apart in both conditions and presented at least 45 degrees apart on an invisible circle.

In both conditions, set size (three, four, five) was manipulated within subjects. Response Order and set size were counterbalanced within each block. In the Patterned condition, the learned item was probed at random across trials. To ensure an equal number of trials in which the learned item was probed across all set size conditions, trials from set size 3 were up-sampled, and trials from set size 5 were down-sampled to match the number of trials in set size 4. Per participant, this resulted in 81 trials in which the learned item was probed overall, and 243 trials in which the non-learned item was probed. The response angle and reaction time (RT) were collected for colour and orientation responses.

Analysis Trials with an RT greater than three times the participant's SD on each feature were removed from the analyses. Response error was calculated as

$$error = \theta_{response} - \theta_{target}, \tag{1}$$

where θ is the colour angle in degrees, and absolute response error as

$$error = |error|. \tag{2}$$

Precision was calculated as

$$\frac{1}{SD(error)}, \tag{3}$$

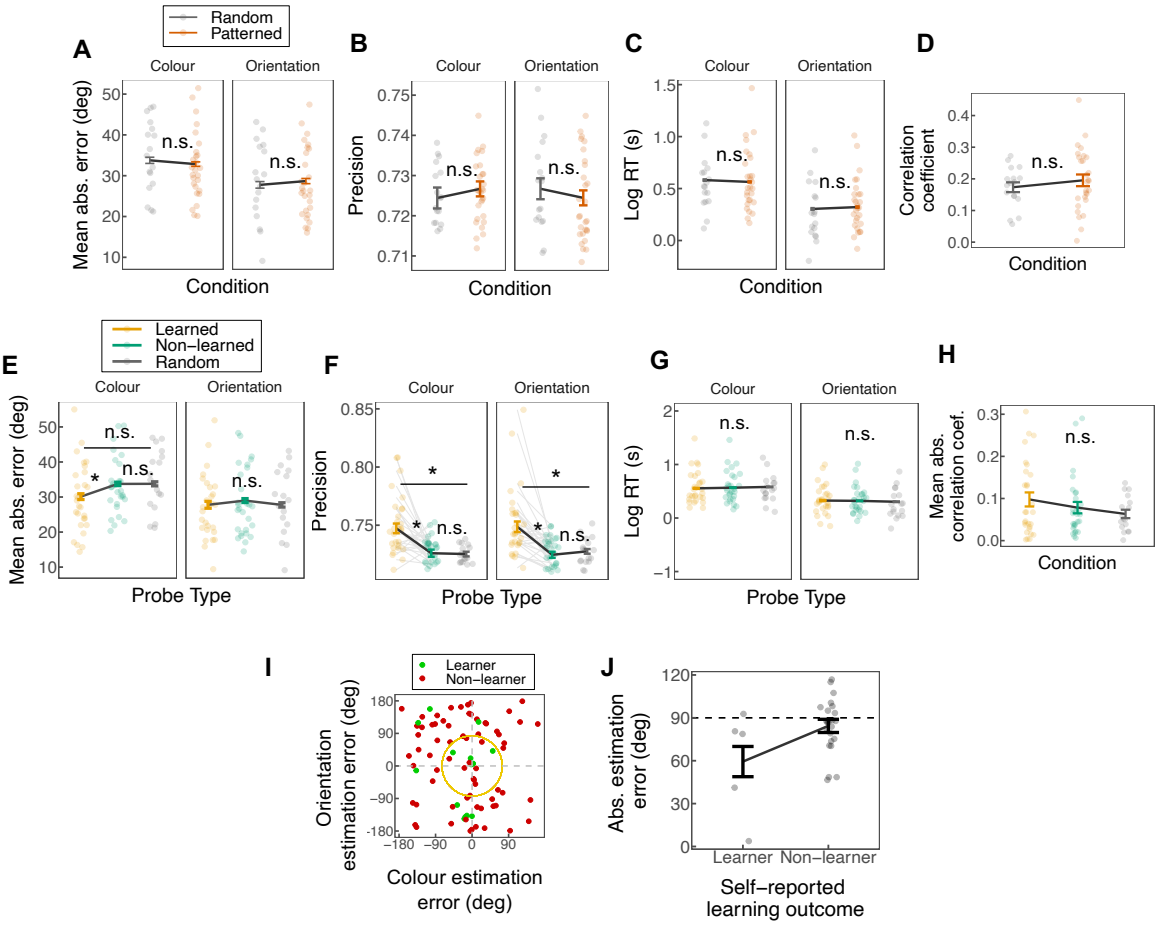
and RTs were log-transformed. Pearson correlations were calculated between colour and orientation response error. Patterned and Random groups were compared using unpaired t-tests. Two-way ANOVAs were performed to examine the effects of Set size and its interaction with Condition on all dependent variables. Trials in which the non-learned item was probed were compared with trials in which the learned item was probed using paired t-tests. Both trials were compared to random trials from the Random condition using unpaired t-tests.

2.3.2 Results

Results for the first experiment are shown in Figure 2.2.

Effects of Condition There were no significant differences between the Patterned and Random condition for colour response error ($t(31.66) = 0.89$, $p = .38$), orientation response error (Fig. 2.2A), colour precision ($t(36.24) = –0.42$, $p = .70$), or orientation precision ($t(31.79) = 1.03$, $p = .31$)(Fig. 2.2B). Equally, reaction times did not differ between groups (Colour $t(36.61) = –0.34$, $p = .73$, orientation $t(31.70) = –0.80$, $p = .0.43$) (Fig. 2.2C), and neither did correlation coefficients (t = –0.90, df = 42.68, $p = 0.37$). In line with previous research, there were set size effects for colour ($F(1.73, 74.41) = 200.96$, $p < .001$) and orientation ($F(1.42, 61.26) = 200.96$, $p < .001$), but no interaction effects with Condition ($F(1.73, 74.41) = 0.39$, $p = .651$) or orientation $F(1.42, 61.26) = 0.26$, $p = .699$).

Figure 2.2
Results from Exp. 1



Effects of condition on **(A)** mean absolute error, **(B)** precision, **(C)** RT, and **(D)** Error correlation. Effects of Probe Type on **(E)** mean absolute error, **(F)** precision, **(G)** RT, and **(H)** Error correlation. Points represent participant means and error bars represent SE. **(I)** Estimation of the repeated features at the end of the experiment. Participants estimated both features three times. Since the three responses diverged, points represent individual choices are shown instead of an aggregate score. Yellow circle represents chance level (90 degrees). **(J)** Estimation error compared between self-reported Learners and Non-Learners.

Effects of Probe Type Within the Patterned condition, colour response error was significantly lower for learned than for non-learned items ($p < 0.001$). This was, however, not the case for orientation error ($p = .52$). Neither of the patterned conditions were different from items in the Random condition (all $p > 0.5$, Figure 2.2E). Results for the Precision metric were similar, with learned items leading to higher colour precision

than non-learned items ($p < .001$) and random items ($p < .001$) . Contrary to the error results, orientation error was also lower for learned versus unlearned items ($p < .001$) and random items($p < .01$). There was no significant difference between non-learned items and random items for colour or orientation (both $p > .05$, Fig. 2.2F). There were no effects of Probe Type on reaction times for colour ($p = 0.24$), nor orientation ($p < 0.65$, Fig. 2.2G). The correlation between colour and orientation error was not significantly different between Probe Types ($t(42.93) = –0.89$, $p = .38$, Fig. 2.2H).

In the Patterned condition, five participants indicated that they noticed an arrow repeating, 23 indicated they did not. While the number of "Learners" (people who noticed the repetition of an item) was not high enough to conduct formal analyses, participants' estimation of the learned colour and orientation do not seem to fully depend on their explicit knowledge of item repetition (Fig. 2.2I): some participants who indicated they did not learn were fairly close to the true values, while others who said they did learn showed a larger estimation error. On average, the "learners" seemed to have lower estimation error, as was expected (Fig. 2.2J).

2.4 Experiment 2

Results from Experiment 1 revealed a consistent benefit for learned items in terms of colour memory, while the benefit for orientation memory was limited to the precision measure. The direction of this effect was as expected. However, there was no benefit for the non-learned items in the Patterned condition, which indicates that there was no space freed up in WM through saving resources on the learned items.

In other words, no true compression effect was found. There was also no significant difference in the correlation of colour and orientation errors, although the direction is as expected, with correlations slightly higher for learned items. There were no significant differences between conditions either. This is likely due to learned and non-learned probe types averaging out in the Patterned condition and diminishing any possible difference, especially since there was only one learned probe being presented together with three to five non-learned probes depending on set size.

Only a minority of participants noticed a repeating arrow, and this did not seem to be related to their ability to reproduce the correct colour and orientation. Therefore, it cannot be determined whether awareness is a prerequisite for creating the benefit. This could be due to the fact that most trials were set size five trials (see Methods) and items interfered too much with each other within and between trials for the repeating arrow to be registered.

Experiment 2 attempted to address these issues by lowering the set sizes to arrays of three items only to see if that would achieve a consistent effect.

2.4.1 Methods

Procedure and stimuli were the same as for the online participants of Experiment 1. N = 43 participants took part in the experiment online (26 female, M_{age} = 29.07, SD_{age} = 4.56, 2 reports missing).

Design In order to test whether the lack of a compression effect was due to overcrowding of the stimulus array, this version of the experiment reduced the set size

to three only. Further, the experiment was conducted purely within-subjects, in order to ensure power. Every participant completed four patterned blocks, followed by one random block (Fig. 2.1C), with 20 trials per block.

Analyses Effects of Probe Type were assessed with one-way ANOVAs for colour and orientation dependent variables separately. The Greenhouse-Geisser correction was applied where the sphericity assumption was violated.

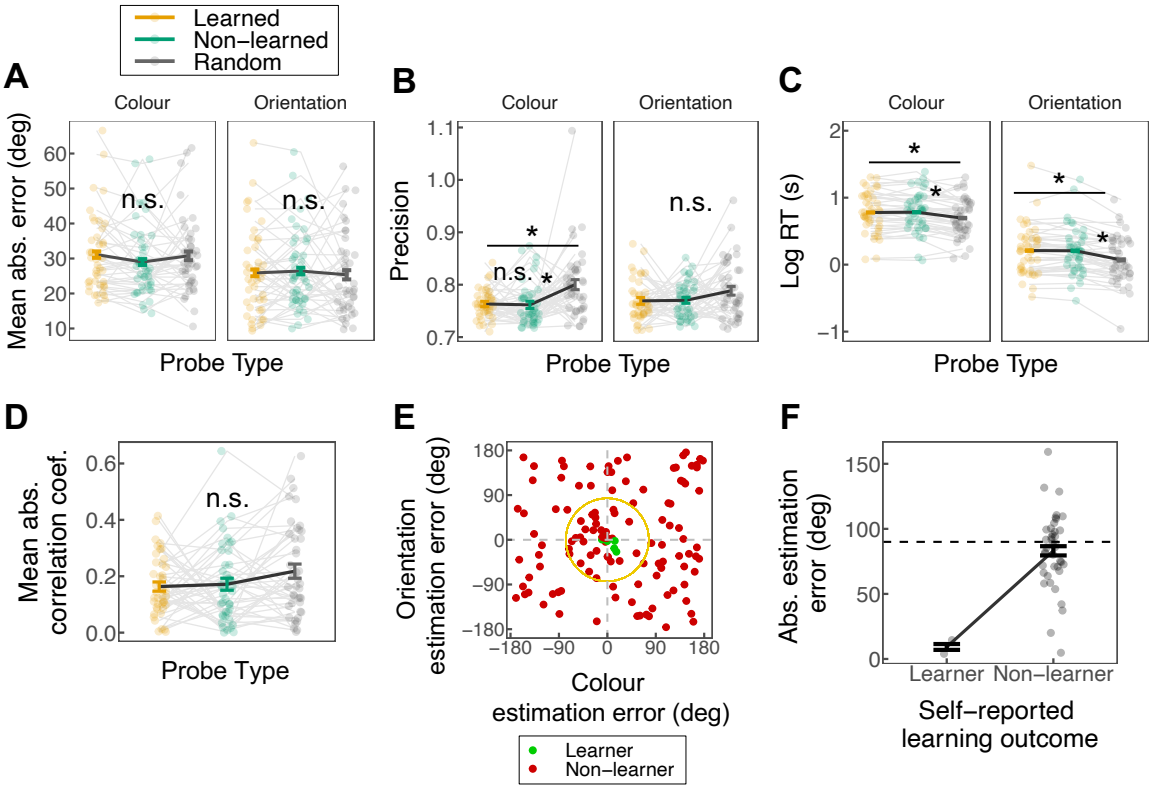
2.4.2 Results

There were no significant effects of Probe Type on either colour ($F(1.83, 75.22) = 1.01$, $p = 0.36$) or orientation error ($F(1.71, 69.93) = 0.002$, $p = 0.737$) (Fig. 2.3A). There was an effect of Probe Type on colour precision ($F(1.50, 61.60) = 7.68$, $p < 0.01$), with post-hoc t-tests revealing a that the random probe was better remembered than the learned ($p < 0.01$) and non-learned ($p < 0.01$) probes. However, this effect remained non-significant for orientation precision ($F(1.83, 74.92) = 2.20$, $p = 0.122$) (Fig. 2.3B). Crucially, there were effects on reaction times for both colour ($F(1.55, 63.36) = 13.47$, $p < 0.001$) and orientation ($F(1.42, 58.34) = 25.65$, $p < 0.001$), with post-hoc analyses indicating that random probes were different from learned ($p < 0.001$) and non-learned ($p < 0.001$) probes (Fig. 2.3C). There was no difference in error correlation between probe types ($F(1.83, 75.22) = 2.11$, $p < 0.13$).

Two participants indicated that they noticed the repeating arrow and could reproduce it (Fig. 2.3 E). As in Experiment 1, some participants who indicated that

they did not notice the repeating arrow were still above chance in reproducing the colour and orientation of the arrow (Fig. 2.3E&F).

Figure 2.3
Results from Exp. 2



Performance was the same for learned and non-learned items in terms of **(A)** mean absolute error, **(B)** precision, **(C)** RTs. **(D)** Colour and orientation errors were similarly related in both conditions. **(E)** Only two participants reported that they noticed the repeating arrow, which was related to their estimation of the repeated features at the end of the experiment, points represent individual choices rather than aggregate scores and the yellow circle represents chance level. **(F)** Estimation error compared between self-reported Learners and Non-Learners. Points are subject means and error bars represent within-subject SE.

2.5 Experiment 3

Results from Experiment 2 showed effects on precision and reaction times, which was an effect contrary to the expected one. Participants performed better for *random* probes than learned or non-learned probes. This may have been due to a general

practice effect, as the experiment was within-subject with only one random block at the end of the experiment. No effects were found that point to a benefit of the learned or the non-learned item. This could again be due to only two people noticing the repeating arrow, despite the lower set size, suggesting there may still have been too much interference between stimuli. Another possible reason might be that there was no incentive in this experiment to detect a pattern in the stimuli, as WM may not have been taxed sufficiently at three items.

Instead of relying on participants noticing the pattern, Experiment 3 explicitly instructed them that one of the presented arrows would repeat on every trial.

2.5.1 Methods

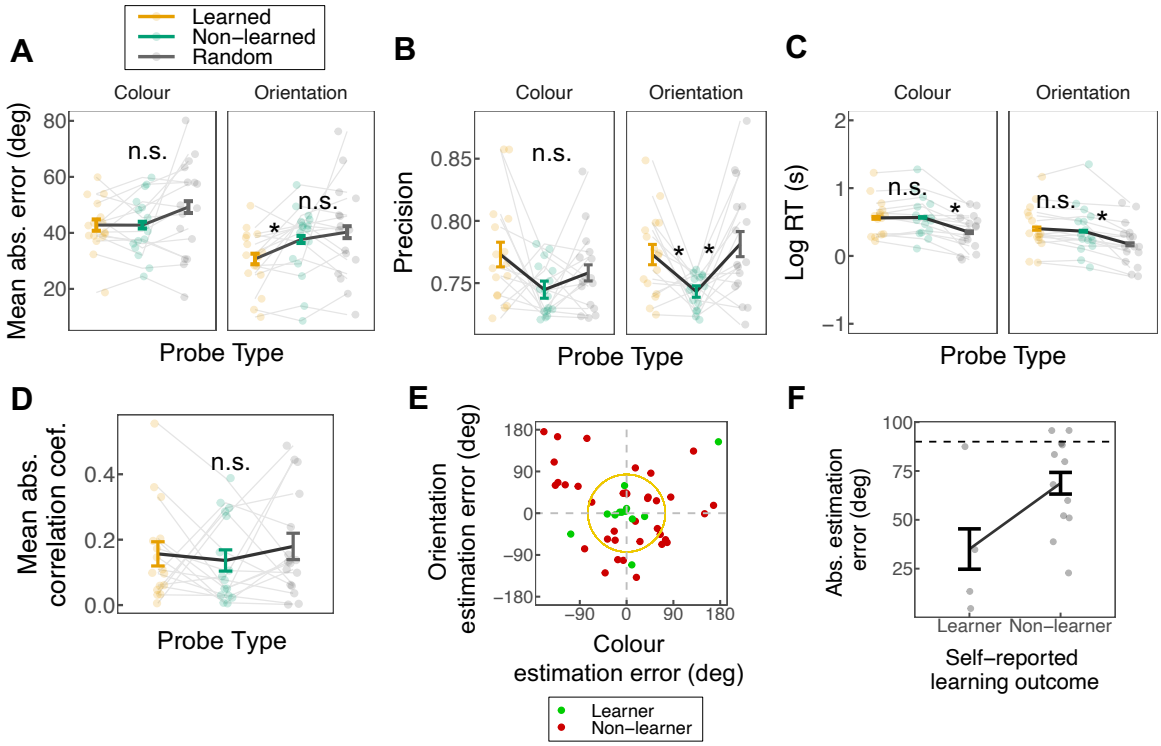
24 participants were recruited on Prolific and completed the experiment online. Seven participants were excluded due to chance-level responses, making the total sample of N = 17 (10 female, $M_{age} = 29.59$, $SD_{age} = 4.72$).

In Experiment 3, performance on the random blocks experienced a general training effect since it was completed in the end of the experiment and does thus not provide a fair comparison with the other conditions. A possible solution would be to return to the between-subjects design in Experiment 1. However, this would require more participants to achieve the same power. Instead, we focused on getting participants to notice the repetition to create the learning effect and thus opted for the same within-subjects design as in Experiment 2.

Stimuli were the same as described in Experiment 1.

2.5.2 Results

Figure 2.4
Results from Exp. 3



Memory performance for orientation was better for learned compared to non-learned items as shown by **(A)** lower error and **(B)** higher precision. **(C)** RTs were faster for random items in the last blocks than both learned and non-learned items in patterned blocks. **(D)** Error correlations were the same between conditions. **(E)** Five participants reported noticing the repeating arrow, but this was not reflected in their estimation of the repeated features at the end of the experiment. Points represent individual responses rather than aggregate scores and the yellow circle represents chance level. **(F)** Estimation error compared between self-reported Learners and Non-Learners. Points are subject means and error bars represent within-subject SE.

There was a significant effect of Probe Type on orientation error ($F(1.55, 23.23) = 3.78$, $p < 0.05$), with post-hoc tests revealing a trend towards learned probes being better remembered than non-learned probes ($p = 0.052$), but not random ($p = 0.10$) probes. The effect on the colour error remained non-significant ($F(1.46, 21.95) = 2.84$, $p = 0.09$) (Fig. 2.4A).

Similarly, the effect of Probe Type on orientation precision was significant ($F(1.70, 25.47) = 6.06$, $p = .01$), with post-hoc tests indicating that learned probes ($p < 0.05$) and Random ($p < 0.05$) probes were remembered more precisely than non-learned. In this case, there was a trend towards learned probes having better colour precision ($F(1.72, 25.77) = 3.06$, $p = .07$, Fig. 2.4B).

There was an effect on colour ($F(1.22, 18.34) = 19.24$ $p = < 0.001$) and orientation reaction times ($F(1.39, 20.80) = 18.15$ p < 0.001), with Random probes leading to shorter RTs than learned and non-learned respectively (all $p < 0.01$) and no difference between learned and non-learned ($p = 0.54$ and $p = 1$, Fig. 2.4C). The effect on correlation coefficients also was non-significant ($p = 0.67$, Fig. 2.4D).

Despite explicit instruction about the repeating arrow, only 4 out of 20 participants reported noticing it. Again, this was not related to whether their estimation at the end of the task was correct (Fig. 2.4E&F).

2.6 Discussion

This study combined a statistical learning paradigm with WM to assess whether there would be a benefit for the learned item such that they take up "less space" (compression and lead to better performance for the non-learned item as well). Experiment 1 overall showed a benefit for the learned item, but not for the non-learned item, compared to random probes. However, only a small number of participants reported learning the repeating item or could report its features accurately. Experiment 2 focussed on improving learning by decreasing the number of items in the memory array,

but this did not change the outcome. In Experiment 3, participants were specifically instructed which item they would repeatedly see. This led to better performance for the learned item than the non-learned item, but not better than the random item, and only for orientation. Despite explicit instruction, only the minority of participants reported seeing the repeating arrow. No experiment showed evidence for a more object-based representation for the learned item.

One aim for this study was to replicate the compression effect (Brady et al., 2009; Ngiam, Brissenden, & Awh, 2019) in a continuous delayed estimation paradigm, hypothesising that if learned items took up less space, non-learned items might be represented at a more granular level. However, Experiments 1 and 3 found a benefit for the learned item, but not for the non-learned items. The learned item did not free up additional space, suggesting a mechanism other than more efficient representation.

The improved performance for the learned item might be due to it being held in a prioritised state compared to non-learned items. For example, it might be held in the focus of attention (Griffin & Nobre, 2003). This would, however, normally require a trade-off of resources and likely lead to worse performance for non-learned items, which was not the case. Additionally, this would presumably make it more likely for participants to make a connection between repeating items, which was not the case. Another explanation is that a pointer is stored in WM that cues retrieval of features, for example, from LTM. The fact that this did not seem to be related to participants' awareness of the repeating arrow suggests that it is an automatic process. However,

due to the lack of "learners" in all experiments, this relationship cannot be ruled out entirely.

Crucially, results did not show improved performance for the non-learned item. There are a few possible explanations for this. Firstly, the pattern was only one repeating arrow, which may mean that any compression effect was simply too small to detect. However, the study was powered to detect the benefit for the learned item, which makes this explanation doubtful. Alternatively, it might be that only the true compression effect depends on explicit knowledge (Ngiam, Brissenden, & Awh, 2019), while the benefit for the learned item functions on an implicit level. Interestingly, although participants denied noticing the repeating arrow, providing explicit instructions in Experiment 3 appeared to enhance performance on the learned item, specifically in terms of orientation. However, if these mechanisms indeed differ in their degree of implicitness, it is counter-intuitive that encoding into WM (non-learned items) would depend more on awareness than the involvement of a distinct memory system, such as LTM (learned items).

The absence of a strong learning effect complicates drawing conclusions about the remainder of the hypotheses since the absence of effects could have been produced by the weak learning effect. For example, there was no effect on the correlations of colour and orientation error, which could mean that it depends on awareness of the repeated item, but this cannot be confirmed. Equally, without a strong effect it was difficult to assess its emergence over time and analyse the long-term learning associa-

tions between features. In terms of set size, there were no differences in Experiment 1 and the manipulation was dropped in further experiments.

Participants failing to notice the arrow repeat is nevertheless an interesting incidental finding itself, and allows for different interpretations. Firstly, it adds to previous research that suggests a learning effect produced through repetition (Hebb, 1961) is difficult in traditional WM paradigms. For example (Musfeld, Souza, & Oberauer, 2024), found that participants had to be probed on every item in a change detection array to produce a learning effect, suggesting that repeated retrieval and not encoding leads to learning. This could explain why participants did not reach awareness in this study as only one item was probed on each trial and it was the repeating one only a fraction of the time. However, Experiment 1 still showed an effect despite this difficulty. It may be that the continuous paradigm was more sensitive on an individual trial basis than the one in Musfeld, Souza, and Oberauer (2024). This goes hand in hand with another explanation, which is that the pattern may have been too weak to notice. In Brady et al. (2009) stimuli were repeated 80% of the time, with new combinations appearing only 20% of the time. In contrast, in this study, the repetition rates were 20%, 25%, and 33% for set sizes 5, 4, and 3, respectively. This means that both repetition and retrieval occurred much less frequently, even in Experiments 2 and 3, which were limited to a set size of 3. A related reason is that there may have been too much interference between items within and across trials, further compounded by the similarity of items, their simultaneous presentation and the position of the repeated arrow changing.

These issues might be addressed by requiring report of every item on every trial or strengthening the learnable structure by increasing the number of repeating items. Another option would be to add a learning phase to the experiment where participants first have to learn associations to criterion to ensure an explicit representation. The disadvantage of this would be that representations in WM during learning could not be analysed, which would be possible with stronger structure.

Observing these difficulties reinforces a longstanding view on the role of WM in cognition at large. It seems that while it interacts with other systems and shows biases rooted in our surroundings, it is not a specialised learning system. It excels in the flexible representation of task-relevant information, and accumulating LTM traces relies on very specific circumstances, such as retrieval practice (Musfeld, Souza, & Oberauer, 2024).

In conclusion, there seems to be an implicit benefit for repeating items in a delayed estimation WM paradigm. However, this did not lead to capacity savings, as evidenced by the lack of a benefit for random feature associations. Participants were extremely unlikely to notice the repetition, possibly due to the flexible nature of WM. The paradigm may be remedied by a stronger structure or an initial training phase in order to produce explicit effects.

3 Characteristics of Working and Long-Term Memory – Evidence from the Any-back Task

3.1 Abstract

Although a great deal is known about the characteristics of long-term memory (LTM) and working memory (WM) separately, it remains unclear what happens when both contribute to task performance. In this study, we investigated how discrete object information and continuous feature information are remembered over short and longer delays when faced with interference. We present the "Any-back" task, a novel paradigm that combines a recognition judgment with a delayed estimation task. Participants viewed a stream of 200 unique images in specific colours (Exp. 1) or orientations (Exp. 2). Interleaved with these encodings of new items were other interfering trials (0–29, corresponding to retention intervals of 1.5–90 s), in which participants had to retrieve an item, reject a lure, or complete a distractor task. At retrieval, participants reported whether they had seen an item before (yes/no) and reproduced the respective feature (on continuous, analogue response spaces). A subset of participants also completed a test battery which assessed general cognition and cognitive domains, including processing speed, flexibility, and short-term memory.

Results revealed that after interference from only one intervening trial, in particular when it was a target retrieval, there was a sharp increase in colour memory error. Performance was sustained above chance-level performance after additional interference from interleaved items, but without any further increase in error. Cluster analyses indi-

cated that a categorical response pattern with few categories might underlie this steep performance decline. Object memory, on the other hand, declined continuously and gradually, and was most susceptible to interfering lures. Moreover, a larger number of items already held in memory was beneficial for recall at longer retention intervals but detrimental for shorter ones. Object memory showed strong correlations with external short-term memory validation tasks and processing speed tasks, whereas correlations for colour memory were fewer and domain-general. Most of these effects were not observed on the orientation task.

Results and limitations are discussed in light of the characteristics of WM and LTM, the possible transfer of information from WM to LTM, and psychometric properties of the Any-back task.

3.2 Introduction

Long-term memory (LTM) is characterised by its large capacity to store information over extended timespans (Brady et al., 2008). In contrast, working memory (WM) is severely limited and operates over short timespans (Luck & Vogel, 1997). Traditionally, paradigms have tried to prevent LTM encoding by using abstract features to isolate WM. There is now a developing interest in studying real-world objects to understand memory processes in more ecologically valid settings. Beyond the comparison of real-world and abstract objects, a central question is how continuous feature information is represented when bound to a discrete real-world object (Chung et al., 2023). In this study, we introduce a novel experimental paradigm to examine how memory for object and feature information develops when maintained over short and intermediate durations. The central aim was to identify behavioural characteristics of WM and LTM contributions to memory performance under interference.

Research has shown that real-world objects lead to better WM performance than abstract ones, probably due to semantic information (Brady & Störmer, 2020a; Torres et al., 2023). As a result, they may benefit from additional encoding time or "deeper" encoding time, such as being presented sequentially (Brady and Störmer, 2020b; Brady et al., 2016, although see Li et al., 2020). Recently, Chung et al. (2023) found that the memory benefit for real-world objects extends to the features they are bound to: capacity for colours was greater when they were remembered as part of recognisable compared to unrecognisable items. Nonetheless, it remains unclear where the benefit

comes from and whether it persists over intervals that exceed the typical capacity of working memory.

One explanation for the benefit for real-world objects is that it incentivises feature information to be encoded in a more coarse-grained format in WM. Continuous variables may be stored as continuous, categorical, or a mixture of both (Bae et al., 2015; Hardman et al., 2017). Colours tend to be remembered more frequently as shades that belonged to a prototypical colour category, such as "blue", despite all target colours being equally likely (Bae et al., 2015). In other words, reported colours were biased towards category centres. Bae et al. (2015) suggested that this type of response pattern may best be explained by a model using a dual-encoding process into continuous and categorical channels and responding with a mixture. Similarly, Hardman et al. (2017) showed that a Gaussian mixture model, which included a term for categorical information, fit the data best. Contrary to Bae et al. (2015), they found that assuming categorical and continuous information to be independent, rather than integrated, yielded a better fit. This suggests that participants choose one of the information sources when responding. Using their model, they also showed participants stored one item in a continuous fashion, perhaps corresponding to the focus of attention (Griffin & Nobre, 2003). The contribution of continuous representations tended to decrease with increasing set sizes (Hardman et al., 2017). Despite these findings, none of the WM studies using real-world objects have assessed the format of storage for continuous information.

Does the binding to discrete information, such as real-world objects, cause feature information to be encoded in a more categorical fashion, thereby leading to performance benefits? Alternatively, it could be encoded in a continuous fashion first, and re-encoded into LTM in a categorical fashion thereafter. Such a transfer of information from WM to LTM has been previously suggested (Forsberg, Adams, & Cowan, 2021), with the underlying brain function assumed to involve the medial temporal lobe (MTL) due to its role in both WM and LTM tasks (Ranganath, 2006; Squire et al., 2004). The hippocampus in particular has been shown to be involved in high-precision memory binding (Olson, 2006; Yonelinas, 2013), as well as conditions of high cognitive load, when the hippocampus may play an increased role (Jeneson & Squire, 2012; Jeneson et al., 2012). In addition, the parahippocampal region appears to process novel information (Schon et al., 2016), possibly providing a marker for information that has not yet been encoded into LTM.

These findings offer a potential neural point of connection for a "hand-off" from a WM to an LTM representation. If LTM is already involved in the task (Lewis-Peacock & Postle, 2008; Ranganath et al., 2004, 2005), this transition might occur during encoding, i.e. information might be encoded in WM and LTM simultaneously. However, in tasks where LTM is not necessary, it is less certain when, or if at all, this hand-off takes place. One possibility is that information is encoded in LTM as soon as it leaves the focus of attention and becomes "activity-silent" (Beukers et al., 2021; Stokes, 2015). Active WM maintenance is supported by persistent neural activity in

the MTL (Kamiński et al., 2017), thus creating a signature for encoding in longer-term traces in this area.

Here, we developed a new paradigm to address some of the gaps in current research. The task was designed to allow for contribution from both WM and LTM to object and colour memory during short and longer delays, while using established measurements for both. Participants were tasked with sequentially viewing a stream of objects that varied in features, such as colour. A critical feature of the paradigm was the inclusion of interference: In between encoding and retrieval, items held in memory incurred interference from up to 29 memory-relevant and irrelevant items. At retrieval, they first gave an old/new recognition response, a method commonly used in episodic memory studies (Yonelinas, 2001). Second, a delayed estimation response was employed, allowing for the fine-grained measurement of colour memory precision (Bays & Husain, 2008). Furthermore, the experiment introduced a novel measure: the number of items retained in memory. The continual nature of the task allowed for the assessment of the number of items participants held in memory at any given moment.

The analysis was guided by three main questions. First, we were interested in differential markers of LTM and WM. Our goal was to investigate how object and colour memory performance would change over time and whether they would be impacted differentially by interference. We expected colour memory, as a proxy for WM, to degrade faster and experience a larger impact from interference than object memory (LTM). Additionally, we expected that object memory would be affected by the total number of items in memory. The second question was whether a "hand-off" between

memory systems could be pinpointed. We hypothesised that this would produce a sudden, not gradual, shift in our measures. For example, if WM is assumed to be more precise than LTM, we would expect a decrease in feature precision that could signal a hand-off to LTM. Another potential marker for this transition might be the emergence of categorical feature information in memory. Finally, we examined the psychometric properties of this new paradigm, assessing its external validity against tasks from a cognitive battery.

3.3 Experiment 1

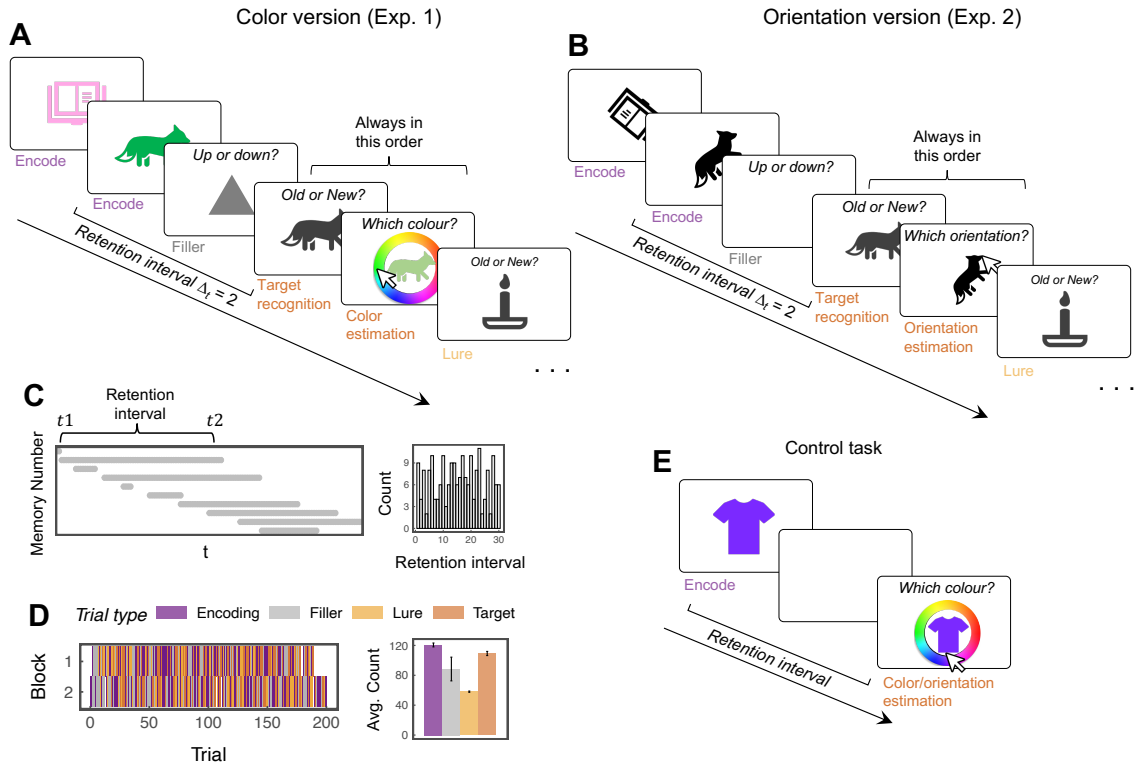
3.3.1 Methods

Participants 65 participants were recruited on Prolific and completed the experiment online. Nine participants were excluded due to chance-level responses, leading to a total sample of N = 56 (27 female, M_{age} = 28.72 (SD_{age} = 4.19, 5 reports on demographics missing). Out of these, all participants performed the main experiment (Any-back task) and 21 participants additionally performed a control version of the experiment and the validation tasks.

Experimental design: Any-back task The Any-back task was a memory task designed to test object and feature memory (Fig. 3.1A). It was programmed using MATLAB (The MathWorks Inc., 2022) for stimulus generation and PsychoPy (v2022.2.4, Peirce et al., 2019) for experiment presentation.

In the task, participants were confronted with a stream of events. There were four event types. On *Encoding* events, participants were shown an object of a specific

Figure 3.1
Any-back task schematic for colour and orientation versions



(A) Any-back task, colour version. Participants had to remember around 200 coloured objects across delays of 1.5 s to 66 s. They viewed a stream of trials with an ITI of 1.5 s. There were four trial types. On Encoding trials, participants had to encode a memory; on Retrieval trials, they were asked whether they had seen an item before and to then retrieve its colour, always in this order. On Lure trials, they were presented with an item they did not see. Filler trials asked a simple perceptual judgment question. **(B)** Any-back task, orientation version. Everything was the same as the colour version, except that participants had to remember item orientations instead of colours. **(C)** Left panel: Depiction of the task in abstract terms. Memories had an encoding time, *t1*, and retrieval time, *t2*, with the retention interval in between. Right panel: Retention intervals were distributed uniformly. **(D)** Left panel: Example of trial types in one experiment with two blocks. Right panel: Average counts for different trial types for the complete experiment across participants. **(E)** Control task without interference. Participants viewed a coloured object and had to maintain it for a delay matched to the duration of 1–5 interfering trials. They then reproduced the object's colour. The colour version is shown, but there was an orientation version to match the orientation experiment.

colour, which they were instructed to remember. On *Retrieval* events, participants were instructed to retrieve the object with the prompt "Have you seen this item?". Regardless of their response (Yes or No), they were then prompted to retrieve the object's colour from memory and select it from a colour wheel. On *Lure* events, participants were asked

the same question ("Have you seen this item?"), but were shown an object they had not seen before. Finally, on *Filler* events, they were asked to judge whether a triangle was pointing upwards or downwards. Filler trials were included to fill gaps between other trials, but also served as an attention check.

There were two experimental blocks per experiment. The distribution of trial types in each block was carefully balanced, so that trial types were evenly spaced (see below). This was designed to create a "steady state" environment, where the proportion of trial types remains constant throughout the block.

Each memory stimulus was associated with an encoding and a retrieval time, which specified the trial it was presented and retrieved at (see schematic in Fig. 3.1C, left panel). The time in between, the retention interval, was uniformly distributed between 1 and 30 items, with a retention interval of 1 item having no interfering items and a retention interval of 30 having 29 interfering items. Inter-trial intervals were 1.5 s long. This meant retention intervals on average ranged from 1.5 s (interval 1) to 66 s (interval 30), taking into account participants' average response times.

Prior to testing, trial event timings were selected programmatically and separately for each participant as follows. First, an encoding time was sampled without replacement from all possible encoding times, with remaining available times tracked in a list. Then, for one memory at a time, a retention interval was sampled from a uniform distribution, with $\alpha = 1, \beta = 30$ determining the range. Adding the retention interval to the encoding time point produced the potential retrieval time point. A list of all possible retrieval time points was maintained which was used to check whether the

time point was still available. If it was already taken by another memory's encoding or retrieval time, a new retention interval was sampled. This was repeated until a suitable retrieval time was found. After all memories were assigned an encoding and retrieval time, we checked whether the distribution of retention intervals was still uniform across the block. This was done using the Kolmogorov-Smirnov-Test, a non-parametric test that quantifies the distance between two probability distributions. The test returns a significant p-value when the distributions are different. We compared the sampled memory times against an empirical distribution of 10,000 samples from the same uniform distribution. We repeated the entire process until p-value > 0.9, robustly indicating no significant difference between the two distributions. Different distributions were generated in this way for each participant. An example of an empirical distribution of retention intervals is shown in Figure 3.1C (right panel).

To include all event types, empty spaces were first assigned to Lure trials, except for the first five trials as no memories would have been encoded, making the answer to a Lure trial a certain "No". We then added Filler trials to all remaining event slots. Images and colours were assigned to memory and lure items at random. Finally, we removed all Filler trials before the first encoding trial as well as after the last relevant item was retrieved from the block as they would unnecessarily prolong experimental time.

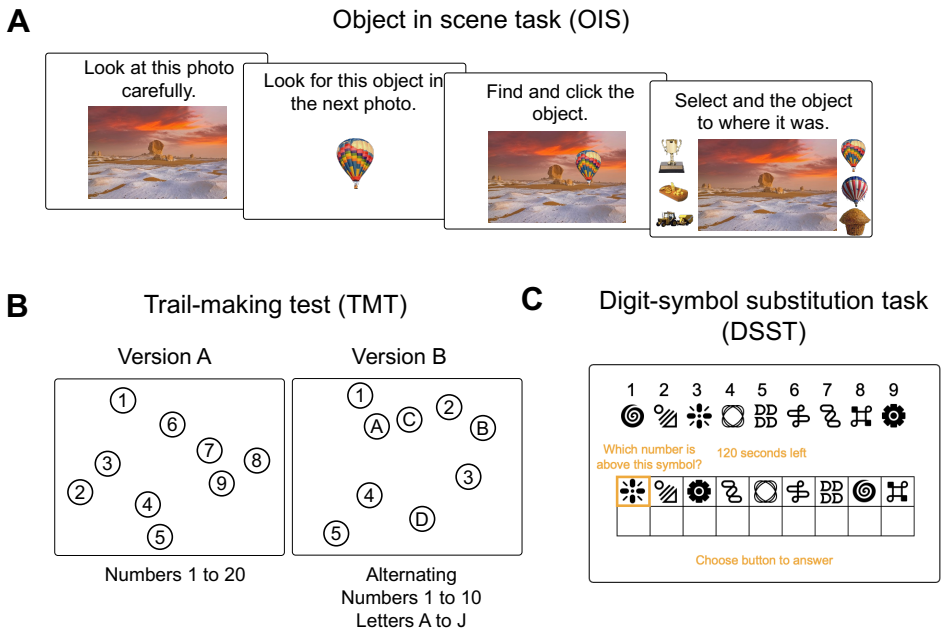
This procedure would yield two blocks per experiment of intermixed trial types, with total trial numbers per block ranging from 158 to 217 (see Fig. 3.1D, left panel for an example). An average count of the trial types that were encountered by participants

is shown in Figure 3.1D (right panel). These were consistent across participants for Encoding, Retrieval, and Lure trials, with some variation for Filler trial numbers, which meant participants completed similar amounts of different trial types. Notably, there were more Encoding than Retrieval trials because Encodings continued throughout the block, but not all items were retrieved. Once the desired threshold of 200 items per participant was reached, including Lures, the block ended.

21 participants additionally performed a control task after the main experiment. The control version of the task was created to assess memory performance without interference for retention intervals between one and five events. In the main task, memory delays were determined by all events occurring in between encoding and retrieval of an item. This included the ITI (1.5 s) plus the presentation time of interfering items (1 s) in case of Encoding trials or participants' response time in case of Target, Lure, or Filler trials. For example, a retention interval of two items with one interfering Target retrieval would be ITI (1.5 s) + RT (trial-dependent). To match memory delay times in the control task to those in the main task, we averaged response times in the main task across participants. For retention intervals of one to five events, the respective delay times were 1.5 s, 4.8 s, 8.1 s, 11.4 s, and 14.7 s. For this task, we used items that did not appear in the main task, and colours were randomly assigned.

Experimental design: Validation tasks We included three validation tasks of different cognitive domains to assess which domains the Any-back task might be related to. The same 21 participants who performed the control task also completed

Figure 3.2
Validation tasks



(A) Object-in-scene Task (OIS). A memory task where participants were prompted to find a specific object in a scene and remember its location. They then had to select the correct item amongst distractors and drag it to its original location. **(B)** Trail-making Test (TMT). A task measuring processing speed and executive function. Participants had to connect numbers in ascending order (version A) or switch between letters and digits (i.e. 1-A, 2-B..., version B). **(C)** Digit-Symbol-Substitution Test (DSST). A task measuring processing speed and general intelligence. Participants translated a number of symbols into numbers using a key.

these validation tasks after the main task and the control task. We used selective tasks implemented in the Oxford Cognitive Testing Portal (Toniolo et al., 2024).

The Object-in-scene task (OIS) tests object and spatial short-term memory (Fig. 3.2A, Toniolo et al., 2024). On every trial, participants were first presented with a complex scene (e.g. a landscape). They were then shown an object (e.g. a hot air balloon), and on the next screen prompted to find and click on the object in the scene. On the next screen, they were required to select the target from 20 objects and drag it to the same location they remember it seeing it in the previous scene. Distractors were com-

prised of 19 different objects and one object of the same kind as the target. Selection accuracy was assessed as the proportion of correctly chosen targets, and location memory error was measured as the Euclidean distance between target location and response location. There were 20 different objects and 4 scenes. There was one practice trial and the instructions were displayed on every screen throughout the task.

The Trail-making test (TMT) is an established test to assess processing speed, mental flexibility, and executive function that involves connecting numbers and letters in ascending order (Lezak, 2004; Toniolo et al., 2024). First, participants did a control task in which they connected the 1 and 2 digits to assess their motor speed. They then connected digits 1 to 5 as a practice trial. In the main task, version A, they connected numbers one through 20. In version B, they connected digits one through 10 and letters A through J, alternating between digits and letters (1 – A, 2 – B, etc.). Each participant completed both versions 4 times. Performance on version A was a measure of processing speed and performance on version B assessed processing speed and task switching or mental flexibility. To get a pure measure of flexibility, we subtracted performance on version A from performance on version B.

Finally, the Digit-Symbol-Substitution task (DSST), is a measure of general cognitive performance (Toniolo et al., 2024; Wechsler, 1939). Participants were presented with a legend of 9 digits and 9 associated symbols. They then had to translate symbols into their associated digits with the legend shown on the screen. The goal was to translate as many symbols as possible within 120 s and their score was the number of symbols translated.

Stimuli Stimuli in the Any-back task were 200 images depicting real-world item silhouettes (see Table S15) presented in a colour randomly drawn from equiluminant CIELAB colour space ($L = 70, b = 20, a = 20$). On each event, one item was presented in the centre of the screen at roughly 320 $\times$ 320px.

Procedure The experiment was hosted on pavlovia.org and distributed via Prolific. After signing up on Prolific, participants completed the experiment at home on their personal laptop devices. They were told to sit one arm's length, or 50cm, away from the screen. Participants started with the Any-back task. They were instructed that they would see a number of objects sequentially, and that they should remember them and their shade of colour. It was explained to the participants that, after a while, they would be shown an object and prompted to indicate whether they had seen it before, followed by recreating the colour in which they had seen the object. The image was shown in white until they selected a colour. They were told to be as accurate as possible and completed five practice trials in the following sequence: Encoding of item 1, Filler, Retrieval of item 1, Encoding of item 2, and Lure. The order of practice trials was the same for all participants.

A subset of 21 participants completed the control task and the validation tasks after completing the main experiment. At the beginning of the main experiment, these participants were instructed that they would do four tasks and complete them in the following order: Any-back task (including the control task), OIS, TMT, and DSST.

After finishing all assigned tasks, participants were returned to Prolific to collect their credit.

Analysis Trials where reaction times (RT) were higher than $2.5SD$ of the participant's RT were excluded from analysis. Moreover, participants were excluded from analysis if they scored less than 90% on average on Filler trials, less than 66% on object memory, or more than 75° on colour estimation.

For all analyses, the first 30 trials were discarded to achieve a roughly even split between trial types. This was necessary to avoid analysing retention intervals with only Filler trials interfering, a pattern that was very common in the beginning of a block due to the lack of items available for retrieval.

Absolute colour memory error was calculated as

$$error = |\theta_{response} - \theta_{target}|, \tag{4}$$

where θ is the colour value in degrees. As an additional independent variable to assess the effect of interference, we counted the number of interfering items grouped by trial type (Encoding, Target, Filler, and Lure). To assess the number of items in memory, we calculated the difference between the number of encoded items and the number of retrieved items at every trial of a block. This measure was introduced for two reasons. Firstly, we wanted to see whether it provides a better measure of items in memory than number of items encoded. Secondly, we wanted to illustrate that number of items definitely exceeded typical WM capacity limits. This is not entirely obvious from the duration of the retention interval. We treated items that participants retrieved incorrectly as "retrieved" nonetheless, as incorrect retrieval would mean that the item was forgotten and not part of the set of active memories.

One caveat is that participants were not explicitly told that they could forget an item once it had been retrieved, and may have held the item in memory even after. An explicit instruction to "forget" items was not included in order to keep the task simple, although future versions might examine the effect of such an instruction.

Object memory accuracy and colour memory error data were analysed by fitting mixed-effects logistic and linear regression models, respectively. We used the lme4 package with default optimisers to fit models (version 1.1-28, Bates, 2010), and lmerTest (v3.1-3, Kuznetsova et al., 2017) to calculate p-values. Random effects included only random intercepts due to a lack of trials per participant resulting from fine-grained variables, e.g. a large range of retention intervals (1–30). For all analyses, we performed model comparisons adding the predictor in question and adding other variables of interest successively. Minimal random effects were added, leading to a general formula $y \sim x + (1 \mid id)$. A more complex random effects structure was not possible due to limited trial numbers. Statistics or model comparisons and parameter estimates of the respective winning models are presented in the Appendix (7.3).

Where necessary we followed up with ANOVAs and post-hoc pairwise t-tests using the stats package (version 3.6.2) included in base R. Greenhouse-Geisser and Huynh-Feldt corrections for departure from sphericity were applied where necessary. Multiple comparisons in post-hoc tests were adjusted using the Holm method.

Circular clustering was used to assess categoricalness in colour and orientation estimation responses. This was done using the CirClust function from the OptCirClust package in R, using the fast and optimal "FOCC" algorithm (version 0.0.4, Debnath

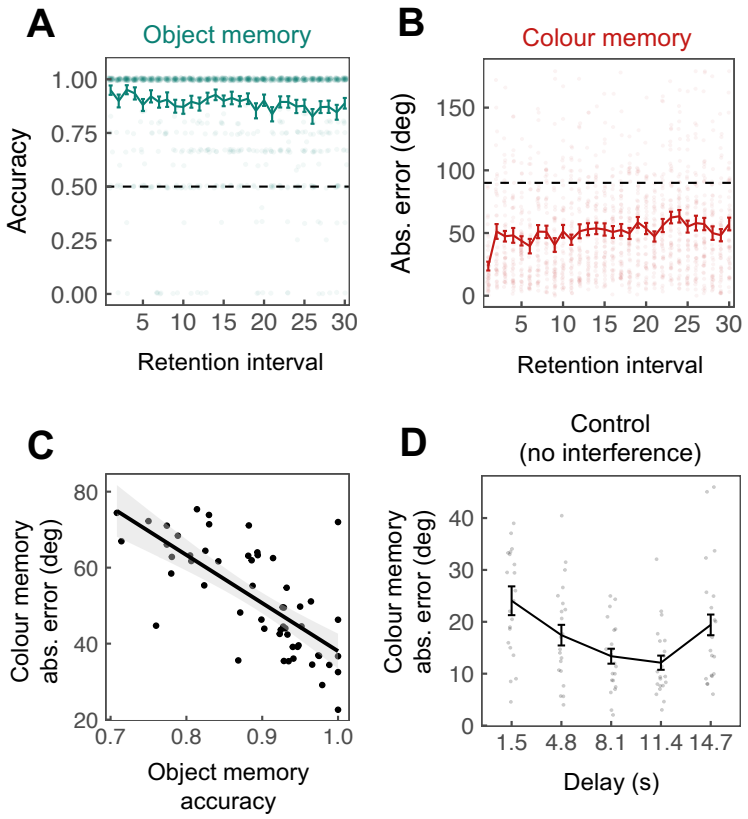
and Song, 2021). Parameters of the circular normal (von Mises) distribution were estimated using functions from the circular package (version 0.4-93, Lund et al., 2024). In the clustering analysis we explored up to eight clusters per response variable. We report the number of clusters that was optimal according to a scree-plot and the option that presented a better visual fit with cluster means, aligning with cluster modes in the response distribution. In addition, we assessed whether participants were using a categorical colour representation, and whether they did so for specific response intervals. A "cluster-based" response was created by choosing the cluster that would be closest to each target value and calculating response error as the difference between them. Paired t-tests were run to see whether empirical and cluster-based error distributions differed at every retention interval.

Finally, we assessed correlations between the Any-back task and validation tasks (TMT, OIS, DSST) using Pearson correlations.

3.3.2 Results

Items and their colour were maintained in memory above chance-level across retention intervals Object memory performance was overall very high for all retention intervals empirically: 95% at retention interval 1 to 90% at 30. The winning model was the one which included retention intervals and corrected for retrieval time *t2* (Table S1). This suggests that accuracy decreases slightly with increasing retention interval, by a 0.09% decrease per increment of retention interval ($\beta = -0.02, p < 0.001$, Fig. 3.3A). It slightly underestimates the empirical decrease in memory performance

Figure 3.3
Exp. 1 Effects of retention interval



(A) Object memory decreased slowly with increasing retention intervals. **(B)** Colour memory error increased sharply by 27.59° after only one interfering item (retention interval 2), and afterwards increased at a much slower rate with increasing retention intervals. **(C)** Object memory accuracy and colour memory error were strongly correlated. **(D)** Results from the control experiment without interference. Delays were matched to average delays of the main experiment. Colour error decreased slightly with delay and then increased again. Points represent subject averages; error bars within-subject errors. Dashed lines represent chance performance.

($\sim$ 2.61% instead of the empirical $\sim$ 5%) from retention interval 1 to 30. The model including the interaction term of retention interval and retrieval time ($\beta_{\text{retention interval}} = -0.05, p < 0.01, \beta_{IE} = 0.0003, p = 0.051$) indeed predicted 90%, matching the empirical data better. However, this model did only provide a marginally significant statistical improvement ($p = 0.088$) over the model including main effects only (Table S1).

Colour memory performance was at 23.63° with no interference (retention interval = 1) and decreased to 57° (retention interval = 30). The winning model included only retention interval as a predictor and showed an effect ($\beta = 0.48, p < 0.001$, Table S2, Fig. 3.3B). Error increased by 0.77° per increment of retention interval. Crucially, there was an immediate decrement to colour performance between no interference (interval of 1, $M = 23.64°$), and one interfering item (interval of 2, $M = 51.23°$, $p = 0.003$) with participants performing 27.59° worse. However, rather than plateauing, colour error did increase at a much slower rate after this decrement as shown by a linear mixed model including only retention intervals 2 to 30 ($\beta = 0.33, p < 0.001$). This means that from one interfering item to 29, participants performed 9.57° worse.

For object memory accuracy, there was no significant immediate decrement ($F(1, 173) = 1.21, p = 0.27$).

Object memory and colour estimation performance were highly negatively correlated ($\rho = -0.72, p < 0.01$, Fig. 3.3B): the higher the object memory accuracy, the lower the colour memory error.

In order to assess how performance would compare with no interference, we ran a control task where we matched a delay without interference to an average delay from the main experiment. At the same delay as a retention interval of 1 (no interference), performance was comparable at 24.2° ($t(14) = -0.96, p = 0.35$, Fig. 3.3D). At a delay equivalent to two interfering items (4.8 s) performance was lower than in the main experiment at 17.3° of colour error ($t(13) = 2.77, p < 0.05$). There was a difference between delays ($F(1, 4) = 5.52, p < 0.001$). Post-hoc pairwise comparisons showed that

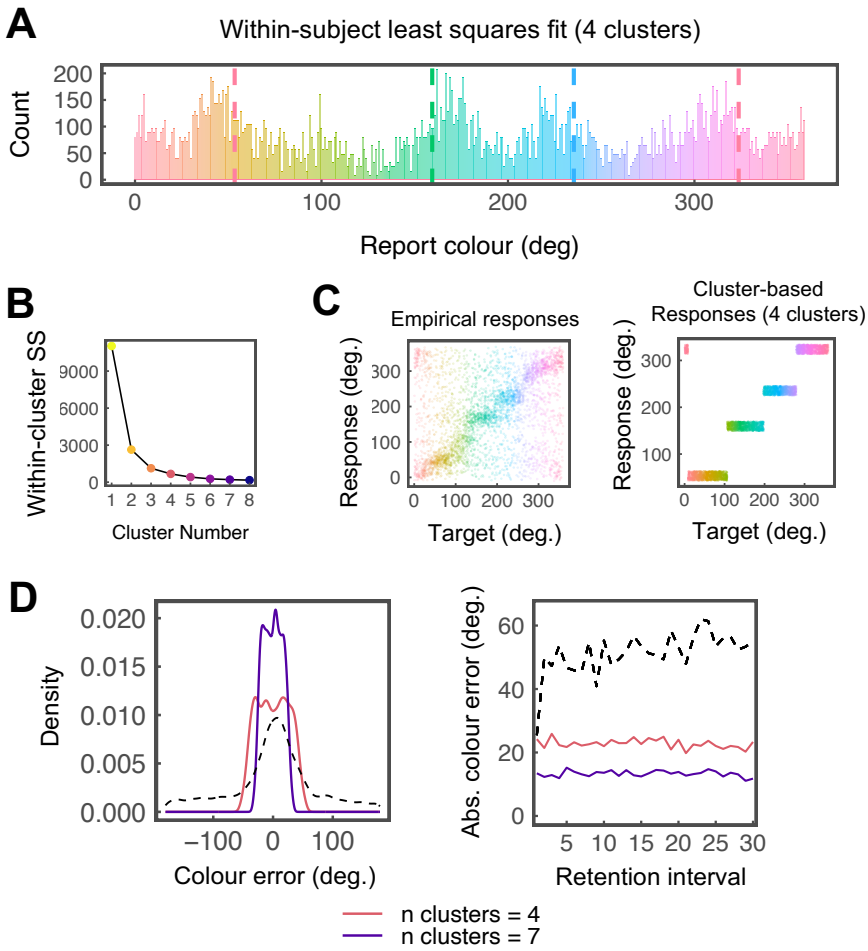
it first decreased when comparing retention intervals 1 with 3 ($t(20) = 10.52, p < 0.01$) and 1 vs 4 ($t(20) = 11.49, p < 0.01$), but no difference between 1 and 5 ($t(20) = 4.29, p = 0.7$).

Clustering reveals categorical responses Remarkably, both object and colour performance were sustained above chance (dashed lines in Fig. 3.3B & C) across retention intervals. We conducted circular clustering to assess whether participants used categorical colour representations to achieve this. According to the scree-plot, a model fit with 4 clusters can be regarded as sufficient to represent the data (Fig. 3.3A).

We then used the detected clusters as a response model to see how responses generated purely based on clusters (illustrated in Fig. 3.4C, right panel) would compare to participants' responses (3.4C, left panel) and whether they would match the characteristic error pattern.

Cluster-based error distributions matched the general shape of participants' error distributions (Fig. 3.4D, left panel). However, the only way of generating responses in this model was cluster-based, omitting the possibility of errors such as guessing responses or swap. There was an upper bound to the error determined by the average SD of all clusters: since every target colour belonged to a cluster, responses could only be a certain distance away. As a result, error density distributions generated from clusters were short-tailed, i.e. clustered around zero, and dropping to zero outside the average SD of clusters. Due to these shortcomings, the absolute number of clusters should not be overinterpreted. Rather, the rationale behind this analysis was to show

Figure 3.4
Exp. 1 Categoricalness of colour responses

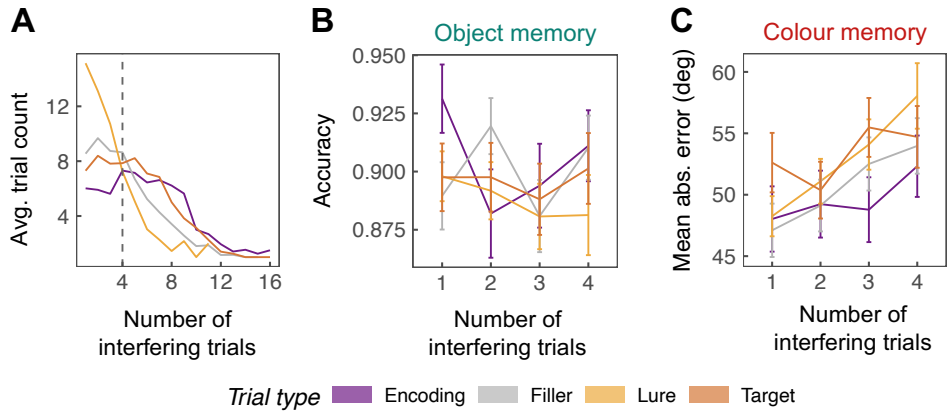


(A) Participants' colour responses over the course of the entire experiment. There were clusters in colour response data corresponding to colour categories (dashed lines). They roughly aligned with distribution modes. **(B)** Result from cluster comparison that minimises the within-cluster sum of squares between data points (WSS). WSS decreased sharply and plateaued after 3–4 clusters. **(C)** Left panel: Illustration of empirical colour responses from participants compared to target values. Clusters are visible as vertical clusters. Right panel: Illustration of responses based on the clustering. Cluster means were used as responses. Jitter was applied for visibility. **(D)** Left panel: Response error distributions for responses based on models with 4 (red) or 7 (purple) clusters. The dashed line represents the empirical error-distribution. Right panel: Absolute error as a function of retention interval. The more clusters the response was based on, the more it underestimated participants' response error. It does not match the pattern of a sharp increase in error followed by a much slower increase in the empirical responses (dashed line).

that, to the extent that participants relied on clusters, the same number of clusters could not underlie their responses at different retention intervals.

When grouping responses as a function of retention interval (Fig. 3.4D, right panel) it was apparent that cluster-based responses did not explain participants' responses entirely. Cluster-based responses produced a uniform response error and therefore cannot account for the sharp rise in error after interference. However, based on the number of clusters, the empirical response (dashed line in Fig. 3.4D, right panel) could either be matched at a retention interval of one item (matched by responses from four clusters, red line) or two items and above (matched responses from seven clusters, purple line). This suggests a change in the number of clusters participants relied on after one interfering item, with participants relying on a smaller set of clusters for retention intervals of 2–30 items.

Figure 3.5
Exp. 1 Effects of Interference



(A) Lines show the average count of interfering trials per item across participants. Each line represents the number of trials where the given trial type occurred in between encoding and retrieval for an item. For example, there were seven trials in which there were four interfering trials of type Encoding. Few items have more than eight of the same trial type interfere with them. **(B)** Effect of interference on object memory. Only Lures interfered with item memory accuracy. Error bars are within-subject errors. **(C)** Effect of interference on colour estimation. All trial types interfere with colour memory, but in particular Lures and Target retrievals. Error bars are within-subject errors.

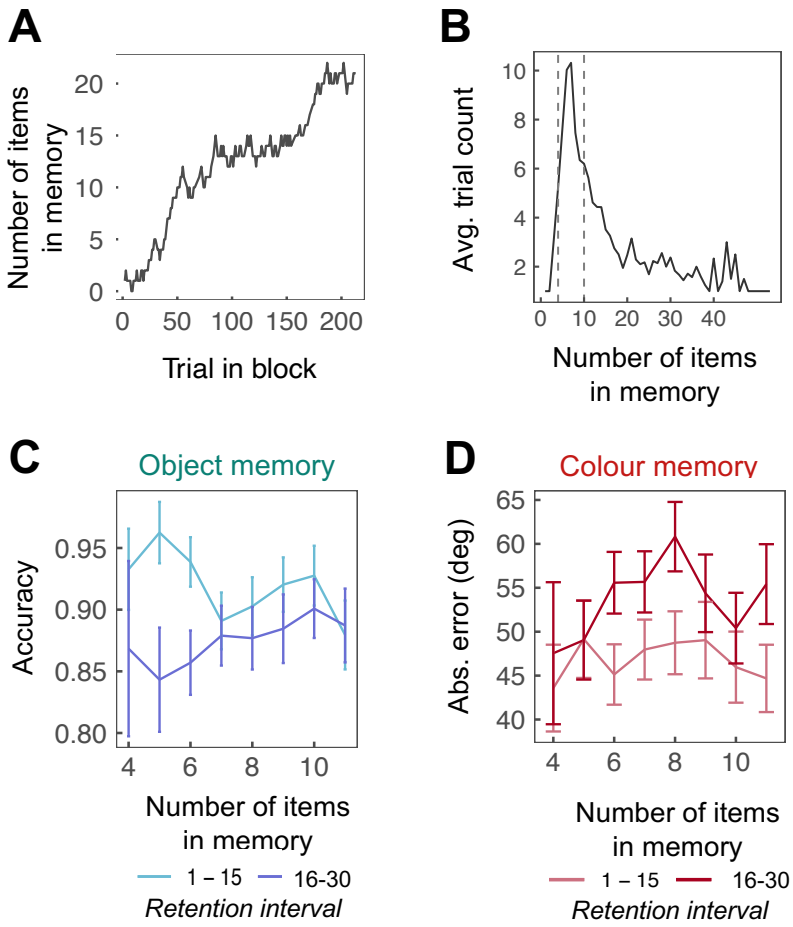
Interfering Lure trials affected accuracy and colour estimation, but Target retrievals affected colour estimation only Next, we looked to investigate what drove the decrement in performance with increasing retention interval. Fig. 3.5A shows the distribution of the average number of trial types. The number of available trials dropped drastically beyond eight trials for Encodings and Target retrievals, and dropped at around four trials for Lures and Fillers. Therefore, only items that had four or less interfering trials from a given type were included in the analysis.

For object memory accuracy, the winning model included trial type as the sole predictor, but none of the levels showed a significant difference (all $p > 0.05$, Table S3). This suggests accuracy was not affected by the number or type of interfering trials.

The model including Number of interfering items and Trial type as predictors performed best for colour memory (Table S4). Colour memory error generally increased with increasing number of interfering trials ($\beta = 2.19, p < 0.001$, Fig. 3.5C). In particular, interfering Targets showed a higher error than Encodings ($\beta = 3.63, p = 0.04$) and there was a trend toward the same for Lures ($\beta = 2.90, p < 0.07$).

Greater number of items in memory affects object memory only for high retention intervals, but affects colour memory error for all retention intervals Next, the effect of the number of items in memory, the amount of items encoded but yet to be retrieved, was examined. This was expected to affect object memory accuracy as there would be more memories to confuse the target with. A visualisation of the number of items in memory variable can be found in Figure 3.6A. Most trials fell within the four and eleven items in memory, so we restricted the analysis

Figure 3.6
Exp. 1 Effects of number of items in memory



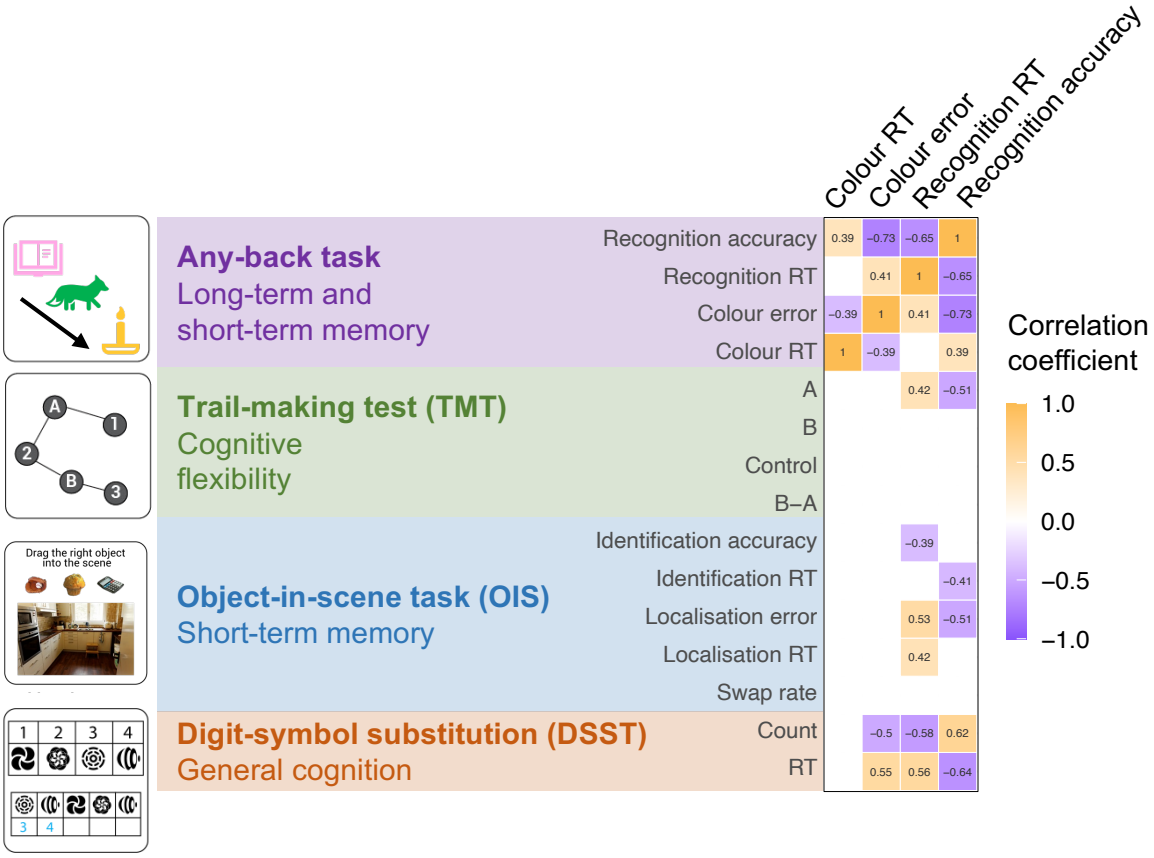
(A) Visualisation of the number of items in memory variable from one example participant for one block. This describes the number of memories encoded but not yet retrieved. **(B)** Trial count for the number of items in memory across participants. Most amounts were available less than six times, so only trials with between four and eleven items in memory were included in the analysis. Note that this is the average across participants for items at item retrievals, and does not match panel A, which displays one specific participant. **(C)** Effect on object memory: Number of items in memory only affected longer retention intervals, for which it worsened performance. **(D)** Effect on Colour estimation: Both short and long retention intervals were affected by greater number of items in memory.

to this range for an adequate trial count (Fig. 3.6B). In order to have enough trials of each factor level of retention intervals and number of items in memory, we binned retention intervals into low (1-15) and high (16-30) groups.

When assessing object memory accuracy, the winning model included number of items in memory and retention interval and their interaction (Table S5, Fig. 3.6C). There was a main effect of retention interval ($\beta = -1.61, p < 0.05$), with accuracy being higher for short than for long retention intervals. There was also a trend towards accuracy decreasing with a greater number of items in memory ($\beta = -0.23, p = 0.06$). However, the significant interaction, the effect of number of items in memory and retention interval ($\beta = 0.15, p < 0.05$) (Fig. 3.6C), suggests that while performance tended to be better for shorter retention intervals overall, holding a greater number of items in memory removed this benefit. As expected, performance started out worse for higher retention intervals. However, it surprisingly improved as the number of items in memory increased.

For colour memory error (Fig. 3.6D), the best model was one including number of items in memory and retention interval, but not their interaction (Table S6). Likewise, there was a main effect of retention interval ($\beta = 7.80, p < 0.001$) on colour memory error. Short intervals showed a lower colour memory error compared to long intervals. While the model including number of items best described the data, the main effect of number of items in memory was not significant ($\beta = 0.18, p = 0.75$). The variance inflation factor for both predictors was ~ 1, suggesting this was not due to multicollinearity. The improved model fit might result from overfitting or a main effect of number of items in memory that was not detected due to the sample size being too small. Alternatively, there might not be a direct effect of the number of items in memory on colour memory.

Figure 3.7
Exp. 1 Correlation with validation tasks



Correlation matrix relating variables from the Any-back task with three external validation tasks: Trail-making test (TMT), Object-in-scene task (OIS), and Digit-symbol substitution (DSST). Performance measures on all three correlated with object memory performance on the Any-back task. Only processing speed (TMT-A) correlated with colour reaction times.

Object memory measures and fewer colour memory measures were correlated with validation tasks This analysis focussed on correlations with the three cognitive validation tasks. The correlation matrix is shown in Figure 3.7.

For TMT, only performance on version A negatively correlated with object memory accuracy ($\rho = -0.51, p < 0.05$) and positive correlated with RT ($\rho = 0.42, p < 0.05$) on the Any-back task. This suggests that participants with a faster processing time on the TMT-A also responded faster on the Any-back, but performed lower at recognition

accuracy, possibly due to a speed-accuracy trade-off. There was no significant correlation of any other TMT measures with Any-back scores (all $p > 0.05$), including TMT B–A, which is taken to be a measure of executive function.

Out of the measures on the OIS, identification accuracy correlated with recognition RT on the Any-back task ($\rho = -0.41, p < 0.05$) and identification RT with recognition accuracy ($\rho = -0.39, p < 0.05$). A similar pattern emerges for localisation performance on the OIS, with localisation error correlated with Any-back recognition RT ($\rho = 0.53, p < 0.05$) and accuracy ($\rho = 0.53, p < 0.05$); and OIS localisation RT correlated with recognition RT ($\rho = 0.42, p < 0.05$). Overall, this suggests that participants with higher performance on the OIS would also perform better at the object recognition on the Any-back task, with performance being either in terms of speed or accuracy.

Finally, recognition accuracy was also related to DSST in a similar fashion, with recognition RT and error on the Any-back correlated with Count (RT:$\rho = -0.58, p < 0.05; \text{error} : \rho = 0.62, p < 0.05$) and RT (RT: $\rho = 0.56, p < 0.05; \text{error} : \rho = -0.64, p < 0.05$) on the DSST. This suggests a speed-accuracy trade-off for measures of both tasks, with RT and accuracy correlated across tasks.

Unlike the Any-back recognition measures, colour memory measures only correlated with DSST: the higher participants error, the higher their reaction time ($\rho = 0.55, p < 0.05$) and the lower their performance ($\rho = -0.50, p < 0.05$) on the DSST.

In summary, we found the recognition memory portion of the Any-back task, but not the colour memory version, to be consistently related to validation.

3.4 Experiment 2

This experiment was designed to investigate whether the observed effects were specific to colour memory or generalisable to other attributes, such as orientation. To this end, we maintained the overall task design, but replaced colour with orientation, using rotated white silhouettes as stimuli. The aim was to determine whether similar patterns of interference, retention dynamics, and storage effects would be observed.

3.4.1 Methods

Stimuli, procedure, and analysis were the same as in Experiment 1. Total trial numbers per block ranged from 164 to 224. When prompted to reproduce the original item orientation, participants were shown the item in a random orientation at the start.

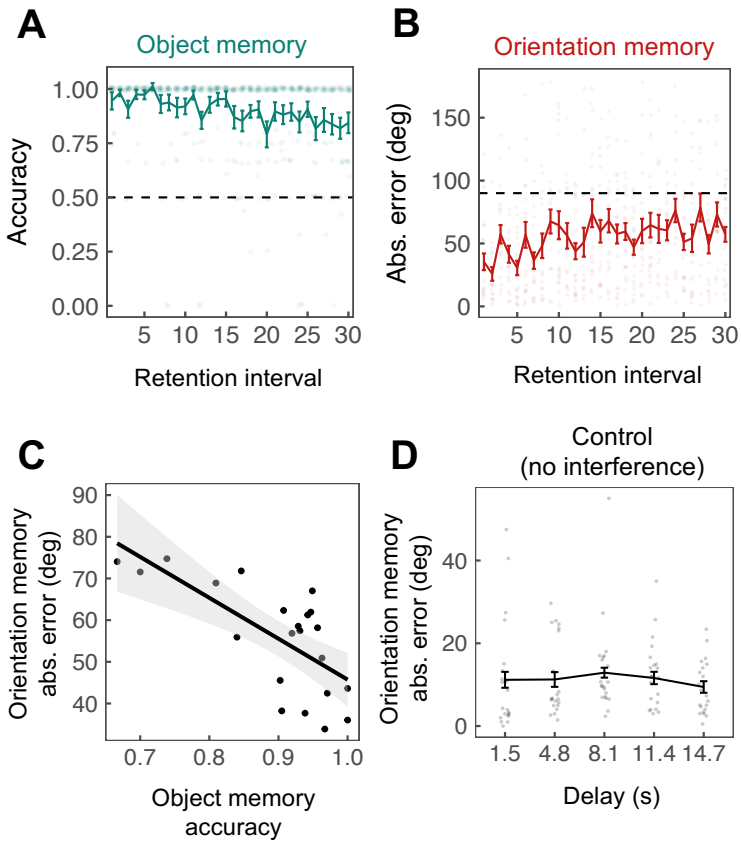
Participants 38 participants were recruited on Prolific and completed the experiment online. All participants performed the main experiment (Any-back task), as well the control experiment and the three validation tasks.

18 participants were excluded due to chance-level responses, or not following the instructions correctly. Cut-offs were 90% for Filler trials, 66% for object memory accuracy trials, and higher error than 75°. Participants who consistently reoriented images right-side-up rather than the to-be-memorised orientation were excluded. This led to a final sample of N = 20 (8 female, $M_{age} = 29.25$ $SD_{age} = 4.41$).

3.4.2 Results

Effect of retention interval on object memory and colour memory error Object memory and orientation estimation performance were again highly negatively

Figure 3.8
Exp. 2 Effects of retention interval



(A) Object memory error decreased slowly with increasing retention intervals. **(B)** Orientation memory error increased sharply by then with increasing retention intervals. **(C)** Correlation between old/new object memory accuracy and colour memory error. Points represent subject averages. **(D)** Results from the control experiment without interference. Delays were matched to average delays of the main experiment. Delay did not affect orientation memory error. Points represent subject averages; error bars within-subject errors.

correlated ($\rho = -0.70, p < 0.01$, Fig. 3.8A): The higher the object memory accuracy, the lower the orientation memory error.

As with colour in Experiment 1, the winning model included only retention interval. Object memory accuracy was high for all retention intervals, starting at 93% accuracy. According to the model, it decreased with increasing interval by 0.12% per

increment ($\beta = -0.06, p < 0.01$, Fig. 3.8B, Table S7). Empirically, it decreased to 84% at a retention interval of 30.

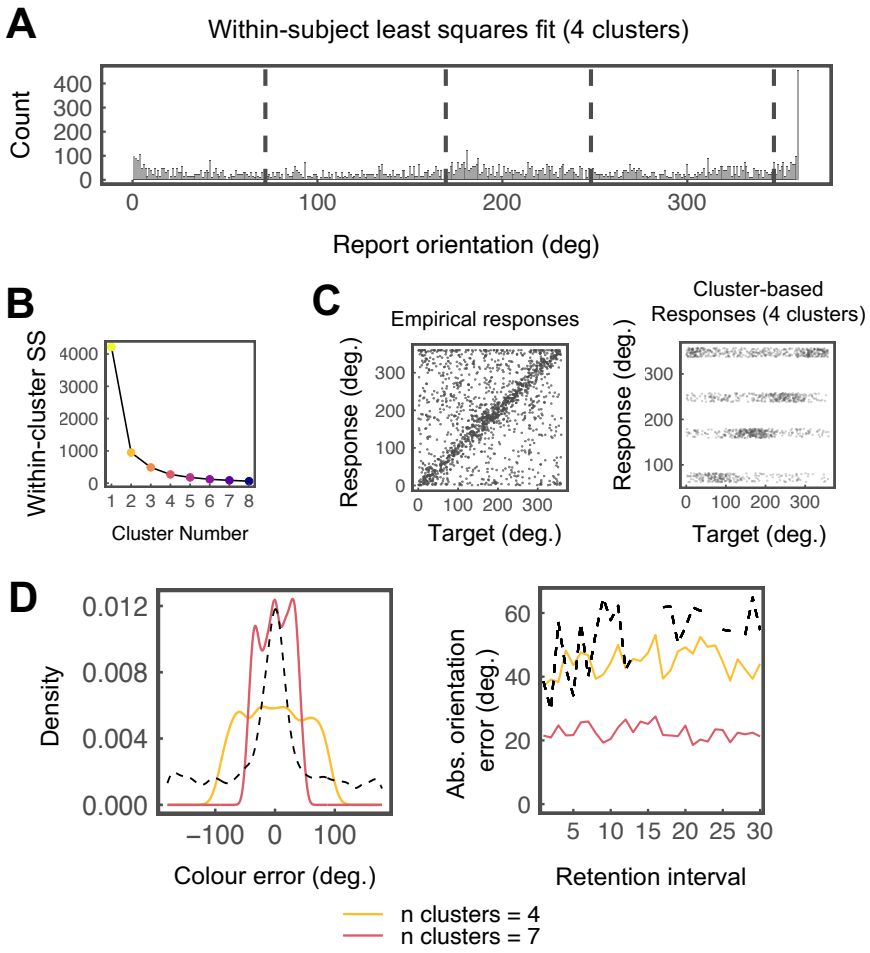
Orientation memory error started at 42.23° with no interference. The best model included the same predictors for orientation memory and also showed an effect of retention interval ($\beta = 0.69, p < 0.01$, Fig. 3.8C, Table S8), suggesting that orientation memory error increased by roughly 0.68° for every increment of retention interval. Empirically, it went down to 54.07° for a retention interval of 30. Unlike for colour, there was no immediate decrement to orientation performance between no interference (interval of 1), and one interference item (interval of 2, $F(1, 70) = 1.93, p < 0.17$). In the control task without interference (Fig. 3.8D), orientation memory error was lower at a delay equivalent to no interference ($M = 11°, t(17) = 3.50, p < 0.01$), and marginally significantly lower at interference from one event ($M = 11°, t(15) = 1.89, p = 0.08$). There was no difference between delays ($F(1, 4) = 0.83, p = 0.51$).

Clustering We again conducted circular clustering to assess whether there were categorical orientation choices in the data. The fit again returned 4 clusters as optimal (Fig. 3.8A, responses displayed in Fig. 3.8B, upper panel). From the figure it seems more likely that participants used only 2 clusters: Most orientations were right-side-up (0°), followed by upside-down (180°).

The cluster-based model was then used to generate responses (Fig. 3.8C, upper panel) and compare them to participants' empirical responses (Fig. 3.8C, lower panel).

As with colour, cluster-based error distributions are shown in Figure 3.4D (left panel). Again, cluster-based responses could not explain the general empirical data

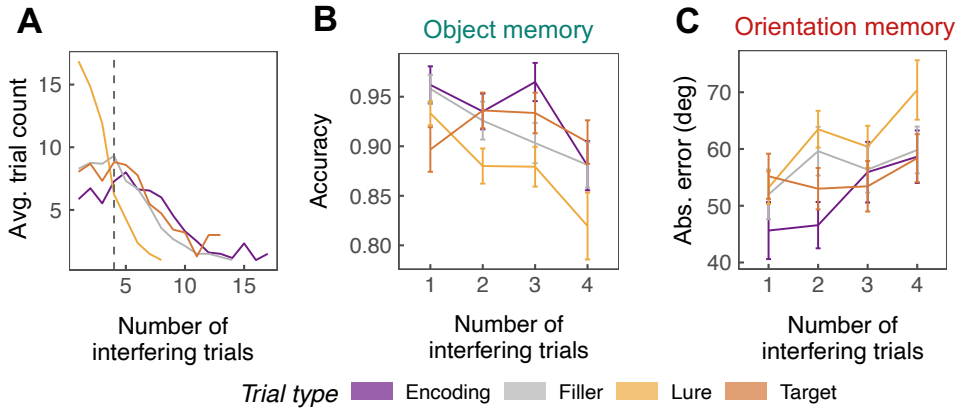
Figure 3.9
Exp. 2 Categoricalness of orientation responses



(A) There were clusters in orientation response data corresponding to direction categories. **(B)** Result from cluster comparison that minimises the within-cluster sum of squares between data points (WSS). As with colour, WSS decreased sharply and plateaued after 3–4 clusters. **(C)** Left panel: Illustration of empirical orientation responses from participants compared to target values. Clusters were visible as vertical clouds. Right panel: Illustration of responses based on the clustering. Responses were cluster means. **(D)** Left panel: Response error distributions for responses based on models with four (red) or seven (purple) clusters. The dashed line represents the empirical error-distribution. Right panel: Absolute error as a function of retention interval. Neither cluster-based response matches the pattern of a sharp increase after one item followed by relatively stable performance in participants' responses.

pattern (Fig. 3.4D, right panel dotted line). This suggests that participants were using a cluster-based strategy either for retention interval 1 or 2-30, but not both.

Figure 3.10
Exp. 2 Effects of Interference



(A) Lines show the average count of interfering trials per item across participants. Each line represents the number of trials where that trial type occurred in between encoding and retrieval for an item. For example, there were seven trials in which there were four interfering trials of type Encoding. Few items have more than eight of the same trial type interfere with them. **(B)** Effect of interference on object memory. Lures, Fillers, and Encodings affected orientation performance. Targets did so to a lesser extent. Error bars were within-subject errors. **(C)** Effect of interference on orientation memory. All trial types interfered with orientation memory, but Lures in particular interfered. Error bars were within-subject errors.

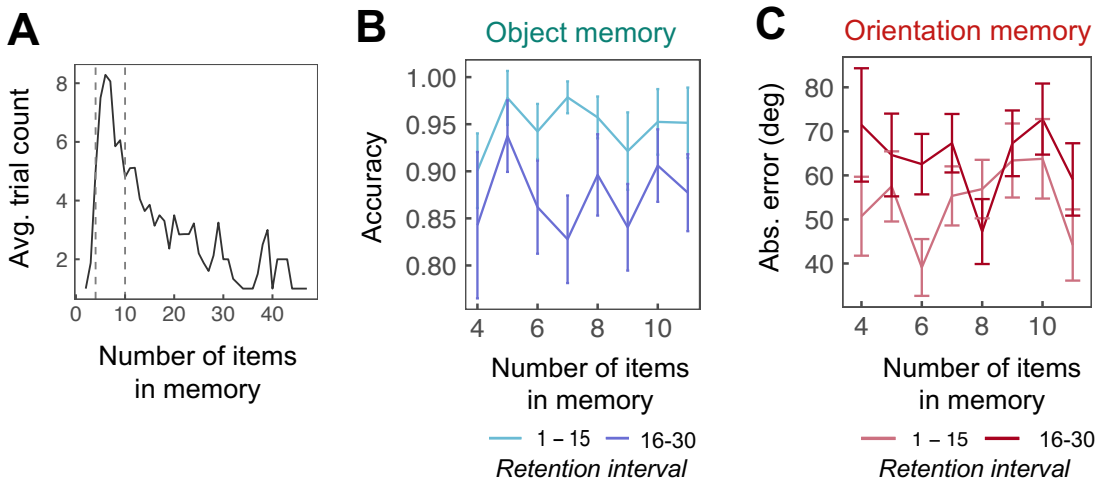
Interference As with colour, we only used trials with numbers of interfering trials of 1–4, because there were below eight trials for every trial type for higher counts (Fig. 3.10A).

The winning model was one including trial type as well as number of trials and the interaction term (Table S9). There was a main effect of number of trials interfering with object memory accuracy regardless of trial type ($\beta = -0.42, p < 0.05$, Fig. 3.10A). There was an interaction effect of number of trials with Targets ($\beta = 0.43, p < 0.05$), suggesting that while Lures, Fillers, and Encodings led to a decrement, Target retrievals did so to a lesser extent.

For orientation memory error, the best model was one with number of items in memory controlling for retention interval (Table S10). There was a main effect

of number of trials on orientation memory error ($\beta = 3.19, p < 0.001$, Fig. 3.10C), suggesting performance was affected by all interfering trials regardless of trial type. Lures showed worse performance compared to Encodings ($\beta = 9.60, p < 0.001$) and there was a trend towards a main effect for Fillers ($\beta = 5.32, p = 0.09$), suggesting the same for Filler trials.

Figure 3.11
Exp. 2 Effects of number of items in memory



(A) Trial count for the number of items in memory. Most amounts were available less than six times, so only trials from four to 11 were included in the analysis. **(B)** Effect on object memory. Number of items in memory affected performance slightly, but only for longer retention intervals. **(C)** Number of items in memory did not have an effect on orientation estimation.

Number of items in memory As with colour, most trials fell within the range of four to 11 items in memory, which we thus restricted the analysis to (Fig. 3.11A). Retention intervals were divided into groups containing ranges 1–15 and 16-30 for this analysis.

The best model was the one including retention interval and number of items in memory, but not their interaction (Table S11). There was a main effect of retention

interval ($\beta = -1.16, p < 0.01$, Fig. 3.11B), with shorter retention intervals showing higher accuracy than longer ones, but the effect of number of items in memory did not become significant ($\beta = -0.01, p = 0.91$).

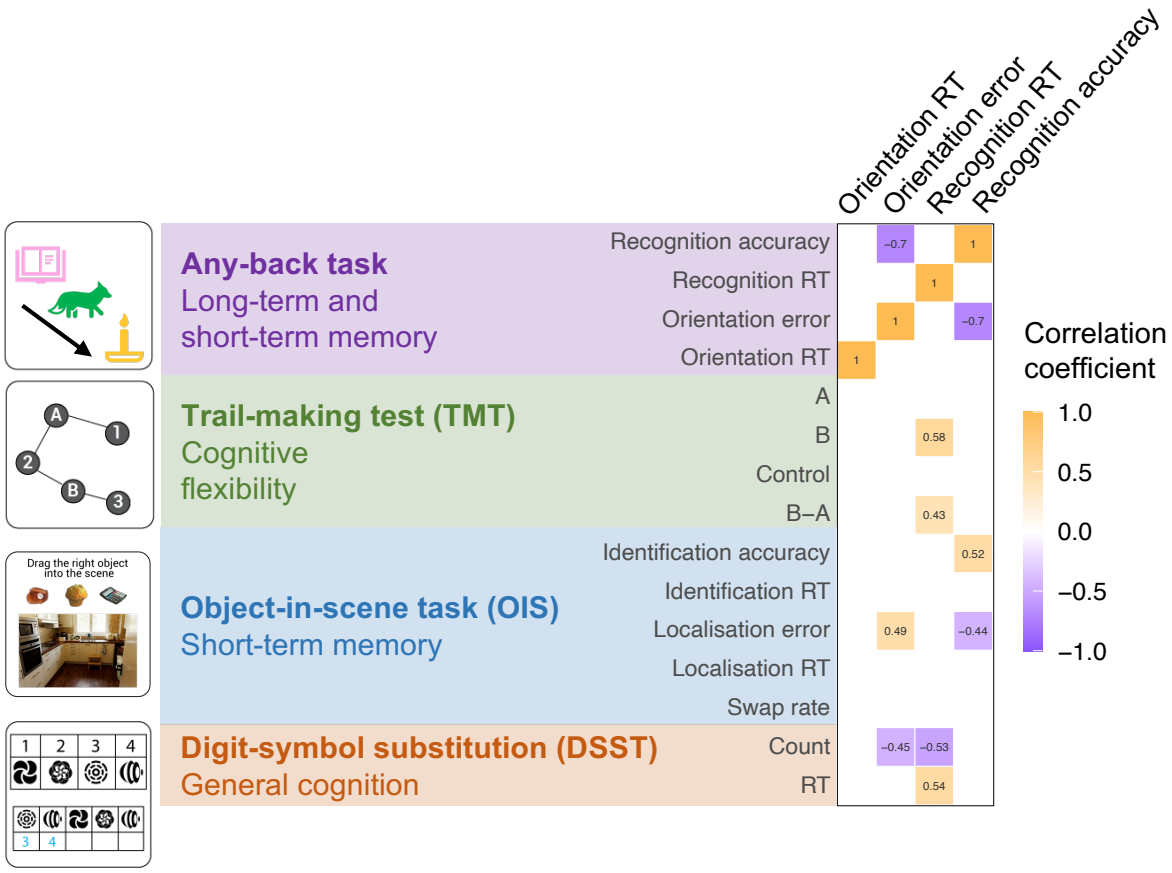
For orientation memory error, the best model was the one including retention interval and number of items in memory (Table S11). However, only the main effect for retention interval became significant ($\beta = 10.03, p < 0.05$, number of items in memory $\beta = 0.56, p = 0.59$), suggesting that number of items in memory did not affect orientation memory.

Validation tasks There were overall fewer significant correlations with the validation tasks than in Experiment 1 (Fig. 3.12).

Recognition RT on the Any-back task correlated with TMT executive functions measures B ($\rho = 0.58, p < 0.05$) and B–A ($\rho = 0.43, p < 0.05$), possibly suggesting that participants with lower cognitive flexibility also struggled responding fast on the Any-back object recognition.

Further, recognition accuracy on the Any-back task correlated with OIS measures of identification accuracy ($\rho = 0.51, p < 0.05$) and localisation error ($\rho = 0.51, p < 0.05$), as well as Any-back colour error correlating with OIS localisation error ($\rho = 0.51, p < 0.05$). This suggests that participants who peformed better at identifying and placing objects on the OIS and points towards both tasks tapping into similar cognitive capacities. The correlation between OIS and orientation memory is in contrast to the colour version (Exp. 1), which found no correlation between OIS and colour memory.

Figure 3.12
Exp. 2 Correlation of the Any-back task orientation version with validation tasks



Correlation matrix relating variables from the Any-back task with three external validation tasks: Trail-making test (TMT), Object-in-scene task (OIS), and Digit-symbol substitution (DSST). Performance measures on the OIS correlated with object memory performance on the Any-back task. Reaction times on OIS and TMT correlated with Any-back reaction times.

This might simply be due to the fact that OIS requires recreation of location which may be more similar to orientation as a spatial feature than colour.

3.5 Discussion

We introduced the Any-back task, a novel experimental task developed to characterise signatures of LTM and WM when they both contribute to memory performance. Participants had to switch between encoding and retrieving items from a continuous

sequence of around 200 unique images in fine-grained colour shades. They maintained items in memory and protected them from a varying number of interfering items of memory-relevant (other memoranda) and memory-irrelevant (Fillers) nature. Results showed that the introduction of just a single interfering item caused a sharp increase in colour error, highlighting its sensitivity to interference. For further interfering trials, however, error increased far less, demonstrating above-chance precision sustained even throughout the longest interval (90 s). In the orientation version of the experiment, feature information was still maintained above-chance performance, however, the sharp rise after one interfering item was absent. Object recognition in both colour and orientation versions, in contrast, showed a gradual decline with increasing retention intervals, retaining high accuracy throughout all delays.

Further analysis of responses found that a higher number of colour clusters corresponded to responses at a retention interval of one item for colour and lower retention intervals for orientation than at longer retention intervals, suggesting a more fine-grained representation. Secondly, the effects of interference varied between object memory and feature memory. Target retrievals, the process of recalling items, affected the precision of subsequently recalled colour memory. However, in the orientation version, it was mainly affected by Lures. In contrast, object memory was not affected by interference in the colour version of the experiment, and affected by Lures, Fillers, and Encodings, but not by Target retrievals in the orientation version. Another distinction was found in terms of the number of objects in memory. It influenced object memory differently depending on retention intervals: at lower intervals, higher memory

loads were detrimental, whereas at higher intervals, they became beneficial. For colour memory, increased memory load particularly affected error rates at longer retention intervals. These effects were not observed in the orientation version.

Finally, we observed correlations between object recognition memory with external validation tasks for the general cognition (DSST), as well as the cognitive domains of processing speed (TMT) and spatial and object short-term memory (OIS). Colour memory correlated with processing speed. Correlations in the orientation version were only found between processing speed (DSST) and short-term memory (OIS) and both object and orientation memory in the Any-back task. The results were generally more robust for colour (Exp. 1), so this will be discussed before addressing the differences in the orientation version.

Different signatures of WM and LTM involvement Several differences between object and colour recognition performance suggest that we were able to capture characteristics of LTM and WM systems.

Error signatures that could indicate differences between WM and LTM include a sharp increase in colour error after one item of interference. Additionally, there are differences in how prone WM and LTM are to interference. For example, Target retrievals seem to affect colour estimation, with Target retrievals and Lures showing greater interference. This is consistent with previous research showing that WM acts as a retrieval platform (Fukuda & Woodman, 2017), and interference occurs when comparing information in WM to perception (Saito et al., 2023). Moreover, the level of interference increases as the number of interfering trials grows. Although it was

expected that object memory would be susceptible to interference from Lures, this was only the case for the orientation version of the task and not for the colour version. However, this could be due to a power issue, as few trials per group were available for this analysis. This could be addressed with a version that balances interference types.

In terms of number of items in memory, it was initially expected that the addition of more items would impair object memory. This was confirmed by the findings, however, only for shorter retention intervals, for which performance worsened. Paradoxically, at higher retention intervals, performance improved. A possible reason for this is that the distraction caused by additional items diminishes over time, allowing for better distinction ability. For colour memory, interference only occurred during longer retention intervals. This might be attributable to colour information becoming more coarsely represented as time passes. However, it is unclear why this should coincide with better object recognition. One caveat to this analysis is that participants were not explicitly instructed that memories would no longer be necessary after retrieving them. This could potentially have muddled this measure and may be remedied by future versions including explicit instructions.

Transfer of information to LTM Our results suggest a possible time point for a "hand-off" of information from WM to LTM. For colour memory, following a sharp initial decrement after one interfering item, performance was remarkably stable for all remaining retention intervals (Fig. 3.2A). This is surprising because information in WM is prone to interference and has limited capacity. Some researchers consider that, in the absence of interference, WM could be sustained actively over longer time spans

(Oberauer & Lewandowsky, 2013; Oberauer et al., 2016), including the longest retention interval (90 s). However, at the longest retention interval, 29 items will have interfered with the initial memory. In addition, the number of items in memory ranged from zero to 30 and would have vastly overwhelmed the established capacity limit of WM of four items at any given time point after 50 trials. This suggests that in order to maintain this level of performance, information would have had to be transferred into LTM. Such a possibility would be in line with findings that showcase the enormous capacity of LTM (Brady et al., 2008; Standing, 1973). However, we found object recognition to be highly negatively correlated with feature error on both tasks on a participant-level. This may point towards individual differences in general memory performance or indicate a strategy of using the object as a cue for feature information. Future analysis might extend this to trial-by-trial analyses.

Our findings suggest that this transfer happens as soon as WM is interfered with after one item due to the shift in accuracy. In addition, our analysis of categorical traces in delayed estimation responses show that more colour clusters might underlie responses before interference compared to after. This suggests that either a recoding of information happened within WM or information was transferred to LTM. The latter explanation is consistent with a recent theory that argues that the memory of an item is supported by episodic memory as soon as the item leaves the focus of attention (Beukers et al., 2021).

Nevertheless, we cannot rule out that LTM could have been involved during the encoding phase. Indeed, colour responses with no interference are consistent with only

four colour clusters – fewer than might be expected with continuous representations (Hardman et al., 2017). This would suggest that long-term, categorical information was involved from the outset, possibly encouraged by the involvement of real-world objects. A more accurate term might therefore be to describe the initial storage as "WM-dependent": the information in WM might be supported by activation in LTM, but not vice versa.

Psychometric potential of the Any-back task We were able to establish some external validity for our paradigm through subject-level correlations with cognitive measures. Most of the external measures were prominently related to object memory rather than colour memory.

Correlations for object memory were related to all three validation tasks. There were significant correlations with general cognition (DSST) and overlapping mechanisms such as reaction times on validation tasks and the Any-back task, as well as general processing speed (TMT A&B). Crucially, in both experiments we observed that participants who were better at object recognition in the Any-back task also performed better at both object identification and localisation in the OIS task. This is important since contextualised or scenic memories are core aspects of episodic memory. Thus, this correlation suggests that the Any-back task was successful in involving episodic LTM with its object recognition task.

For colour memory, the only correlation observed was between colour error (Any-back) and general cognition (DSST). This might seem unsurprising, given no direct analogue of colour estimation was included in the external testing battery. In contrast,

orientation error (Any-back) correlated with localisation error (OIS). This pattern of correlations points, perhaps unsurprising, towards orientation memory to be related more closely to orientation than colour memory, and provides some external validation for the feature memory part of the Any-back task.

Differences between colour and orientation versions There were notable differences between colour and orientation in our study. One significant observation is that the sharp decrement seen in the colour error curve was missing from the orientation version, in which information seemed to degrade more gradually with increasing retention intervals.

Even so, another possible reason for the differences could be that orientation affects recognition in a distinct way. Object rotation, especially at degrees flipping the image upside-down, might have impaired recognition. In previous research on real-world objects, especially faces, turning stimuli upside down eliminated performance benefits compared to meaningless stimuli (Lorenc et al., 2014). If objects had been unrecognisable, however, we would have expected a lower accuracy on object recognition performance. Given that the overall performance was still very high for both object recognition and colour, this seems unlikely.

The issue might instead stem from the limited sample size (N = 20). Due to the wide range of retention intervals, it takes a large sample size to gather enough trials. This is also true for analyses on interference and the number of items in memory, as they equally require division of trials among many conditions. Consequently, our sample

size may be too small to draw strong conclusions about the orientation version of the experiment.

Further, colour and orientation appear to be processed differently. Cluster-based analysis showed that performance with no interference was already less fine-grained for orientation compared to colour. This is unsurprising, given that different perceptual features vary in their prevalence in real-world environments, leading to different expertise levels. Thus, expecting the same effects for colour and orientation might not be a valid assumption. Future versions of this paradigm, however, should consider exploring other features that do not impact item recognisability.

Limitations and future directions A key limitation of this study is that, while there is some evidence for episodic memory being involved, we did not commit to a particular type of LTM in our design. Specifically, we did not differentiate between semantic and episodic memory, which are known to involve different processes and mechanisms. Another limitation is that we did not assess memory using a remember/know judgment, instead relying on a generic old/new question or a simple confidence distinction (Henson et al., 1999). This approach does not capture the more nuanced aspects of memory retrieval and should be adapted in future studies. Additionally, we did not include a recognition test after the memory task, which would have allowed us to more clearly differentiate between WM and LTM. For example, if some items are initially maintained in WM before indeed being "handed off" to LTM, a trace of these items should be more detectable in a subsequent recognition task. The absence of this limits our ability to interpret this effect.

Another limitation lies in our clustering analysis approach. Participants are likely to use a mixture of categorical and continuous colour formats, as suggested by (Hardman et al., 2017). They found that people were particularly likely to report colour based on continuous representations. Therefore, it would be more effective to fit a mixture model that accounts for both cluster-based and continuous representations. Additionally, had there been more trials in a control experiment, it would have been useful to fit the data to this as a baseline, to better understand how these strategies emerge and interact. In our study, we found that participants seemed to be using a maximum of four colour categories (consistent with Bae et al., 2015). However, our methods only show that responses are consistent with four clusters, not that they were actually made on the basis thereof. Further modelling, perhaps applying the model by Hardman et al. (2017), is needed to address this question.

Moreover, clusters are already present in perception, meaning that even the most "continuous" responses in WM will inherently contain clusters. In our data, we identified a basis for 4, which is certainly consistent with (Bae et al., 2015). However, the approach used here relied solely on clustering, which meant we could only model a purely cluster-based strategy.

The precise relationship between WM and LTM remains complex. For example, what is the mechanism behind binding feature information to object information? Besides this, if information is indeed relayed from WM into LTM, it is unclear how the underlying representation is modified from a precise WM representation into a more categorical representation. Further research is needed to address these questions.

The introduction of a novel paradigm enabled us to delineate WM and LTM contributions to colour and object memory, respectively. Our findings are consistent with a model of WM-LTM interactions in which WM initially supports object and fine-grained feature memory, and transfers it into LTM when this information is interfered with, specifically when WM is needed in the interfering event, such as in the retrieval of a different item. This enables reliable storage of object and feature information over timescales that exceed WM. Future research might investigate this transfer further, including its underlying representational formats and brain mechanisms using the Any-back task.

4 Non-Essential Maintenance of State Representations in Working Memory Facilitates Reinforcement Learning

Abstract

Memory and learning have traditionally been studied as separate cognitive processes, but recent research suggests that the neural computations of reinforcement learning (RL) are influenced by working memory (WM). While previous studies have shown a potential trade-off between these two systems, as well as support from WM in resolving key issues in RL, it remains unclear whether WM of the learned stimuli itself facilitates RL. In four experiments, participants learned action-stimulus mappings through probabilistic rewards. Although WM was not necessary for learning, encouraging WM maintenance of the RL stimulus following the RL decision (Exp.1 a&b) enhanced learning. This was not linked to increased adherence to a win-stay-lose-switch (WSLS) strategy. Disruption of the memory of the RL stimulus, however, did not disrupt RL in either probabilistic (Exp. 2a) or deterministic reward settings (Exp. 2b). Moreover, the WM enhancement effect found in Experiment 1 could not be replicated in Experiments 2a&b, possibly due a strategy shift in response to interleaved suppression blocks in which WM was unreliable. A tentative conclusion is that WM may aid RL when it is reliable and conducive to the task. Our results provide some support for a model of cognition where RL uses representations within WM. This framework produces a number of empirically testable predictions about the interaction of learning and WM.

4.1 Introduction

Working memory (WM) functions as a temporary storage system for holding and manipulating information, while reinforcement learning (RL) focuses on learning action-value associations to maximise rewards. Recent interdisciplinary efforts have advanced our understanding of how they relate to one another, which is mutually informative for computational frameworks of RL and neuroscience (Yoo & Collins, 2022). Although WM and RL work together in dynamic environments, they can be differentiated. For instance, there are notable trade-offs in their contributions to learning (Collins, 2018). Additionally, there is some evidence for the potentially collaborative nature of these systems. Individuals with greater WM capacity tend to show improved RL performance (Ben-Artzi et al., 2022), and the representational format of information in WM can influence and guide RL processes (Shibata et al., 2024). Related to these findings, this study investigates a fundamental question: whether maintaining information in visual WM, when it is relevant but not necessary to RL, can facilitate learning outcomes.

RL refers to the process of learning from experience through reward and describes computational and algorithmic principles of this process (Sutton & Barto, 1998). It is rooted in the study of reward learning and classical conditioning in biological systems (Rescorla & Wagner, 1972). Biologically, dopaminergic neurotransmission has been found to underlie this form of learning, with the discrepancy of reward expectation and outcome (reward prediction error, RPE) used to transfer value from the reward outcome to the reward-predictive preceding state and action (Schultz et al., 1997). WM,

in contrast, is a system for temporarily holding items in memory over a short period to enable their manipulation (Cowan, 2017; Ma et al., 2014). It is classically considered capacity-limited and, unlike RL, not traditionally viewed as a learning system.

In the literature, WM and RL have been found to contribute independently or interdependently to learning processes. A prominent theory in the field suggests that WM is engaged at low set sizes, but that RL is driving the learning stimulus-action associations when capacity is exhausted (Collins, 2018; Collins & Frank, 2018). While the RL-based process is slower, it also led to more retention in a subsequent LTM test, while WM provided faster learning at the expense of long-term retention (Collins, 2018; Rac-Lubashevsky et al., 2023). The trade-off has been supported by findings from EEG studies that highlight increased RPE signals in high, versus attenuated RPE signals in low set size conditions (Collins & Frank, 2018; Rac-Lubashevsky et al., 2023). This distinction is further evidenced in patients with schizophrenia who show impairments in WM-dependent learning while retaining intact RL capabilities (Collins, Albrecht, et al., 2017; Collins et al., 2014). Formalised RL models that incorporate a WM component effectively capture these dynamics (Collins, Ciullo, et al., 2017b). However, it is worth noting that experiments in these studies used deterministic rewards, making a stimulus-action association learnable within one exposure. Typically, probabilistic paradigms are used (Niv et al., 2015; Shibata et al., 2024), which prevent solution by WM alone.

Apart from this, other studies report bidirectional interdependence between the systems. RL can influence WM performance through various mechanisms. For ex-

ample, pre-learned item-value associations have been found to lead to rewarded items disrupting WM representations of unrewarded items (Infanti et al., 2015), suggesting a role of reward in enhanced distractor filtering in attentional selection. Moreover, reward not only benefits the encoding of items by improving the representation of relevant items (Wallis et al., 2015), yet it may also enhance WM for equally rewarded stimuli by encouraging slower response times and thereby shifting the speed-accuracy trade-off (Grogan et al., 2022).

On both computational and biological levels, RL faces important challenges like credit assignment, high-dimensional task complexity, and integrating information over time (Gershman & Daw, 2017). Incorporating episodic memory into RL frameworks has been shown to address these issues, enabling effective decision-making in high-dimensional, sparse, or long-term consequence scenarios (Gershman & Daw, 2017).

The credit assignment problem, a key unresolved issue, is the challenge of determining which actions are responsible for observed outcomes and thus should hold higher value. It can be managed by "eligibility traces", temporary records of events, such as visiting a state or taking an action, that allow the system to associate event occurrences with subsequent outcomes (Sutton & Barto, 1998). WM can facilitate post-decision processes by maintaining task-relevant stimuli and thus may facilitate credit assignment. For example, post-decision activation in the visual cortices is specific to the object category preceding a reward, and thus could act as an eligibility trace (Schiffer et al., 2014). Memory processes also play an important role in value inference. For example, humans use episodic memory to infer the value of unchosen options based on

the learned value of chosen ones (Biderman & Shohamy, 2021; Biderman et al., 2023). Similarly, WM supports updating values for irrelevant features after decisions have been made, further contributing to RL processes (Doody et al., 2022) and individuals with greater WM capacity have been shown to be better able to distinguish between outcome-relevant and irrelevant items (Ben-Artzi et al., 2022).

RL often faces intractability when defining values for all possible states or state-action pairs, particularly in high-dimensional spaces or sparse data scenarios. Achieving this computationally is a challenge in biological and artificial intelligence. However, collaboration with other cognitive systems may make this more manageable. It has been shown that attention can reduce the dimensionality of tasks, making them more tractable for RL to process by highlighting which learning components to focus on (Niv et al., 2015). Similarly, information in WM can also bias learning by also influencing the representational formats used in RL, working akin to attentional templates: learning is enhanced when the formats in WM (objects versus features) align with the rule learned (Shibata et al., 2024).

Finally, a critical function of WM is binding information together, allowing for the combination of features to form coherent representations (Vogel et al., 2001; Wheeler & Treisman, 2002). This process is necessary for RL, where actions, states, and rewards must be linked to guide future decision-making. This is especially pivotal when not all components are available at once. Forms of human memory in general have been considered to be a candidate for bridging this gap since it allows for accumulation of information over time (Balkenius et al., 2020). We recently conducted a study that

showed that WM is involved in integrating the components necessary for RL, when these are not available at the same time and need to be held in WM (Shibata et al., Unpublished manuscript). RL has also been reported to be able to operate on candidate states available in WM facilitating the integration of features over short durations (Silva et al., 2018). However, due to the reliance on WM in these circumstances, RL suffers from interference to WM, which results in disrupted integration of information, and slower, learning (Shibata et al., Unpublished manuscript; Wimmer & Poldrack, 2021).

While these findings highlight the supportive role of WM in structuring and maintaining representations for RL, they focus on using WM to solve specific problems in RL. However, the arguably simpler question of whether maintaining simple learning-relevant information in WM can help improve learning is not established. "WM consolidation" is a time-dependent process in which, after encoding, stable WM representations are created (Ricker et al., 2018). If information is interfered with too early in the maintenance interval, this process is disrupted. Equally, if WM is interfered with early during maintenance, this leads to diminished long-term memory (LTM) retention (Bergmann et al., 2013). Moreover, a role of active maintenance in WM in promoting LTM formation has been highlighted (Ranganath et al., 2005).

The current study aims to test whether RL utilises representations within WM by assessing the effect of holding the relevant RL stimulus in WM on learning and reinforcement of the last action. In our design, participants learned to relate four RL stimuli with two possible actions through reward feedback. On each trial, after the participant was presented with a stimulus and chose an action, we enhanced (Exp. 1)

or disrupted (Exp. 2) the maintenance of the RL stimulus. The goal was to assess the effect of this manipulation on the reinforcement and accuracy of the RL decision on the next trial. We hypothesised that if WM contributes to learning, active maintenance of WM should improve learning by increasing participants' adherence to a "win-stay lose-switch" (WSLS) strategy, whereas disrupting WM should impact learning and reduce such a strategy.

4.2 General Methods

4.2.1 Participants

Participants for all experiments were recruited on the Prolific platform. Participants were excluded from the analysis if they did not meet the criterion of 55% overall RL performance, as well as 80% on WM performance. The final sample size included in the analysis was: Experiment 1a: N = 30 (12 female, M_{age} = 28, SD_{age} = 4.54); Experiment 1b: N = 26 (10 female, M_{age} = 28.04, SD_{age} = 3.9), Experiment 2a: N = 34 (15 female, M_{age} = 29.69, SD_{age} = 3.68, 2 demographics reports missing); Experiment 2b: N = 47 (10 female, M_{age} = 29.37, SD_{age} = 4.64, 3 demographics reports missing). All participants were fluent English speakers, reported no history of neurological or psychiatric illness, and were compensated for their participation. All experiments were conducted with the approval of the Research Ethics Committee and informed consent was obtained from all participants through an online questionnaire.

4.2.2 General procedure

All experiments were programmed in PsychoPy (v2022.2.4, Peirce et al., 2019) and hosted on Pavlovia (pavlovia.org). The participants were asked to complete the tasks on their own computer devices, with tablets and phones not allowed. At the start of the study, they were instructed to position themselves at arm's length from their screen and respond using their mouse or trackpad during the task. After the experiment, they were returned to Prolific and given a completion code to collect their compensation.

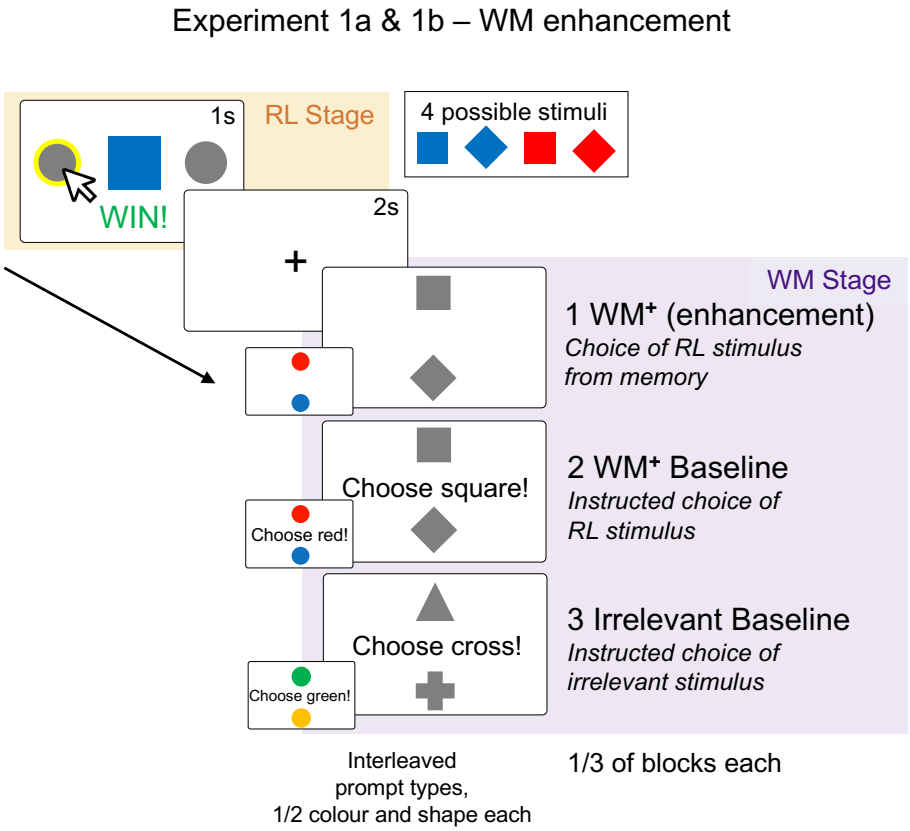
4.2.3 Stimuli

In all experiments, the learnable stimulus-response mapping was based on either colour or shape. There were 4 states (colour: red or blue; shape: square or diamond), 2 actions (left or right) and 2 reward types (win or lose). All stimuli were shown on white background. RL stimuli (states) were presented as squares or diamonds of red or blue colour (Fig. 4.1) and took up 20% of the screen height. Actions were displayed as circles to either side of the RL stimulus at a size of 12% of screen height. Rewards were displayed above the RL stimulus as "WIN!" in green in case of a reward or "LOSE!" in red in case of no reward. Throughout the experiment, a running total of collected points (the "Bank") was displayed at the top of the screen. Stimuli used for the Irrelevant baseline in Experiment 1b and the Suppression condition in Experiment 2b were triangles and crosses in either green or yellow and were displayed at 20% of screen height.

4.3 Experiment 1a

4.3.1 Methods

Figure 4.1
Task schematic



Task schematic for Experiments 1a and b. In the RL stage participants were presented a stimulus and chose the action they believed was associated with it. They were given reward feedback which was 80% probabilistic, omitting reward on 1 in 5 correct responses. Which stimulus-action association was rewarded was based on colour or shape, and changed every 32 trials. After a delay, they were prompted to report the colour or shape of the RL stimulus from memory (1, WM^+ condition), or were told the correct option and have to click it (2, Relevant Baseline condition), or were told which one of two irrelevant stimuli they should click (3, Irrelevant Baseline condition). Experiment 1b included only WM^+ and Relevant Baseline conditions.

Experimental design We employed a two-stage task where an RL task and a WM task were run in conjunction for each trial (Fig. 4.1). The RL stage was presented first on every trial and was the same across conditions. In this stage, participants

learned stimulus-action associations through probabilistic reward. They were presented with the RL stimulus for 1 s which could be a square or diamond and red or blue (all stimuli shown in Fig. 4.1). They then had to choose an action (left or right button press) and were shown their reward outcome ("WIN!" or "LOSE!"). The stimulus-response mapping to receive a reward was defined by colour (colour rules: "red = left, blue = right" or vice versa) or shape (shape rules: "diamond = left, square = right" or vice versa). This reward rule changed every block of 32 trials and RL stimuli were equally probable within every 32 trials (dotted lines in Fig. 4.2A). Reward was 80% probabilistic, meaning that correct responses would not be rewarded in 20% of cases, whereas false responses yielded a reward in 20% of cases. Every "win" RL response was worth five points. In order to incentivise WM performance as well, every correct WM response (see below) was worth one point. Points accumulated throughout the experiment and were visible as a total count ("Bank") at the top of every screen.

The second part of each trial was the WM stage, designed to strengthen participants' WM of the RL stimulus. After a 2 s delay, they were asked to report either the colour or shape of the RL stimulus they had seen at equal probability. In the enhancement condition (WM^+), they had to do so from memory. There were two baseline conditions in which participants were told the correct answer (Relevant baseline) or told which one of two irrelevant stimuli they should choose (Irrelevant baseline). These instructions were always 100% reliable. Stimuli in the WM stage were either shapes or colours and counterbalanced within blocks. Conditions were blocked and there were

four miniblocks per condition, two each for colour and shape rules, resulting in 12 miniblocks. Miniblocks were grouped at random into two main experimental blocks.

Procedure Participants completed 10 practice trials for the WM^+ condition. They were informed that the rule for the practice would be a colour rule of type "blue – left". They had to get 80% correct to move on to the main experiment.

In the main experiment, participants completed a total of 384 trials, divided into two blocks of 192 trials. They could take a break in between blocks. Participants were instructed that their goal was to maximise reward by learning which image belonged with which action and to click on the image they thought would yield a reward and that they would be rewarded five points for correct responses. They were told that the action corresponding to the object might change throughout the experiment and that rewards might depend on colours or shapes. They were also told to remember the image and that they would be asked to recall it after a delay, which would earn them one point. For the WM stage, they were instructed to select the stimulus they had seen before, and to follow the instruction written on the screen in Baseline conditions.

Analysis Data were analysed using R (v4.1.2, R Core Team, 2019). Mixed-effects models were fit using the lme4 package (v1.1-28, Bates, 2010) and p-values were calculated with lmerTest (v3.1-3, Kuznetsova et al., 2017). Contrasts between conditions as well as pairwise post-hoc t-tests were obtained using the emmeans package (v1.8.2, Lenth et al., 2024).

Only trials with a correct WM response and within 2.5 SD from the participant's mean RT were analysed. Additionally, every trial after a removed trial was excluded

as well since we were aiming to assess the effect of the WM manipulation on the next trial.

Logistic mixed-effects models were fit to assess the effect of Condition on RL accuracy in the following trial. There is considerable discussion surrounding the specification of random effects in mixed-models with some arguing for a maximal random-effects structure afforded by the experimental design (Barr et al., 2013) versus only including random intercepts to facilitate model convergence and interpretation (Bates et al., 2015). We therefore report both the model with only random intercepts as well as the model including random slopes

$$\text{Accuracy (correct/incorrect)} \sim \text{Condition} + (1 \mid \text{id}), \tag{5}$$

and

$$\text{Accuracy (correct/incorrect)} \sim \text{Condition} + (1 + \text{Condition} \mid \text{id}). \tag{6}$$

Model coefficients are reported as odds ratios and means are reported as accuracy in percentages. Since the results from this approach were inconclusive, we additionally report a one-way repeated-measures ANOVA with Condition as a factor.

Reinforcement was quantified by assessing the effect of reward and Stimulus Change on staying behaviour (repeat action vs. switch to the other action) on the next trial. One strategy reinforcement could operate through is WSLS. Under this strategy, participants should repeat a rewarded action when the stimulus stays the same on the next trial and switch the action when the stimulus changes, or vice versa after a non-rewarded trial. This behaviour can be quantified as reinforcement or the probability of

repeating a choice after reward. We calculated this as the difference in the probability of staying after a reward and the probability of staying after a non-reward trial

$$p(\text{stay}_t) = p(\text{stay}_t|\text{win}_{t-1}) - p(\text{stay}_t|\text{lose}_{t-1}), \tag{7}$$

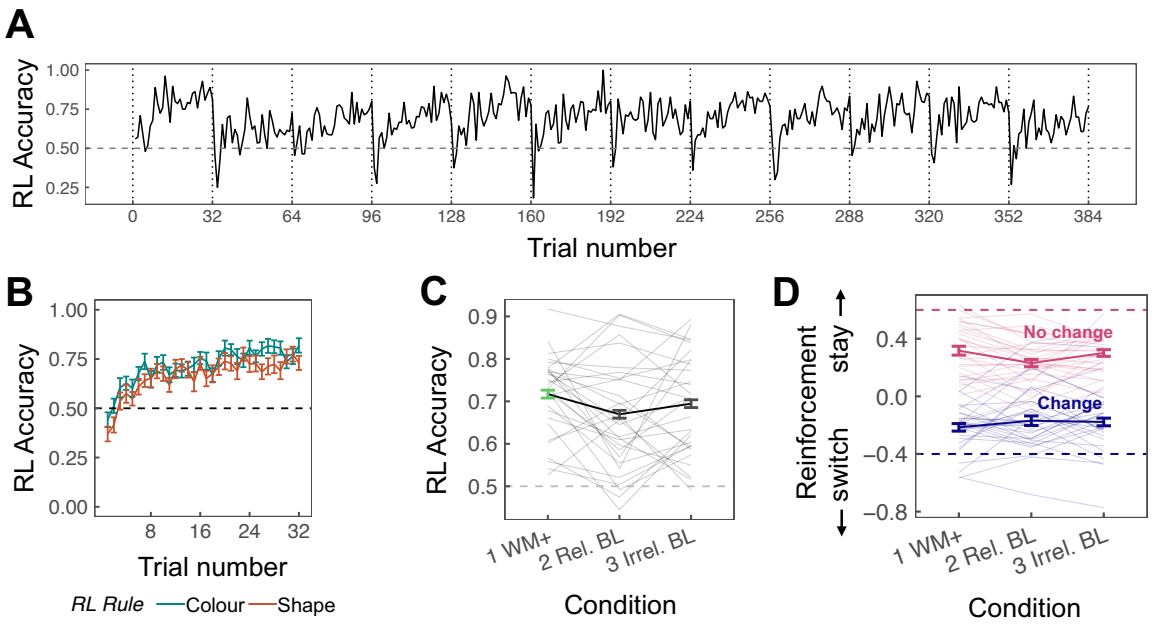
where t is the trial number. This measure indicates the extent to which a participant aligns with the WSLS strategy. If a participant followed the strategy and was sensitive to reward, we would expect a greater value of p(stay) due to more "stay" trials after reward than after no reward. The opposite, i.e. a lower p(stay) would be expected in case of a Stimulus Change. Whether a Stimulus Change was relevant depended on the reward rule in the respective block. Under a colour rule, changes in colour should prompt a switch after reward; whereas a change in shape but not colour should not.

To assess the effect of Condition on p(stay) we entered Condition and Relevant Stimulus Change (Change/No Change) as predictors into an ANOVA with p(stay) as the outcome. Relevant Stimulus Change indicated whether the feature relevant to the reward rule currently being learned was changed from the previous trial or not.

4.3.2 Results

1. Participants learned stimulus-action associations over time Figure 4.2A shows average performance in the RL stage across the whole experiment. At a rule switch (dotted line), participants performance expectedly dropped to below chance behaviour. This was due to the fact that the new rule would on purpose be different to the old rule. Adherence to the old rule in the first trials of a new block would therefore lead to more incorrect than correct responses until participants noticed the change.

Figure 4.2
Results Exp. 1a



(A) Performance in the RL stage across the entire experiment. After a rule switch (dotted lines), participants started below chance performance and improved during each block. **(B)** Average RL performance across all blocks grouped by RL Rule (colour or shape). Participants learned both rule types while it was in action. **(C)** RL accuracy compared between WM^+ (enhancement), relevant and irrelevant baseline (BL) conditions. Performance was better in the WM^+ condition than in the relevant baseline. **(D)** p(stay) by condition. Positive values indicate action repetition after reward, whereas negative values indicate action switching after no reward. Participants based their actions on whether the stimulus changed, but this did not differ between conditions. Coloured lines indicate perfect adherence to a WSLS strategy.

The dotted line represents chance performance **(A–C)**. Each line represents a participant and error bars are within-subject **(C, D)**.

Performance increased gradually thereafter, indicating that participants learned the rule over the course of a block. Averaged for the whole experiment, participants learned both colour and shape rule well, with their asymptotic performance averaging at around 79% for colour and 72% for shape (average across last 12 trials in each block, Fig. 4.2B).

2. Holding RL stimulus features in memory improves RL accuracy

Our hypothesis was that participants would be more likely to choose the correct RL action in blocks where they had to report the RL stimulus from memory compared to

where the response was explicitly provided. In order to assess this, we ran a mixed-effects logistic regression with only random intercepts and one including random slopes as stated in the Methods.

Results are shown in Figure 4.2C and Table S16. Model comparison indicated that the more complex random effects structure was preferable ($\chi^2(5) = 73.62, p < 0.001$). According to the full model, there was a trend towards worse performance in the Relevant Baseline ($\beta = -0.20, p = 0.06$), but not the Irrelevant Baseline ($\beta = -0.09, p < 0.21$) condition compared to the WM^+ condition.

Due to the known caveats of using random slopes logistic mixed models outlined in the methods section, we also report coefficients of the intercept-only model. Here, the model estimated better performance in the WM^+ condition compared to both the Relevant Baseline ($\beta = -0.24, p < 0.001$) and the Irrelevant Baseline ($\beta = -0.24, p < 0.001$).

In order to help distinguish between these two results, we also ran a traditional repeated-measures ANOVA. There was a main effect of Condition ($F(2,84) = 4.2$, $p < 0.05$) and post-hoc pairwise comparisons indicated a difference between WM^+ and the Relevant Baseline ($t(84) = 2.89$) but not Irrelevant Baseline ($t(84) = 1.7$, $p = 0.45$). There was no difference between Baseline conditions ($t(84) = –1.68$, $p = 0.22$). The effect size of this effect as measured by Cohen's d was d = 0.43, suggesting a mild to medium effect.

Overall, these results support a difference between WM^+ and the Relevant Baseline condition, and, to a lesser extent, the Irrelevant Baseline condition. Participants showed greater RL accuracy after holding the RL feature in WM.

3. Behaviour aligns with a WSLS behaviour, but this is not modulated by WM We assessed the effect of holding the RL stimulus in WM on the probability of repeating the same action on the next trial when a relevant Stimulus Change occurred versus when it did not. Positive values indicate that reward increased the number of stays, which we expected to be high for rewarded trials where the stimulus remained the same, if participants followed a WSLS strategy. Negative values, on the other hand, indicate that a lack of reward decreased the number of stays in favour of switches, which we expected to happen after a stimulus change.

We entered Condition (WM^+, Relevant Baseline, Irrelevant Baseline) and Relevant Stimulus Change (Change/No change) into a 3×2 ANOVA with p(stay) as the dependent variable. Results are shown in Figure 4.2D. As expected, there was a main effect of Stimulus Change ($F(1,186) = 30.40$, $p < 0.001$). There was also a trend towards an interaction effect of Condition × Stimulus Change ($F(2, 186) = 2.59$, $p = 0.08$). The main effect of Condition did not become significant ($F(2, 186) = 0.79$, $p = 0.46$).

Post-hoc tests indicated that participants were more likely to repeat their action after reward than no-reward as indicated by the difference in p(stay) in WM^+ ($t(186) = -0.48, p < 0.01$) and Irrelevant Baseline ($t(186) = -0.48, p < 0.01$) conditions. This was not the case in the Relevant Baseline condition ($t(186) = -0.11, p = 0.93$). This

suggests learning was stimulus-specific: reward prompted participants to more stays in case of no Stimulus Change from the previous trial and more switches in case of a stimulus change.

However, p(stay) did not differ between conditions in either Stimulus Change group (all $p > 0.05$). In the Change group, there was a slight trend towards lower p(stay) for the WM^+ condition compared to the Relevant Baseline condition ($t(186) = -2.38, p = 0.16$), as well as a trend towards a lower p(stay) in the irrelevant Baseline condition compared to the relevant Baseline condition ($t(186) = 2.45, p = 0.15$).

These results suggest that WM benefits were not due to an increase in WSLS behaviour.

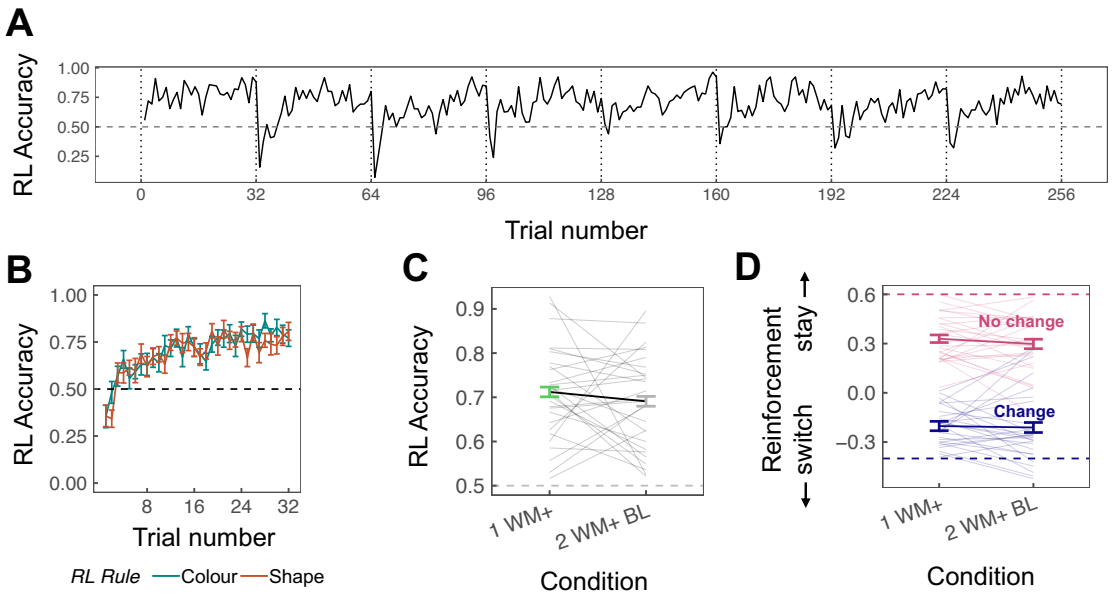
4.4 Experiment 1b

Experiment 1a found that RL performance benefitted from holding the RL state in WM. Since WM was not required in our task, we expected the effect to be small, and therefore wanted to reproduce it in an independent sample. Experiment 1b thus aimed to replicate the effect between WM^+ and Relevant Baseline Conditions in Experiment 1a.

4.4.1 Methods

Participants completed 4 blocks at 32 trials each, for a total of 256 trials. They had to get 70% of practice trials correct to move on to the main experiment. All other methods were the same as Experiment 1a.

Figure 4.3
Results Exp. 1b



(A) RL choice performance across the entire experiment. At the start of each block, participants started below chance performance and improved during each block. Dashed line indicates chance; dotted line indicates a rule change. **(B)** Average RL performance across all blocks grouped by RL Rule (colour or shape). Participants learned both rules well. **(C)** RL accuracy by condition. Performance was higher in the WM$^+$ condition than in the baseline condition. **(D)** p(stay) by condition. Participants based their actions on whether the stimulus changed, and this was the same between conditions. Coloured lines indicate perfect adherence to a WSLS strategy.

The dotted line represents chance performance **(A–C)**. Each line represents a participant and error bars are within-subject **(C, D)**.

4.4.2 Results

1. Participants learned the reward rule Participants learned the rule over the course of each block (Figure 4.3A). On average, their asymptotic performance at the end of each block was 79.8% for colour and 74.9% for shape (Figure 4.3B).

2. Effect of WM on RL performance was replicated As in Experiment 1a, there was a main effect of Condition, with participants performing worse in the Baseline condition than in the WM$^+$ condition ($\beta = -0.11, p < 0.05$, Figure 4.3C,

Table S17). The effect, as measured by Cohen's d was d = 0.20, indicating a small effect.

3. Stimulus-based reinforcement, without effect from WM We entered Condition (WM^+, WM^+ Baseline) and Stimulus Change (Change/No Change) into a 2×2 ANOVA with p(stay) as the dependent variable. Findings are shown in Figure 4.3D. As in Experiment 1, there was a main effect of Stimulus Change ($F(1,100) = 70.52$, $p < 0.01$). Both the main effect of Condition ($F(1,100) = 0.09$, $p = 0.77$) and the interaction effect $F(1,100) = 2.21$, $p < 0.14$) stayed non-significant. This again suggests the extent to which participants used a WSLS strategy was not affected by the WM manipulation.

4.5 Experiment 2a

The first two experiments showed that when participants are incentivised to hold features of the RL state in WM, they show better performance when compared to a baseline using the same stimuli. However, this was not driven by increased use of a WSLS strategy. This is curious, as this is a simple strategy that relies on memory of the last trial only, i.e. remembering stimulus, action, and reward outcome on the last trial guides the decision on the current trial. Incentivising a memory-based strategy should thus have led to stronger consistency with this strategy.

In the following experiment, we wanted to investigate whether disrupting WM would have the opposite effect on RL performance.

4.5.1 Methods

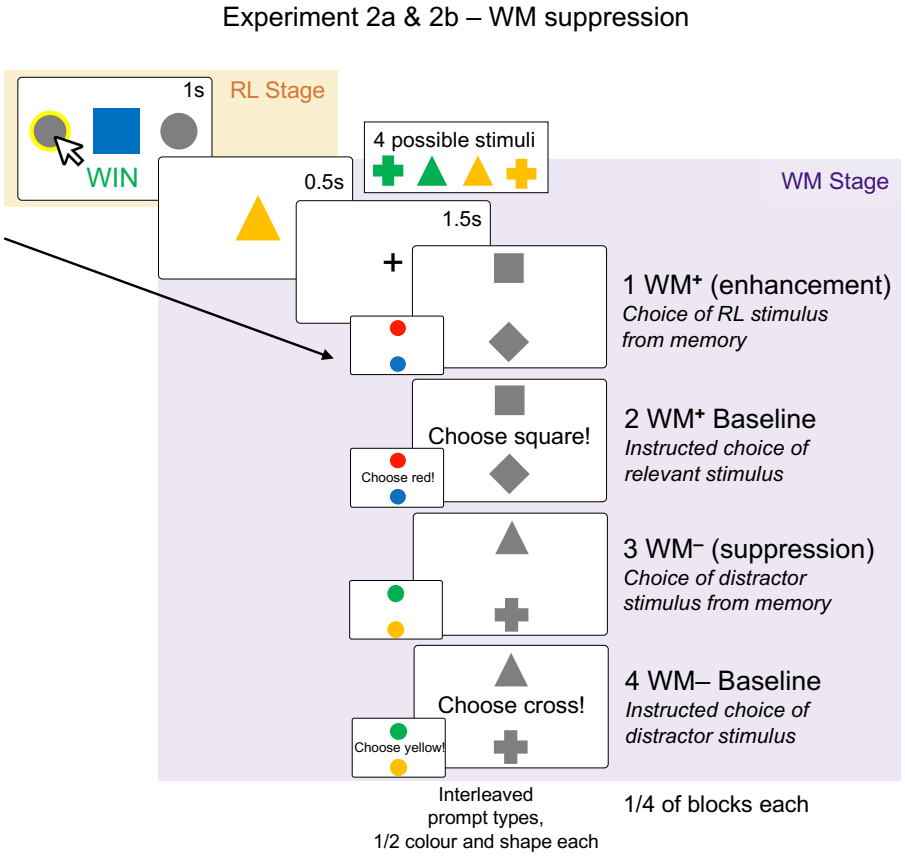
Experimental design This experiment contained the same WM^+ condition and the Relevant Baseline condition (here Baseline WM^+) as Experiment 1. In addition, we included a WM suppression condition WM^- and a baseline (Baseline WM^-). All conditions are shown in Fig. 4.4.

The new condition (WM^-) was designed to disrupt WM. The RL stage was identical to the WM^+ condition. After participants had made their RL choice, memory of the RL stimulus was interfered with by presenting an irrelevant distractor stimulus for 0.5 s to hold in WM during a delay of 1.5 s. They then were prompted to report the distractor stimulus from memory (WM^-) or were told which one was the correct response (WM^- Baseline). The separate baseline condition was introduced to control for perceptual input from the distractor.

Conditions were blocked and participants completed 4 blocks per condition. Each block contained 32 trials, making the experiment 512 trials in total.

Procedure Instructions were the same as for Experiment 1 for the WM^+ condition. Participants completed 20 practice trials in total, 10 from the WM^+ condition, for which they were told the rule. For the WM^- condition, they were instructed that sometimes they would have to remember a different object they would see after the RL stimulus. They practised 10 trials of the WM^- condition with a different reward rule than before which they were explicitly told ahead of starting the practice. They had to get 80% trials correct in order to move on to the main experiment.

Figure 4.4
Task schematic

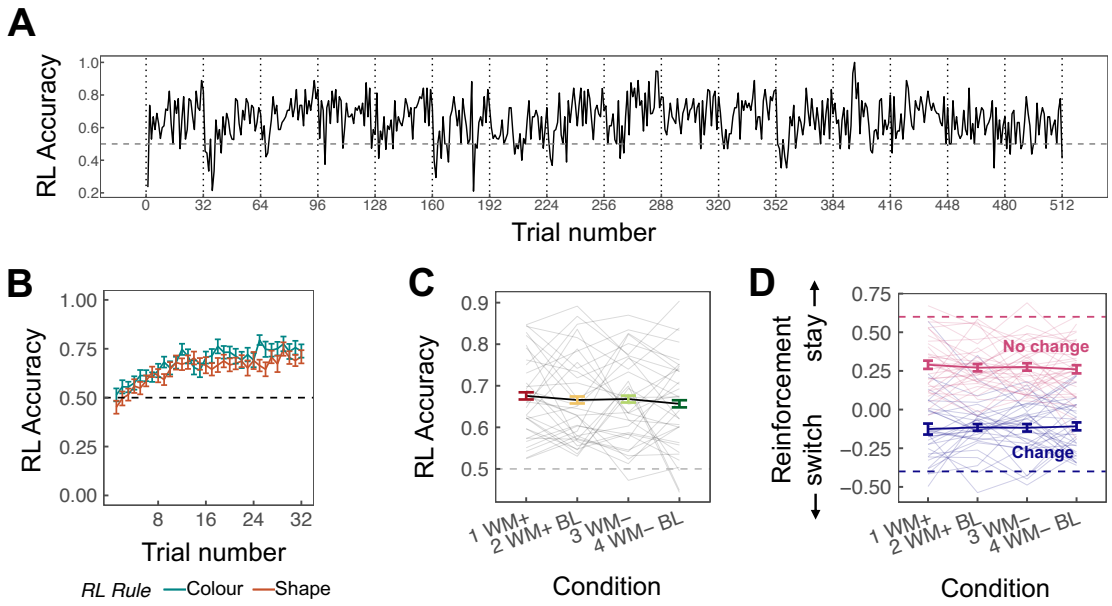


Task schematic for Experiment 2a&b. In addition to the WM^{+} and Relevant baseline conditions in Experiment 1, we introduced a condition designed to suppress WM (4, WM^{-}) and its baseline (3, Baseline WM^{-}). After the RL choice, participants saw a distractor stimulus that they needed to hold in memory instead of the RL stimulus. They were prompted to retrieve the distractor stimulus colour or shape from memory (4, WM^{-}) or told the correct choice (3, WM^{-} baseline).

4.5.2 Results

1. Participants learned the rules As in Experiment 1 participants learned the rule over time (Fig. 4.6A). However, their asymptotic performance was slightly lower than in previous experiments, at an average of 75.6% for colour and 69.4% for shape (Fig. 4.6B).

Figure 4.5
Results Exp. 2a



(A) RL choice performance across the entire experiment. At the start of each block, participants started below chance performance and improved during each block. Dashed line indicates chance; dotted line indicates a rule change. **(B)** Average RL performance across all blocks grouped by RL Rule (colour or shape). Participants learned both rules well. **(C)** RL accuracy by condition. Performance was similar across conditions. **(D)** p(stay) by condition. Participants based their actions on whether the stimulus changed, but this did not differ between conditions. Coloured lines indicate perfect adherence to a WSLS strategy. The dotted line represents chance performance**(A–C)**. Each line represents a participant and error bars are within-subject **(C, D)**.

2. No significant effect of suppressing WM of the state on RL accuracy

We hypothesised that interference with the WM of the RL stimulus would reduce RL performance. However, compared to its matched baseline, performance was similar in blocks where WM was suppressed (contrast WM^- and WM^- Baseline $z = -1.10, p = 0.69$, Fig. 4.6C). Crucially, unlike in Experiment 1, holding the stimulus in WM did not lead to a benefit in RL accuracy as evidenced by comparing the WM^+ condition to its Baseline ($\beta = -0.05, p = 0.33$, Table S18).

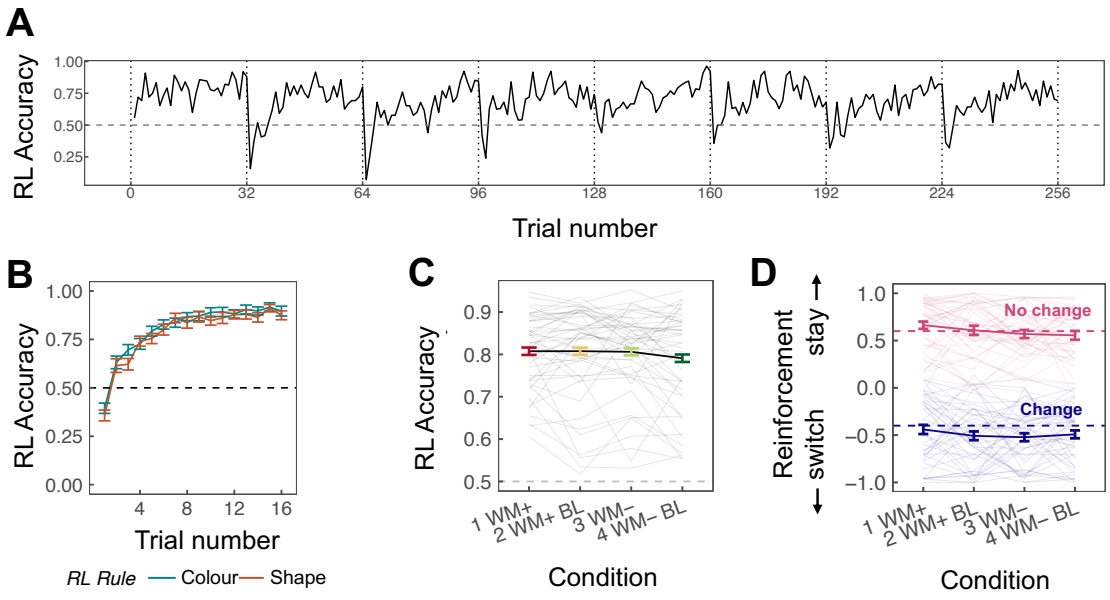
3. Adherence to WSLS strategy is not affected by WM disruption We entered Condition (WM^+, WM^+ Baseline, WM^- and WM^- Baseline) and Stimulus Change (Change/No Change) into a 4×2 ANOVA with p(stay) as the dependent variable. Results are shown in Figure 4.6D. There was a main effect of Stimulus Change ($F(1,256) = 30.40$, $p < 0.001$), suggesting general adherence to the WSLS strategy. Both the main effect of Condition ($F(3,256) = 0.32$, $p = 0.81$) and the interaction effect $F(3,256) = 30.40$, $p = 0.63$) stayed non-significant.

Taken together, these results suggest that neither WM enhancement nor disruption had an effect on RL accuracy. Contrary to our hypothesis, disruption of WM did not diminish RL performance. The lack of an enhancement effect relativises our results of Experiment 1a and b.

4.6 Experiment 2b

Experiment 2a showed no effect of WM suppression on RL performance. In addition, the effect of WM enhancement found in Experiments 1a and b was not present either. Moreover, performance was relatively low at 68%. In order to simplify the experiment, we ran the experiment again, but with deterministic reward: every correct RL choice was rewarded and every incorrect choice was not. We expected that this would make the task easier overall, and, importantly, it would make it perfectly solvable using memory only. We hypothesised that if people used only their memory, this would make effects of memory enhancement and suppression evident.

Figure 4.6
Results Exp. 2b



(A) RL choice performance across the entire experiment. At the start of each block, participants started below chance performance and improved during each block. Dashed line indicates chance; dotted line indicates a rule change. **(B)** Average RL performance across all blocks grouped by RL Rule (colour or shape). Participants learned both rules well. **(C)** RL accuracy by condition. Performance was similar across conditions. **(D)** p(stay) by condition. Participants based their actions on whether the stimulus changed, but this did not differ between conditions. Coloured lines indicate perfect adherence to a WSLS strategy.

The dotted line represents chance performance **(A–C)**. Each line represents a participant and error bars are within-subject **(C, D)**.

4.6.1 Results

Participants learned the rule, averaging at about 90.9% for colour and 88.5% for shape by the end of each block (Fig. 4.6A&B).

Despite the overall better performance, there was no difference between manipulation and baseline conditions. Participants performed similarly in the WM^+ baseline compared to the WM^+ condition ($\beta = -0.01, p = 0.94$, Fig. 4.6C, S19). Equally, there

was no effect when comparing the WM^- condition to its baseline (contrast WM^- and WM^- Baseline $z = -1.43, p = 0.48$, Fig. 4.6C).

We entered Condition (WM^+, WM^+ Baseline, WM^- and WM^- Baseline) and Stimulus Change (Change/No Change) into a 4×2 ANOVA with p(stay) as the dependent variable, Findings are shown in Figure 4.6D. There was a main effect of Stimulus Change ($F(1,355) = 252.90$, $p < 0.001$). Both the main effect of Condition ($F(3,355) = 0.62$, $p = 0.60$) and the interaction effect ($F(3,355) = 0.16$, $p = 0.92$) stayed non-significant. This suggests participants were following a WSLS strategy, but this did not vary between conditions. As Figure 4.6D suggests, adherence to the strategy was near perfect, likely due to the perfectly reliable deterministic reward feedback.

This corroborates the accuracy finding that there was no enhancement or suppression effect through WM.

4.7 Discussion

In four experiments, participants completed a combined probabilistic RL and WM task to investigate the interaction between these two systems. The experiments explored the effect of actively maintaining the RL-relevant stimulus in WM (Exp. 1a&1b) or holding an RL-irrelevant stimulus designed to disrupt memory of the relevant one (Exp. 2a&2b). Experiment 1a showed that RL performance improved simply by holding the RL state in WM compared to baseline conditions, a finding that was replicated in Experiment 1b. However, this improvement was not driven by a WSLS strategy, suggesting that the relationship between WM and RL was more nuanced than follow-

ing simple trial-based rules. Unexpectedly, we observed no significant effect of WM suppression on RL performance neither with probabilistic (Exp. 2a) nor deterministic reward (Exp. 2b). Additionally, the enhancement of WM seen in Experiments 1a and 1b was not replicated. Taken together, our results provide limited evidence for WM contributing to RL performance, with this contribution only observed under specific circumstances.

Experiments 1a and 1b showed a significant effect of WM in RL tasks, suggesting that WM maintenance can facilitate RL. Importantly, the task was not perfectly solvable with WM alone due to the probabilistic nature of the rewards, meaning that memory traces had to be gradually accumulated across trials. This finding contrasts with previous research by Collins et al., which suggested that WM and RL trade off in learning tasks depending on set size, with WM working in low-set-size conditions (Collins, 2018; Collins, Ciullo, et al., 2017a). Here, however, even when the task was low in set size and could theoretically be, albeit imperfectly, solved by using WM, RL processes were facilitated by the active maintenance of relevant stimuli in WM (Exp. 1a&b).

Several potential mechanisms could explain how WM contributes to RL in this context. As mentioned, one driving factor might be active maintenance of stimuli helping to consolidate memory traces (Ricker et al., 2018). This could allow them to last longer and be more effectively transferred into long-term traces to accumulate information over time, which is essential for maintaining value functions for RL.

Moreover, the WM retrieval practice after the RL stage might have acted similar to a retro-cue (Griffin & Nobre, 2003), bringing back not only the RL stimulus, but the associated action and reward outcome. This could aid in using WM to integrate this information (Shibata et al., Unpublished manuscript). Alternatively, it is possible that the WM stage primed and enhanced their attention for the next trial. Although we do not think that attention was enhanced for specific rewarded stimuli, as this likely would have amplified the use of WSLS strategies, we cannot rule out that active WM maintenance had a general attention-enhancing effect. Future studies could incorporate manipulations that can disambiguate between these two options. For example, a paradigm introducing interference to tax WM between trials might prevent the use of WM from the previous trial on the next.

Several aspects of these findings are noteworthy. Firstly, the effect of WM maintenance was found to be significant only when compared to the relevant reward baseline which used the same stimuli as the enhancement condition, but not when compared to the irrelevant baseline which used different stimuli. This seems counterintuitive, as simple repetition of relevant material, as present in the relevant baseline, can improve retention (Hebb, 1961), such that a smaller difference would have been expected between enhancement and relevant baseline conditions. There is a possibility that, because the task was simply to follow the instructions, the relevant baseline did not engage participants enough to keep them focussed on the task, resulting in diminished learning. The seems unlikely because the irrelevant baseline was similar in structure, unless the novelty of the irrelevant stimuli made it more interesting. Finally, it might

simply be that the effect was too small to detect at our sample size. Since the WM was not a strict requirement for this task, this would naturally result in a small effect, which we saw even in the difference between relevant baseline and WM^+ condition. Moreover, the set size in the task was small and the delay brief, resulting in low WM demands, limiting the potential contribution of WM.

Secondly, the absence of a significant effect in the WSLS strategy is puzzling. This strategy, which relies on memory of the last trial, involves remembering the stimulus, action, and reward outcome from the previous trial to guide the decision on the current trial. If WM was expected to facilitate consistency in following this strategy, we would have expected a higher use of this in WM enhancement conditions. One explanation is that this is due to WM only enhancing the stimulus, but not the action and reward outcome, both of which are needed to follow the WSLS strategy based on memory alone. This makes a mechanism of WM integrating the three components (see above) less likely.

Alternatively, it is conceivable that, given the small set size, participants were not only recalling the previous trial but considering earlier trials to infer a pattern. For example, if participants understood that the response to earlier stimuli was correct but unrewarded due to probabilistic reward outcomes, they might disregard the current trial as a potential learning opportunity and refrain from responding in a WSLS manner. This explanation aligns with the fact that participants' behaviour perfectly aligned with a WSLS strategy when rewards were deterministic in Experiment 2b. Overall, this suggests that when the reward structure was simpler, participants were more likely

to rely entirely on memory of the last trial. When rewards were probabilistic, on the other hand, the complexity of integrating information from multiple trials seemed to reduce reliance on just the most recent trial, making the contribution of WM less pronounced.

Contrary to our hypothesis, we did not find an effect of WM when holding an irrelevant stimulus in memory. This could either be due to the disruption of the irrelevant item being too weak, or participants being able to hold the irrelevant stimulus in WM alongside the relevant stimulus. Indeed, this baseline bears some resemblance to WM complex span tasks, in which memory is maintained throughout a short distractor task. Participants are typically required to hold and process multiple items in memory and hold around four items in WM (Cowan, 2001), meaning that the set size in our task was within the capacity limit. Additionally, the small effect observed in the WM^{+} condition might indicate that, while there was some contribution of WM to RL, the manipulation may not have been strong enough to make a noticeable difference in the overall task performance.

Surprisingly, we were not able to replicate the absence of a significant effect in the WM enhancement effect Experiment 1&2, either. This could be due to the alternation between blocks of requiring suppression and enhancement of WM, leading participants to view WM as an unreliable source of information. Since WM was not strictly required, they might have chosen to pursue a strategy of using WM and RL in isolation. This is consistent with Collins (2018) findings that when WM is overwhelmed in higher set sizes, RL takes over.

Given the inconclusive nature of the results, we propose the following, tentative, interpretation. If WM is a reliable source, and thus conducive to the overall goal of RL, it is used to enhance RL. In cases where WM is unreliable or not necessary for task completion, participants might rely on RL mechanisms more, effectively "splitting" the two processes.

To strengthen these findings and further explore the mechanisms behind the enhancement effect, we suggest the following modifications to the experiment. Making the task more difficult in terms of stimuli and WM demands could make the contribution of WM more obvious. For example, this could involve larger set sizes, longer delays, or more complex tasks where WM's role is more integral. In these cases, we would expect increased reliance on WM which would hopefully allow for more discernible effects of WM enhancement and disruption.

There are several limitations of the current study that may have affected the results. A significant portion (approx. 37%, 31%, 31%, 25%, respectively for Exp. 1a, 1b, 2a, and 2b) of the online participants failed to learn the probabilistic RL rule, which could have compromised the validity of the findings. This could be due to distractions inherent in online experiments. Conducting in-person experiments could reduce distractions and increase engagement, potentially leading to better performance and more reliable results. Alternatively, making the task more engaging could help overcome this limitation in online settings.

While the study focuses on human behaviour, RL is a computational framework that can be more deeply explored through formal models. The combination of WM

and RL has been successfully modelled in previous research (Collins, Ciullo, et al., 2017b; Collins et al., 2014). Fitting an RL model with a WM component (Collins, Ciullo, et al., 2017b) to the data could improve the validity of the approach and help better understand the interaction between WM and RL, as well as how information is processed and used across tasks.

Unlike previous studies suggesting that WM trades off with RL, our results demonstrate that actively maintaining the learned stimulus in WM can facilitate RL, challenging the prevailing view of competition between these systems. We propose that WM may function as a "stimulus buffer", providing a temporary substrate for RL, and caution that this may only happen when WM is considered a reliable source of information. By holding relevant information in WM, the system may enhance RL efficiency, for instance, by consolidating learning or enhancing attention to task-relevant features. These results motivate the development of computational models that explicitly incorporate WM as an integral component of learning. Further research is needed to clarify the circumstances under which WM can aid RL as well as the mechanisms underlying this to refine models of learning and decision-making.

5 When Are Two Systems Better Than One? A Normative Neural Network Account of Short-term and Long-term Memory Stores

5.1 Abstract

Extant models of working memory (WM) and long-term memory (LTM) have described the adaptive nature of both memory systems. Yet, the benefits of combining them are unclear and a formal implementation is lacking.

Here, we developed a neural network architecture that combined components representing WM (a recurrent Buffer) and LTM (a Hopfield model), designed to exhibit established hallmark features of the respective systems.

Results showed that memory stores behaved in the predicted manner when used separately, with the Hopfield model performance declining as its capacity was exceeded, and Buffer error increasing with interference from incoming memories and decay from neural noise (Simulation 1). When combining both Buffer and Hopfield models and encoding memories into one of the stores at a time, there appeared to be an optimal parameter governing memory allocation to the stores (Simulation 2).

Using this switch parameter, we manipulated the number of memories encoded into each store based on their prospective durations. Memory error of the model was evaluated in test environments stochastically diverse in density (time between memory arrivals) and retention duration (time between encodings and retrievals).

The Buffer appeared to be specialised for "slow" environments characterised by low memory arrival rates, whereas "bursty" environments of faster arrival times

overwhelmed it with interference from new memories. The Hopfield net, while having a higher performance baseline, experienced an attenuated effect, shifting the optimal switch parameter (Simulation 3).

Buffer performance further suffered when required to store memories for a long duration, due to increased potential for interference and accumulating noise during the delay, making it more beneficial to encode long-duration memories in the Hopfield model (Simulation 4). A major insight was that usage of the combined model was advantageous in all environments.

Our research thus provides a normative account of system specialisation for adaptive memory as well as a model implementation to further investigate this topic. We discuss possible modifications to the model that might extend its set of use cases.

5.2 Introduction

While both are memory systems which separate information from its immediate sensory source, working memory (WM) and long-term memory (LTM) have specialised roles in cognition. WM is considered flexible and dynamic, used to temporarily hold and manipulate information, while LTM is considered stable and enables memories that can last a lifetime. Many models exist to model behavioural and neural phenomena of WM (Compte, 2000; Zipser, 1991), and LTM (Hopfield, 1982; Howard & Kahana, 2002; McClelland et al., 1995) separately, and recently researchers have begun to connect established models in order to model interactions between the two systems (Hedayati et al., 2022; Lindsey & Litwin-Kumar, 2022). Here, we rethink the specialisation of the two systems, exploring the normative question of why they exist in their current form. This issue can be thought of as a fundamental allocation problem: Given a pool of neurons, how should we allocate them to build a system with specific attributes? We developed a simple model that combines WM and LTM modules to encode memories given by different environments to answer this question.

WM and LTM are complementary in many ways. To begin with, WM has a limited capacity deemed to be around four or seven items (Cowan, 2001; Luck & Vogel, 1997; Miller, 1956), while LTM has a seemingly endless capacity, with humans being able to recognise and distinguish a vast amount of items after seeing them only once(Brady et al., 2008; Standing, 1973). Furthermore, WM has been shown to support high-precision memory that can be flexibly allocated, according to task demands (Bays

et al., 2009; Ma et al., 2014). It can be sustained in different representational formats and states of priority (Fougnie et al., 2010; Griffin & Nobre, 2003; Myers et al., 2018; Vergauwe & Cowan, 2015). In contrast, while some claim similar levels of precision for LTM, at least under certain circumstances (Brady et al., 2013; Miner et al., 2020), LTM is thought to have lower precision than WM (Biderman et al., 2019). Finally, it relies on long-term potentiation for storage of memories, which is considered to create robust representations that are costly to change, requiring retrieval and reconsolidation (Alberini & LeDoux, 2013).

Finally, despite supporting high precision memory and active maintenance of information for long periods given it is rehearsed, it is very fragile in the face of different types of interference, including externally from new sensory information and internally between items already stored in WM (Oberauer et al., 2016). It is thought to be sustained by persistent activity (Constantinidis et al., 2018; Goldman-Rakic, 1995) as well as activity-silent short-term synaptic changes (Mongillo et al., 2008; Stokes, 2015) in prefrontal cortex (PFC) and other brain areas. Even in the absence of interference, noise in neural populations leads WM activity to drift, and thus the original representation to decay, resulting in higher memory error (Panichello et al., 2019; Rademaker et al., 2018). LTM, on the other hand, is regarded as having relative temporal stability, although this is marked by gist-based rather than detailed knowledge (Nagy et al., 2020; Schacter et al., 2011).

Many models of working memory exist at different explanatory levels. A prominent neural network model is the bump attractor model, which shows that persistently

firing models can sustain a stable state through recurrent excitation and inhibition between neurons (Compte, 2000; Durstewitz et al., 2000; Zipser et al., 1993). Matching empirical findings showed that activity in these networks diffuses over time due to neural noise, posing a limit on memory (Burak & Fiete, 2012). Degradation due to neural noise has further been found to impact feature precision and binding in neural population models (Bays, 2014, 2015; Matthey et al., 2015; Schneegans & Bays, 2019).

Various models of episodic memory have been proposed, often centring on the hippocampus in particular (Gluck et al., 2003; Norman et al., 2008; Schapiro et al., 2017; Sučević & Schapiro, 2023). The Hopfield model, a biologically plausible autoassociative network (Hopfield, 1982; Rolls, 2010), has been especially popular to model human episodic memory (Schapiro et al., 2017; Sun et al., 2023; Whittington et al., 2021).

Both WM attractor models and LTM Hopfield models have been combined with other network architectures to explain the interplay between cognitive systems. For example, several models of WM connect recurrent attractor networks as a proxy for PFC to sensory stores in order to explain feature binding and interference, as well as mixed selectivity in PFC (Manohar et al., 2019; Swan & Wyble, 2014), or model the sensory store as a bump attractor to explain sustained activity in sensory cortices (Bouchacourt & Buschman, 2019). For LTM, several models have used a Hopfield model as a proxy for the hippocampus and an episodic memory component in conjunction with other network components (Helfer & Shultz, 2020; Sun et al., 2023) when explaining the rapid encoding and slow consolidation of memories into neocortex posited by complimentary learning systems theory (McClelland et al., 1995; Schapiro et al., 2017).

However, few of these models have integrated the components of WM and LTM. Recently, Hedayati et al. (2022) combined an autoencoder, a deep learning model that learns latent long-term representations by minimising the memory reconstruction error, with a feature binding model intended to represent WM (Swan & Wyble, 2014), which proved fruitful in modelling several WM effects. Moreover, (Lindsey & Litwin-Kumar, 2022) proposed a theory that relies on recall-gated plasticity. The key component in their model was a mechanism by which long-term learning prioritises storage of information that was consistent with updates in the short-term memory system. In other words, repeated storage of the same information in WM leads to storage in LTM. This mechanism shields LTM from spurious synaptic changes and makes it focus on more reliable signals in the environment (Lindsey & Litwin-Kumar, 2022).

Another line of research frames WM as an information theoretic issue. Rather than focussing on the implementation level, these studies view memory as a noisy information channel (Koyluoglu et al., 2017; Sims et al., 2012). The goal here is to optimise the channel to efficiently represent as much information as possible, usually relying on including LTM as prior expectations of the memoranda (Bates et al., 2019; Sims, 2018; Taylor & Bays, 2018).

The current study conceptually integrates these approaches. We combined a simplified recurrent attractor designed to stand for a WM "Buffer" with a Hopfield model representing an episodic LTM, each exhibiting hallmark attributes of the two systems. In order to determine under which circumstances the combined model is preferred over the single modules, stochastic environments were constructed to display

different properties designed to tax both modules. Running simulations with both models, we examined performance of the combined model with different proportions of memories assigned to the Buffer versus the Hopfield model.

5.3 Methods

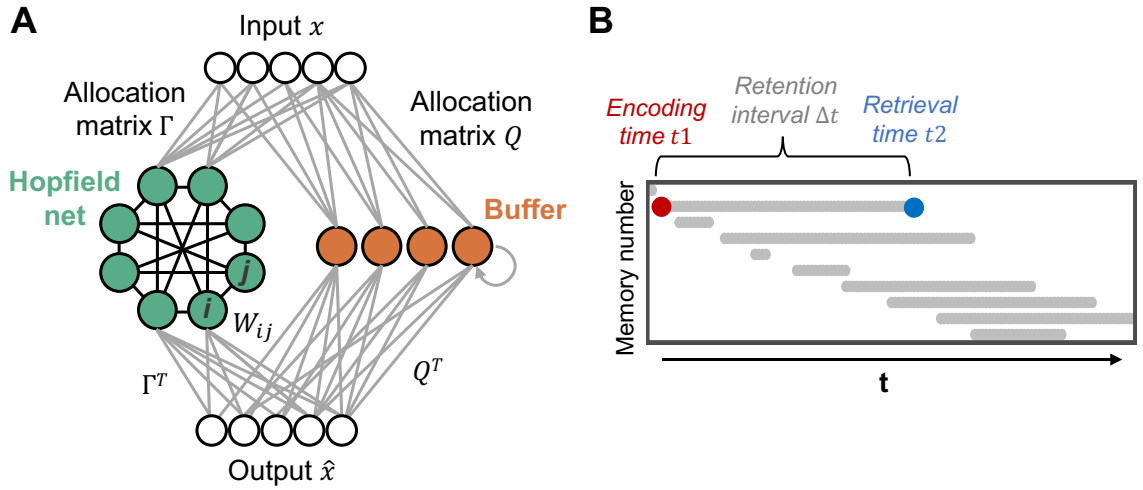
5.3.1 Model architecture

The combined model is shown in Figure 5.1A and consisted of two modules, which are proxies for LTM and WM. The LTM module was a classical Hopfield model (Hopfield, 1982). It can encode a memory after seeing it once, akin to human episodic memory. Hopfield models have often been used to model human hippocampal function in episodic memory (Sun et al., 2023; Whittington et al., 2021). In this model, the Hopfield model thus served as a proxy for LTM. The Buffer, on the other hand, represented a WM system, comparable to a simplified PFC exhibiting recurrent firing rate activity.

In our simplified task, the environment created memories that were presented to the model at time t_1 together with an expected retention duration (Δt). The model thus knew at encoding time how long memories need to be stored for, a choice we made for simplicity (see below, 5.3.3 Assumptions). They then were encoded into either the Buffer or the Hopfield model, but not both at once. Encoding took place according to an expected duration of the memory: "long" memories were encoded into the Hopfield model and "short" memories were encoded into the Buffer. The cut-off for what constituted a "long" versus a "short" memory was determined by an optimisable

parameter θ in the model (see below). Memories were then retained for an interval Δt until they were retrieved at retrieval time t_2.

Figure 5.1
Model architecture and environment



(A) Model architecture. The model consists of two modules, the Hopfield model and the Buffer. Inputs arrive and can be allocated to either the Hopfield model via allocation matrix Γ or to the Buffer via allocation matrix Q. At retrieval, memories are retrieved from their respective storage locations. Not all weights are shown. **(B)** Environment. We consider a simple environment in which the model was presented with a continuous stream of memories. Memories arrive and are encoded at time t_1. They then have to be maintained for a retention interval Δt until they are retrieved at time t_2.

5.3.2 Environment

The environment is depicted in Figure 5.1B. We consider a simplified framework that determines the content and timing of memories. A given memory x has an associated arrival time t_1 and retrieval time t_2. The arrival time t_1 of memories was modelled as a Poisson process, meaning the intervals between memory arrivals follow an exponential distribution: $t \sim \text{Exp}(\lambda)$, where λ was the rate parameter representing the average arrival rate of memories. Here, higher values corresponded to longer wait

times until the next memory arrives, resulting in a "slow" environment. Lower values, on the contrary, yielded a "bursty" environment.

The retention interval Δt was sampled from one of three probability distributions, and was used to determine the retrieval time $t_2 = t_1 + \Delta t$.

1. **Balanced Environment**: All durations were equally likely in this environment, modelled using a uniform distribution $\Delta t \sim U(a, b)$.

2. **Short-term Environment**: A scenario where many memories are needed over a short duration was modelled using a truncated log-normal distribution

$$\begin{aligned} F(\Delta t; \mu; \sigma) &\sim \text{Lognormal}(\mu, \sigma) \\ F(\Delta t | a < \Delta t \leq b) &= \frac{F(\Delta t; \mu; \sigma) - F(a; \mu; \sigma)}{F(b; \mu; \sigma) - F(a; \mu; \sigma)}, \end{aligned} \tag{8}$$

   with a and b determining the upper and lower limit of values and μ and σ representing the mean and standard deviation of the distribution, respectively.

3. **Long-term Environment**: This environment yielded a majority of long memories, while still allowing for some short ones. Retention intervals were drawn from a mirrored truncated log-normal distribution

$$F(\Delta t; \mu; \sigma) | a < X \leq b) = b - F(x | a < X \leq b). \tag{9}$$

The first simulations in this chapter used a balanced environment, until comparing environments became the key manipulation. Parameters were set to $\mu = 1$ and $\sigma = 2$ for cases 2 and 3 and distributions were truncated to restrict values to the interval $[0; 50] (a = 0, b = 50)$ for all distributions.

Hopfield model In our simulations, the Hopfield model (Hopfield, 1982) was designed to work as a proxy for LTM. The associative network usually consists of N_h binary neurons, interconnected by a set of weights $W_{N_h \times N_h}$, which describe the interactions between neurons. Neurons have an "off" and an "on" state, taking the values of -1 and 1, respectively. We adopted these conventions for our Hopfield model. The aim of the network was to store patterns of states across all neurons. In our case, the model was intended to store a vector $x \in R^n$ with binary activity ($x_k \in \{-1, 1\} \forall k$), corresponding to a single memory k. The network's storage capacity was given by $C = \frac{N}{2 \ln(N)}$ since we used a Hebbian learning rule (Storkey, 1997), corresponding to ~ 0.14 of retrievable memories per neuron.

To add a memory k to the Hopfield model, an allocation matrix Γ projected the input vector x onto the Hopfield neuron activity h. This was done to ensure that the memory could be stored in the Hopfield model when the input vector $x[k]$ was of different dimensionality than N_h. The allocation matrix Γ was shared between memories and formed by Hebbian learning between random activity patterns ($h[k] \in \{-1, 1\} \forall k$) for the Hopfield model and the associated memory input $x[k]$

$$\Gamma[k+1] = \Gamma[k] + h[k]x[k]^T. \tag{10}$$

Memories were then encoded into the Hopfield model's weights using a Hebbian learning rule. At each encoding of a memory, the Hopfield activity h_k updates the weights as

$$W[k+1] = W[k] + \sigma_h h[k] h[k]^T \tag{11}$$

where σ_h was the noise the activity in the Hopfield model activity incurred upon encoding.

At retrieval, part of the original memory vector x was used as a probe to initiate pattern completion. Let $\tilde{h}$ be a corrupted version of the memory $h[k]$ with 20% of its components set to zero. Then

$$\hat{h}[k] = sgn(W[k]\tilde{h}[k]) \tag{12}$$

where $\hat{h}[k]$ was the network's reconstruction encoded activity pattern.

Next, the reconstructed activity was transformed back into the original memory space by applying the transpose of the allocation matrix Γ

$$\hat{x}[k] = \Gamma[k]^T \hat{h}[k]. \tag{13}$$

Finally, we assessed the recall performance of the network by evaluating the reconstructed memory $\hat{x}[k]$ against the original memory $x[k]$, using mean squared error (MSE)

$$\text{MSE[k]} = \frac{1}{n} \sum_{k=1}^{n} (x[k] - \hat{x}[k])^2. \tag{14}$$

Buffer In this model, the Buffer serves as a stand-in for WM, comparable to the PFC. It consists of a set of neurons N_b. A set of weights Q was used to input the memory vector x into the Buffer. The Buffer contains four independent slots, each capable of holding one memory at a time. They can be viewed as simplified bump attractors – while excitatory and inhibitory connections between neurons are not explicitly modelled, the Buffer maintains information that due to noise drifts from its initial state. This was based on canonical WM literature showing that humans can maintain up to four items at a time in WM (Luck & Vogel, 1997). The exact number and the representational format of WM are controversial (Ma et al., 2014). However, as a first step, we adopted this model and will discuss alternative approaches (see 5.3.3 Assumptions; 5.5 Discussion). Neurons were evenly distributed across the four slots, with $N_{\text{slot}} = \frac{N_b}{4}$ Buffer neurons per slot.

To encode a memory k in the buffer, the input vector was projected onto the target neurons in slot s, resulting in activity b_{s_t} in the same slot

$$b_s[t] = Q_s x[k], \tag{15}$$

in which Q was an orthogonal matrix we obtained by QR decomposition as implemented in the MATLAB qr function. $Q[s]$ are the columns of Q which select the neurons used to encode the memory in slot s. Due to its orthogonality, it allowed us to store items in different slots independently of one another. As a result, the subset of

neurons in each slot was involved in storing the same memory, and not simultaneously involved in storing other memories.

After a memory had been encoded, the Buffer activity was maintained with added noise

$$b_s[t+1] = b_s[t] + \sigma_b \epsilon \qquad \epsilon \sim N(0,1) \tag{16}$$

in which σ_b was the level of Gaussian noise ϵ applied at each time step, mimicking drift over time seen as a result of neural noise. This was proportional to the number of neurons in the Buffer to keep signal to noise ratio equal across different buffer sizes.

$$\sigma_b = 2\sqrt{\frac{1}{N_b}}. \tag{17}$$

Noise in the neural representations has been thought to degrade memories over time. This was because random noise leads to the diffusion of memories over time (Rademaker et al., 2018; Schneegans & Bays, 2018), which over time increases memory errors (Bays, 2014; Panichello et al., 2019).

When a new "short" memory arrived, this was encoded into the buffer as above (Eq. 16). The oldest memory, i.e. the memory with the longest elapsed time t' since encoding, was overwritten to make room for the new memory.

At retrieval time, the memory was recovered from the slot into which it was encoded

$$\hat{x}[k] = Q_s^T b_s[t_2]. \tag{18}$$

Procedure for One Trial Each trial contained a fixed number of memories $n = 100$, arriving in succession, with inter-memory intervals determined by the arrival time parameter λ. When the current time equalled a memory's arrival time ($t = t_1[k]$), its prospective retention interval $\Delta t[k] = t_2[k] - t_1[k]$ was compared to the model's switch parameter θ. If it was classified as a "short" memory ($\Delta t \leq \theta$), the memory was encoded into the Buffer. Conversely, if it was classified as a "long" memory ($\Delta t > \theta$), the memory was encoded into the Hopfield model, as previously detailed. Thus, the higher the value of θ, the fewer memories would be encoded in the Hopfield model, and the more would be encoded in the Buffer. Note that, for this version of the model, information was stored in one *or* the other component. Communication between the components was not assumed, nor could a memory simultaneously be stored in both (see also 5.3.3 Assumptions). During the retention phase, activity in the Buffer accumulated noise. At each time step, if a memory's retrieval time coincided with the current time ($t = t_2[k]$), retrieval was initiated. As with the Hopfield model, we used MSE to calculate memory error.

5.3.3 Assumptions

To simplify the model and the environment, we make the following assumptions:

1. Unless we are learning information for a time-bound test or setting an intention to remember something for a certain time point in prospective memory, the duration of a memory is rarely known in advance. Here, we assume that the model knows how long it must retain a memory ahead of time (retention interval Δt) and uses

this information to decide which store to encode the memory in. In reality, this duration might be explicitly estimated based on factors such as habits and goals. Alternatively, it might depend on reliability of stimuli in the environment, akin to the process in systems consolidation (Sun et al., 2023), or on the number of previous retentions in WM (Lindsey & Litwin-Kumar, 2022).

2. We assume that the model has perfect executive control: each memory is stored in the correct store based on knowledge of the retention interval and the switch parameter θ. In the simulation there are no encoding errors as such, and no accidental encodings.

3. The model maintains multiple pointers at no additional cost: it knows which store is used for a given memory and, within the Buffer, it knows which slot a memory is stored in. This introduces an additional memory cost that is not explicitly modelled. Computationally, this acts as a shortcut, as the same could be achieved by comparing the probe cue to the stored memory and retrieving the best match. However, using this shortcut prevents modelling swap errors, that is, the retrieval of a different item in memory than the probed.

4. We assume that there is only one type of interference in the Buffer. While incoming memories cause old ones to be overwritten, we do not model inter-slot interference, i.e. interference between already stored items. As a result, the model is not capable of making "swap" errors, frequently seen in human participants (Bays, Wu, & Husain, 2011).

5. Similarly, there is no targeted forgetting in the Hopfield model. As soon as a memory is encoded, it can be reliably retrieved unless the network is at capacity. Although this is not biologically plausible, it suffices for the scope of this project. In our setting, forgetting memories to free up space would be equivalent to increasing capacity or encoding fewer memories in the Hopfield model.

6. The Buffer can currently only store discrete items, i.e. all slots in the buffer are of equal dimensionality (Zhang & Luck, 2008), however, this is at odds with empirical findings showing continuous allocation of WM resources (Bays & Husain, 2008).

7. The environment is simplified in that memories are random and statistically independent of each other. There is no correlation in their content, encoding and retrieval times, or retention intervals. This is unlikely as bits of information are usually temporally and thematically correlated.

While the assumptions listed here may affect biological plausibility, we do not expect them to qualitatively alter our results. Our model captures important differences between the two memory systems and provides a starting point for investigating them together. Further research for example might elaborate on this work by implementing the Buffer as spiking networks or mathematical analytic solutions.

5.3.4 Simulations

To test whether both memory modules behave as expected, we ran them separately at first (see Result 1). Additionally, we ran simulations to demonstrate the failure

of each system (see Appendix 7.3). Subsequently, we allowed the encoding of memories into both stores and assessed how the modules trade off in a balanced environment (see Result 2). We also evaluated the effect of different arrival rates in the environment and how the switch parameter θ needs to be adjusted (see Result 3). Finally, we investigated how the model responds to different distributions of retention intervals and arrival rates (see Result 4). Parameter values for each run are detailed in the respective results sections.

5.4 Results

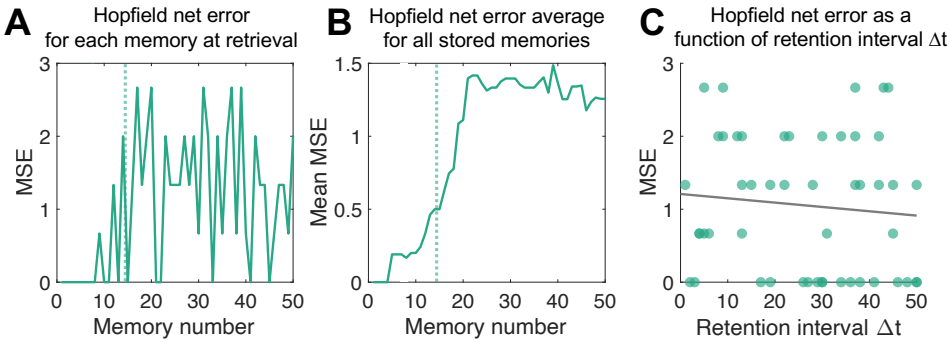
1a The Hopfield model stored memories perfectly, but its performance declined as its capacity was exceeded

The Hopfield model was expected to retain memories up to a set size equal to its capacity, robust to interference. It should be robust to decay once memories are stored and recall the memories it stored below the capacity limit perfectly regardless of retention interval.

In order to demonstrate its properties, we ran the Hopfield model for a single trial each, with parameters N = 100 (C $\approx$ 10.85), a balanced environment, $\Delta t \sim$ Uniform $[0, 50]$, arrival rate $\lambda = 1$, set size = 50.

First, we evaluated each memory at its retrieval time t_2. Figure 5.2A shows MSE for each memory based on how many memories were stored at its retrieval. The Hopfield model retrieved memories with perfect accuracy, until it hit its capacity limit (dotted line in Fig. 5.2A&B), and the memory error increased thereafter. Figure 5.2B

Figure 5.2
The Hopfield model stores memories reliably until it hits its capacity limit



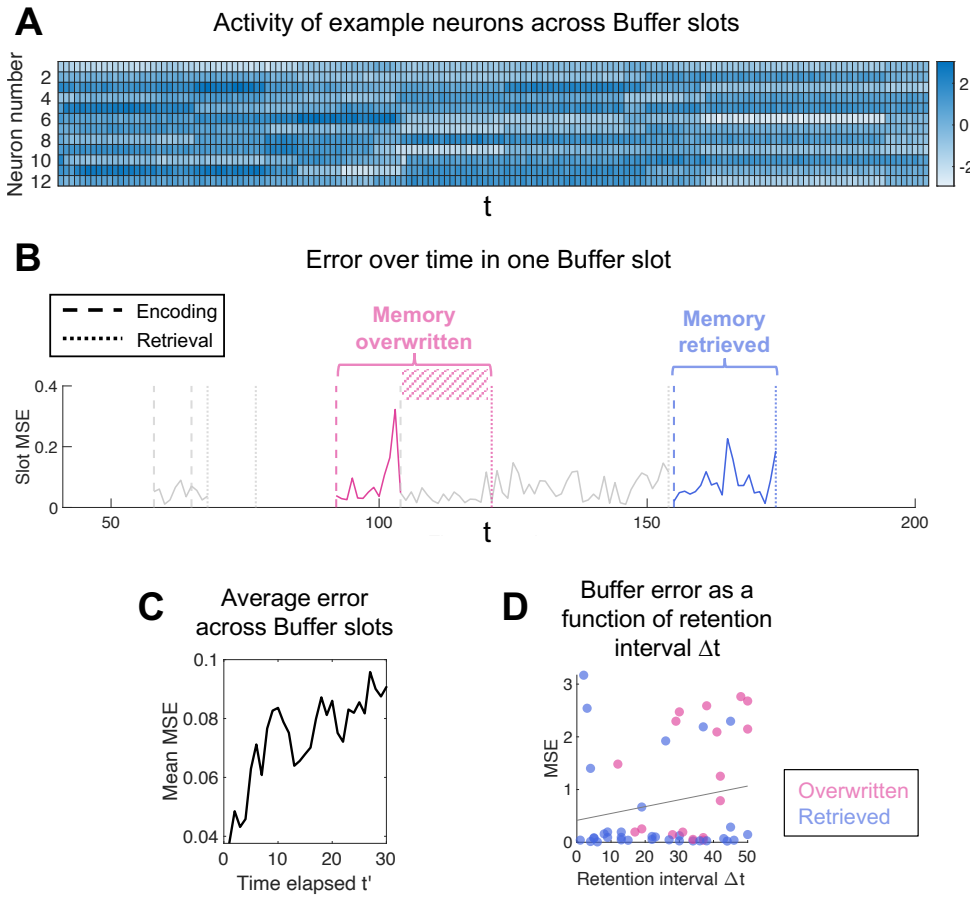
Single trial shown. Dotted line represents the capacity limit. **(A)** MSE in the Hopfield model as a function of the number of patterns stored. MSE was evaluated for each memory at its retrieval time t_2. **(B)** Average mean-squared error in the Hopfield model across previous memories as a function of the number of memories stored. **(C)** Hopfield MSE as a function of retention time Δt. Dotted lines represent the Hopfield model's capacity (~ 0.14 memories per neuron).

shows the average error for all memories previously encoded when retrieving a certain memory, i.e. increment 10 of on the x-axis shows the mean across memories 1 to 10 on the y-axis. As expected, memories that were stored before reaching the capacity limit were recalled perfectly. When the capacity limit was reached, the network failed to recover them. There was no relationship between the performance and retention interval Δt, suggesting that representations were maintained robustly after encoding (Fig. 5.2C). Once memories were stored, they could be retrieved without error, but the network failed when memories stored exceeded its capacity.

1b Memories in the Buffer were susceptible to noise and interference from incoming memories

Next, we tested the Buffer. Its error was expected to increase as a function of time, with memories being overwritten, to reflect the limited capacity of WM.

Figure 5.3
Buffer performance.



(A) Neuronal activity in the Buffer for 20 example neurons across slots. Abrupt activity changes can be seen in some neurons, indicating a new memory being encoded. **(B)** MSE in one Buffer slot evaluated at every time step after encoding. Dashed lines represent a new memory being encoded, dotted lines represent a memory being retrieved. Two cases are highlighted. The blue memory degrades over time but was retrieved at its designated t_2 whereas the pink memory was overwritten before it could be retrieved. **(C)** MSE in the Buffer for all slots as a function of time elapsed after encoding (t'). On average, memories degrade linearly. **(D)** Buffer MSE as a function of retention interval Δt. As the retention interval increases, memories are more likely to be overwritten, increasing their MSE compared to retrieved memories.

As memories are encoded in slots, neuron activity changes abruptly for neurons included in that slot, but continues to drift for neurons not involved in encoding the new memory (Fig. 5.3A). Figure 5.3B shows the average memory error across all memories in the experiment for one Buffer slot. To see how memory error would evolve, we

evaluated it at every time point from the time the memory was encoded (dashed lines in Fig. 5.3B), rather than only at retrieval time points. Error in the slot increased from encoding until the memory was retrieved (dotted lines). When the slot was needed for a new incoming memory, the old memory was overwritten (pink memory). By the time the original memory was retrieved (pink dotted line), it was not available, and the memory error will be higher. However, in the absence of an incoming item interfering, the memory was retrieved with some decay only from random noise (blue line). Across all memories, the error in all slots increases linearly with time due to noise accumulating (Fig. 5.3C).

This shows how the Buffer was susceptible to interference from incoming memories, resulting in higher error after a memory was retrieved. Figure 5.3D shows that this, too, was dependent on time. The longer the retention interval, the more likely the Buffer was to encounter interfering items, overwriting memories. This was exacerbated when retention intervals were extremely long (Fig. S2).

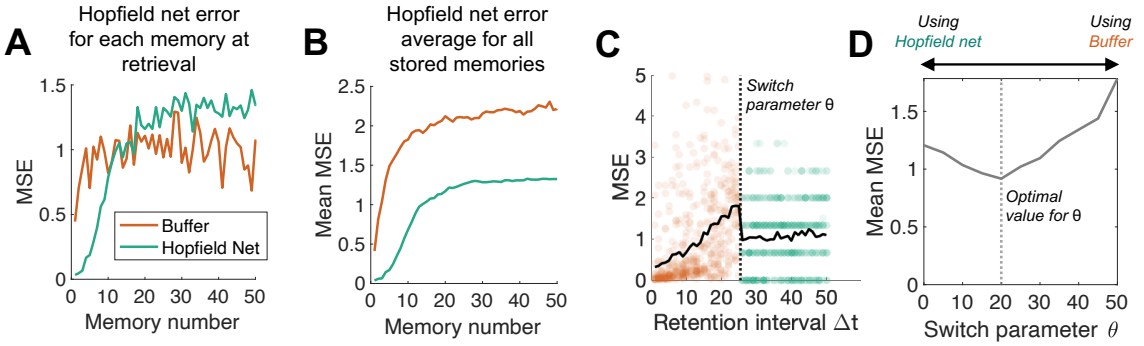
2 Hopfield model and Buffer trade off – the longer a memory was needed, the more beneficial it was to store it in the Hopfield model

To illustrate the differences between the two memory modules, we ran a simulation for 100 trials for each parameter combination. The parameters were set as follows: $N_h = 100,\ N_b = 100, \lambda = 2, \Delta t \sim U(0, 50), n = 100,$ with the workload shared equally between the networks (switch parameter $\theta = 25$).

Figure 5.4A displays the memory error for each memory at the time of retrieval. In this scenario, the Buffer error remained constant after storing the first four items, indicating that since its performance did not depend on the number of memories stored. In contrast, the Hopfield model exhibited an increase in error once its capacity limit was reached. When evaluating the memory error for all previously stored memories (Fig. 5.4B), the Buffer error increased sharply, as any memory could only be stored until the slot was needed again. Any memory stored in the Buffer would only survive three interfering memories before being overwritten. Thus, as memory numbers increased, the proportion of older memories which could be recalled was small. The Hopfield model, on the other hand, demonstrated a slower increase in error due to its higher capacity and the lack of noise accumulating as a function of time.

As shown in Figure 5.4C, the Buffer's performance was particularly sensitive to the retention interval, as error increases with retention intervals. When switching encoding to the Hopfield model (at $\Delta t = 25$, dotted line), the resulting error was smaller compared to the Buffer. This indicates that there was an optimal threshold for distributing work between the Hopfield model and the Buffer. To further explore this, we performed a grid search: we varied the values of θ from 0 to 50 in increments of 5 and ran 20 trials per parameter combination, evaluating model MSE for each. This confirmed that there was an optimal θ, corresponding to the minimum of the MSE function (dotted line in Fig. 5.4D).

Figure 5.4
Performance in the combined model



Switch parameter θ determines that the Buffer stores 50% of the memories.
(A) Both Hopfield and Buffer error rates are shown. As more memories are stored, the error rates for both systems increase. The Buffer's error plateaus sooner and at a lower level than that of the Hopfield model, which was more sensitive to the number of stored memories, leading to a later plateau at a higher error. **(B)** Average MSE for all memories encoded until the retrieval of a specific memory. The Buffer could maintain only four memories at a time, resulting in a drastic increase in average error after exceeding this limit. Even after reaching its capacity, the Hopfield model could still retrieve memories stored earlier, keeping its average error lower. **(C)** MSE as a function of retention interval Δt. The error in the Buffer increased with an increasing Δt, suggesting that some memories should be stored in the Hopfield model, as its error remained lower with half the memories stored. **(D)** MSE as a function of the switch parameter θ. The dotted line indicates the optimal value of θ that minimises overall model error.

3 Buffer error increased in bursty compared to slow environments and shifts the optimal storage criterion

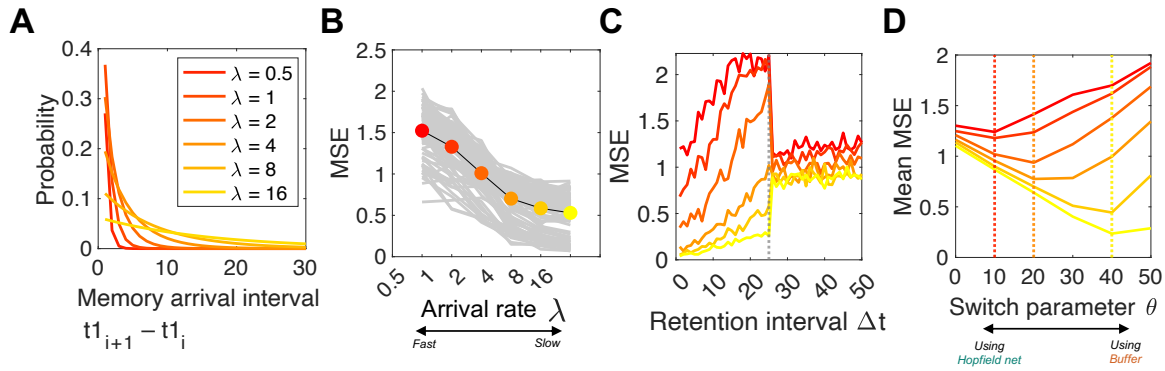
To assess how a faster-moving environment would affect model performance, we simulated 100 trials, varying the arrival rate λ while keeping other parameters constant. Figure 5.5A illustrates the distribution of inter-arrival times resulting from the different λ values used in this simulation. The values of $\lambda = \{0.5, 1, 2, 4, 8, 16\}$ indicate the average arrival rate, with higher values corresponding to longer wait times until the next memory arrives. Thus, a higher λ indicates a "slower" environment, whereas a lower λ indicates a "burstier" environment.

The combined error decreases as a function of λ (Fig. 5.5B), showing that the model was better at handling slower than faster environments. As Figure 5.5C shows, this was overwhelmingly driven by the Buffer error: Buffer error increased as a function of θ, above and beyond the effect of the Δt, whereas the Hopfield model error only slightly increased with λ, but remained unaffected by Δt.

As mentioned, the Buffer error interacted with the retention interval: the longer the retention interval and the shorter the intervals between arrivals, the more likely memories were to be overwritten. Supplementary Figure S4 illustrates this effect for different switch parameter values.

We finally varied the switch parameter θ (from 0 to 50 in increments of 10, with 20 trials per combination) to investigate the effect of arrival rates on the optimal trade-off of memory modules (Fig. 5.5D). As expected, the arrival rate also influenced the optimal switch parameter (dotted lines): since the Buffer was more adversely affected by faster arrival rates than the Hopfield model, shifting the criterion towards using the Buffer resulted in higher error variability. Preferred storage in the Buffer (high θ) proved optimal in slower environments but detrimental in fast-paced settings. Conversely, favouring the Hopfield model (low θ) was advantageous in bursty environments where the Buffer was overwhelmed by excessive memories. However, due to the capacity limit of the Hopfield model, its overall error increases the more it was utilised. Overall, relying solely on either the Hopfield model or the Buffer was not the optimal solution for any arrival rate.

Figure 5.5
Effect of arrival rate λ and switch parameter θ



(A) Distributions of arrival intervals used in this simulation. Arrival intervals are the wait time from the arrival of one memory to the next. They are exponentially distributed with arrival rate λ. The higher the value of λ present, the longer the interval on average, leading to a "slower" environment, whereas lower values lead to shorter intervals and "bursty" environments. **(B)** MSE as a function of λ. Overall error decreases as λ increases. **(C)** When distributing memories equally between both systems ($\theta = 25$), it was evident that error increases more rapidly in the Buffer the faster memories arrive, whereas the Hopfield model was less affected. **(D)** Dotted lines represent optimal values of switch parameter θ given a certain arrival rate λ.

4 The distribution of retention interval Δt shifted the optimal criterion

Having focussed on balanced, uniform distributions, we next examined the effect of different distributions of Δt. Figure 5.6A shows the Δt distributions used in this simulation: balanced (uniform), short-term (log-normal), and long-term (mirrored log-normal distributions). We ran 100 trials with arrival rates kept constant ($\lambda = 1$), a set size of 100, and $N_h = N_b = 100$.

Figure 5.6B shows the MSE as a function of the environment and the switch parameter, with dotted lines representing the optimal switch parameter in the different environments. In the balanced (uniform) environment, using the Hopfield model was preferred, due to the moderate arrival rate affecting the Buffer more, as discussed in the last section (Fig. 5.5). In a long-term heavy case (mirrored log-normal distribu-

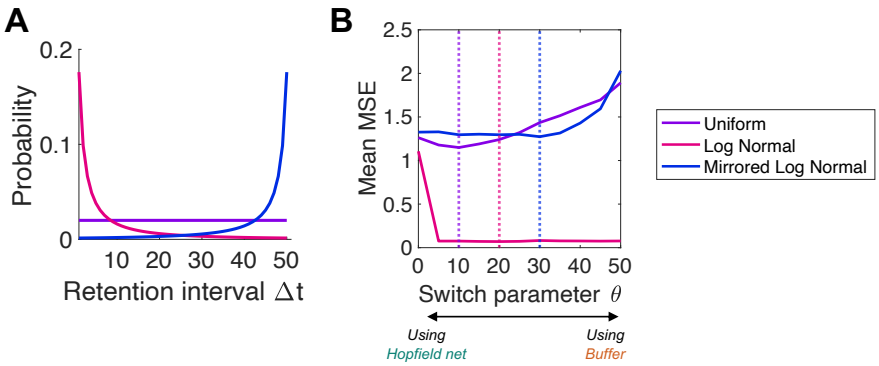
tion), selecting the Buffer was beneficial, reserving the Hopfield model for very long memories. This preference arises from the fact that the majority of the memories have long durations; therefore, a slight shift in θ will result in a significant increase in the number of memories allocated to the Hopfield model. Finally, for a short-term (log-normal) environment, the use of the Hopfield model was favoured. This indicates that the Buffer, despite being specialised for short memories, was overwhelmed when used on its own. Here, the switch was quite drastic, as there were very few longer memories, limiting the impact of a shift in θ. The complete set of simulations varying set size and arrival rate, in addition to the parameters presented here, is presented in Figure S5. The impact of long- and short-heavy log-normal distributions was extremely asymmetrical. This asymmetry results from the model architecture being able to hold four short memories simultaneously, while each longer-term memory counts toward the capacity of the Hopfield model.

Crucially, as in the previous simulation, the optimal switch parameter value was not at 0 or 50 for any environment. This indicates that the combined model overall performed better than each of the modules on their own.

5.5 Discussion

This study introduced a neural network model that combined a WM and a LTM component, represented by a recurrent Buffer and a Hopfield model, respectively. We examined the theoretical question of why neurons should be allocated to each store, given that the system will be confronted with varying environments. As a first step, we

Figure 5.6
Effect of the distribution of Δt



(A) Illustration of the distributions of retention interval Δt tested, uniform (purple), log-normal (pink), and mirrored log-normal (blue). **(B)** Average MSE as a function of switch parameter θ and tested environments. Dotted lines represent the optimal switch parameter θ. $\theta = 0$ resulted in the only the Buffer being used, while $\theta = 50$ resulted in only the Hopfield net being used. In all three cases, relying on both systems together was beneficial. In the uniform environment, it was best to use both stores in a balanced manner. In a fast-paced environment (pink), overusing the Hopfield model (lowering the θ) drives up the error, as the memories will exceed its capacity. In a slow-paced environment, on the other hand, overusing the Buffer (increasing θ) increases the error due to increased interference.

demonstrated that the combined model exhibited the respective hallmarks of both WM and LTM. In the Hopfield model, memories were stored perfectly, but its performance declined when its capacity was exceeded. Memories in the Buffer were susceptible to interference from incoming memories as well as to noise, with the effects being worst in long retention intervals and fast environments. By using the Hopfield model and the Buffer in conjunction to store relative short and long memories, we demonstrated that there was a trade-off between the two. The longer a memory was needed, the more beneficial it was to store it in the Hopfield model. This was due to the Buffer error increasing with retention interval due to extended opportunity for interference and noise accumulation. The Hopfield model was not sensitive to time, but performance decreased as a function of the number of memories stored. Finally, we compared the

combined model's performance between environments. In bursty environments, where memories arrive in fast succession, Buffer error was found to be higher compared to slow environments shifting the optimal storage criterion towards using the Hopfield model. Lastly, the distribution of retention intervals was shown to affect the optimal storage strategy. A key finding here was that usage of the combined model was advantageous in all environments. This demonstrates a normative account of system specialisation.

Our model incorporates important features of WM and LTM. Firstly, it reflects the limited capacity of WM, allowing only a small number of items to be held simultaneously (Luck & Vogel, 1997), as well as the much larger capacity of LTM (Brady et al., 2013). Similarly, it shows the capability of accurately storing information for a longer period in LTM than in WM, which faces interference from new incoming memories which disrupt the maintenance of older information. Additionally, it includes the time-based decay that causes memory activity in the Buffer to diffuse from the original state (Panichello & Buschman, 2021; Rademaker et al., 2018). However, the model does not address the interference that might occur between items already stored in memory.

Overall, the model effectively captures key features of the two memory systems. Furthermore, the model creates patterns of failure that are associated with the inherent limitations of both systems. It highlights the balance between the ability to quickly hold and retrieve information (flexibility) and the capacity to retain knowledge over time (longevity). Thus, the model contributes to our understanding of why a specialisation of memory systems is beneficial.

Nevertheless, some important features are missing from the model, and should be included in future versions. For example, the format of the neural activity in both stores should be reconsidered. WM is flexible, allowing for the switching between task representations (Hardman et al., 2017; Vergauwe & Cowan, 2015). In the model, the representations are always limited to -1 and 1, however, this could be modified.

A related issue is the flexibility in the allocation of resources. The current solution in the model is to store memories in discrete slots (Zhang & Luck, 2008). However, this is a controversial issue, as there is substantial evidence suggesting that WM can support the storage of information at varying fidelities (Bays, Gorgoraptis, et al., 2011; Bays & Husain, 2008; Bays, 2014). This could be addressed by changing the function of the memory allocation matrix to always use all available neurons, and reallocate them as new memories need to be encoded.

Another discrepancy is that representations in medial temporal cortex underlying episodic memory have been found to be sparse, with a small number of neurons firing strongly in response to a specific stimulus (Quiroga et al., 2008). In our model, the stored memory is a random binary pattern. Sparsity could either be achieved by using one-hot encoding vectors, where most of the values are set to zero, or using methods like compressed sensing (Ganguli & Sompolinsky, 2012) or an autoencoder (Hedayati et al., 2022) to transform sensory input into sparse latent representations.

Another aspect worth mentioning is energy efficiency. The cost of synaptic computation is very high due to the initiation of action potentials (Lennie, 2003), while forming longer-term associations reduces this cost in the long run (Harris et al.,

2012). Persistent spiking activity underlies WM, especially when information needs to be manipulated (Constantinidis et al., 2018; Kamiński et al., 2017), while long-term potentiation underlies LTM. Costly spiking neurons are important to perform computations in WM (Masse et al., 2019), while cost-economical synaptic learning is less flexible, providing another perspective on their specialisation.

A main limitation of our modelling strategy includes the lack of explicit mechanisms for the two stores to interact. While there is no direct connection between the stores, the model's retrieved memory could become a new input, and be encoded into the respective other store, creating an indirect connection. This has a potential for connection to the theory of recall-gated plasticity (Lindsey & Litwin-Kumar, 2022) in which encoding into LTM is determined by repeated storage of information in WM. However, WM and LTM are explicitly linked, with WM buffering LTM information upon retrieval (Fukuda et al., 2010), and WM acting as a bottleneck for LTM encoding (Forsberg, Adams, & Cowan, 2021). Thus, future versions of the model should focus primarily on modelling the interaction between systems explicitly.

Another limitation is the overcomplication of the model. It makes a large number of assumptions, for example the storage of pointers to index which memory store was used at no additional memory cost (see 5.3.3 Assumptions). These assumptions were made to simplify the model, allowing us to shortcut more detailed computations to focus on our question. Nevertheless, future work is needed on extending the model to minimise these assumptions, most importantly enabling interaction between the two memory stores.

To conclude, our results demonstrated that the most adaptive memory performance occurred when both memory stores were combined, thus providing a normative justification for the use of multiple distinct memory systems. This work adds a model implementation to previous research on the adaptiveness of memory and supports the idea that specialised memory stores are employed for different circumstances. Future research might extend the model to cover situations in which these memory stores might interact explicitly.

6 Changes in Neural Representational Geometry Reflect Lasting Traces of Induced Centrality Bias in Macaque Working Memory

6.1 Abstract

Biases in working memory (WM) are a well-studied topic, with a rich literature exploring various types. One key type of bias is the centrality bias – the tendency to adjust one's estimate towards the mean of a stimulus set. While a substantial amount is known about this, the underlying neural mechanisms are unclear. Here, we replicate and extend an analysis of a non-human primate (NHP) dataset to examine the neural activity underlying induced colour biases and assess their longevity. Two monkeys performed a retro-cue delayed estimation task: they were required to hold one or two colour stimuli in WM, and were cued to report one of them after a delay. In the first half of each session, colours were drawn from the full colour wheel (No Limit block), whereas in the second half, the colour space was restricted to one half of the colour wheel (Limit block) in order to induce a centrality bias. Results showed that performance was better in the Limit block, as soon as 10 trials after its onset, and that colour responses were biased towards the mean of the restricted colour space. Both performance benefit and biases persisted into the No Limit block of the testing session on the next day for colours within the Previous Limit compared to colours outside of it. When assessing this lasting benefit at the neural level, single-neuron responses showed no significant difference between the proportion of neurons selective for colours in the Previous Limit compared to colours outside. However, the variability of selective neurons' responses

was lower for Previous Limit items in prefrontal cortex (PFC) following the retro-cue, and in the frontal eye fields (FEF) following the onset of the response wheel. In addition, the average response strength of selective neurons was lower in FEF. Further, distinct neural geometry was revealed in the population response using principal component analysis (PCA). As previously established, colour responses tended to lie on a circle in two-dimensional principal component space for the No Limit condition, and on a triangle for the Limit condition, in PFC and FEF. In addition, Previous Limit trials in the following testing session were characterised by similar circle and triangle shapes in FEF, and, to a weaker extent, in PFC. Overall, our results provide evidence for a neural adaptation in frontal brain regions to the restriction of the colour space.

6.2 Introduction

Biases in working memory (WM) are common and different types have been well studied. A common type of bias observed in WM is the centrality bias, also known as the central tendency or contraction bias. Well known across cognitive domains, it describes a shift in representations within WM towards the mean of the stimulus distribution (Hollingworth, 1910). A key question remains regarding how these biases are maintained by the brain and for how long they persist. This study therefore investigates the neural correlates of induced biases and examines whether they persist into the next testing session, using data collected from non-human primates (NHPs).

Central tendency bias describes a form of target bias, reflecting an imperfect representation of the target rather than random guessing. In WM, the two can be distinguished using a Gaussian mixture model by framing the strategic guess as a "swap errors within the ensemble", i.e. model biased responses as swaps for the mean of the stimulus distribution (Huang, 2020). Another model demonstrated that participants' responses are a mixture of both the category mean value and continuous perceptual information (Duffy et al., 2010). For example, color hues are often biased toward the mean of the stimulus set, but retain fine-grained information (Olkkonen & Allred, 2014). Centrality bias is considered adaptive, aligning with Bayesian observer models as an optimal strategy. It develops over time (Jou et al., 2004) and tends to emerge during retrieval or decision-making processes rather than during encoding (Ashourian & Loewenstein, 2011; Jou et al., 2004).

The neural correlates of centrality bias appear to follow a gradient between frontal and posterior regions. In humans, frontal electrodes show coding of both simple features and the ensemble mean, while occipital electrodes do not represent the ensemble mean (Oh et al., 2019). Moreover, the centrality bias is stronger further along ventral visual processing stream and becomes more decodable in early visual areas only when features represented by these areas are task-relevant (Tark et al., 2021). Similar results have been reported in human fMRI. Information can be represented based on task requirements, with visual areas represent feature information, while frontal areas represent more abstract, category information (Lee et al., 2013). Moreover, during stimulus-specific training, frontal areas such as dorsal and ventral PFC over time exhibit heightened activity and items become more decodable from this activity (Miller et al., 2022). The notion of a posterior-anterior gradient is also supported with the prefrontal cortex (PFC) representing categories in NHPs (Freedman et al., 2001). During learning, frontal areas follow different timescales than posterior areas (Miller & Constantinidis, 2024), and even with PFC anterior regions showed enhanced plasticity after learning a WM task (Riley et al., 2018; Trepka et al., 2023). These findings suggest that judgments biased by the mean involve higher-level processing (Oh et al., 2019). These findings support the notion that representations in frontal areas commonly involved in WM can adapt through learning to support context-dependent representations in WM.

Centrality bias is moreover linked to the concept of WM ensemble representations in WM, where statistical information, like the mean size of a set of items, is remembered alongside individual items (Brady & Alvarez, 2011), improving overall ac-

curacy by using the ensemble statistic as an anchor. Humans tend to remember both the mean and variance of stimulus sets (Brady & Alvarez, 2015a), which can inflate WM capacity (Brady & Alvarez, 2015b). The narrower the range within the ensemble representation, the more likely participants are to use it to determine individual features (Utochkin & Brady, 2020). Computational models have demonstrated that ensemble representations are summed distributions of individual items (Robinson & Brady, 2023). Keeping track of summary statistics like the mean allows for strategic guessing or shifting of the response in case of uncertainty. Centrality biases may be supported by the formation of ensemble representations describing a stimulus set.

It is also related to temporal interactions between items from a set. For example, serial biases in WM arise from recent stimulus history and depend on stimulus similarity. For example, when items are encoded sequentially but held in WM at the same time, more similar orientations tend to repel each other, while more dissimilar orientations are attracted (Bae & Luck, 2017). This often results in a performance benefit, due to improved distinction of similar items or recognition of dissimilar items (Chunharas et al., 2019; Lin & Luck, 2009). The higher priority item will have stronger influence than the lower priority item (Bae & Luck, 2017) and the bias is reported to be stronger for features that were required to be reported on the previous trial (Bae & Luck, 2020). Serial biases may underlie centrality bias in that they may help estimate ensemble statistics. Rather than holding stimulus history in memory and estimating a summary statistic, this might result from serial biases. Indeed, a bump attractor network model shows that centrality bias arises from allowing responses to be influenced by recent

stimuli history (Boboeva et al., 2023). This points towards serial biases as one way of developing centrality bias.

Humans also acquire centrality biases naturally by learning or being presented with categories. For example, one study showed that in a delayed hue comparison task, hues tended to be biased towards the mean of the artificial category, a stimulus set on a yellow-green to blue colour spectrum (Olkkonen et al., 2014). Generally, knowledge about colour categories appears to influence WM in a way that colour reports are biased towards the category mean and away from category edges (Bae et al., 2015). This bias is stronger in WM than in perception (Bae et al., 2015), possibly due to mechanistically being supported by neural attractors that pull the memory representations towards stable fixed points in the feature space to protect them from interference, but in turn creating biases (Panichello et al., 2019).

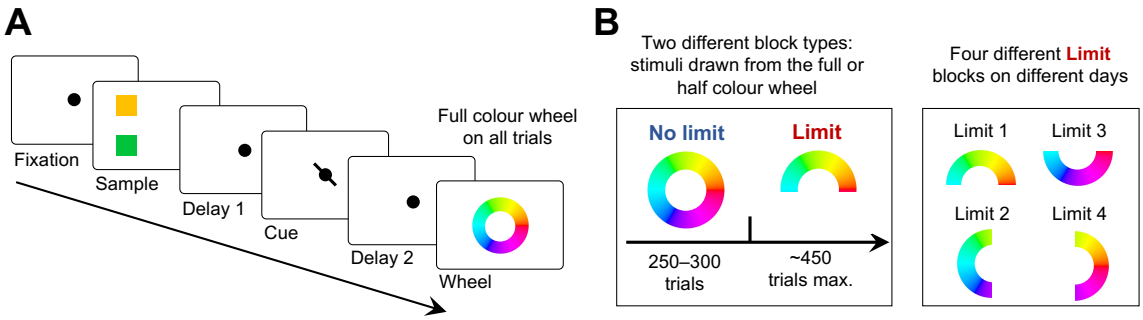
In this study, we explored the neural correlations underlying long lasting WM biases. We reanalysed an existing NHP dataset (Uchimura et al., Unpublished manuscript). In the experiment, monkeys performed a colour delayed estimation task. During the task, centrality bias for colours was induced by restricting the colour space samples were drawn from. Brain activity was concurrently recorded from several areas, such as the PFC, frontal eye fields (FEF), motor areas, posterior parietal area 7a, and the visual area comprising visual V4 and posterior inferotemporal areas (V4pIT). All these areas have been shown to be involved in WM. The PFC, as mentioned, plays a critical role in reflecting and processing categorical information (Freedman et al., 2001, 2003) and is crucial for WM to guide goal-directed behaviour (Miller & Cohen, 2001). FEF

is located in PFC, and is involved in generating voluntary eye movements, including saccadic and smooth pursuit eye movements (Schall et al., 1995; Tehovnik et al., 2000). Posterior parietal area 7a is involved in sustaining or anticipating activity in visuospatial tasks. It reflects both spatial locations and movement, providing critical input for processing spatial information (Constantinidis & Steinmetz, 1996; Pasternak & Greenlee, 2005). Finally, area V4 in processing colour and is a step in the ventral visual processing (Motter, 1994; Pasternak & Greenlee, 2005; Pasupathy et al., 2020).

In their analysis of the same dataset Uchimura et al. (Unpublished manuscript) found that centrality bias in behaviour resulting from a restriction of the colour stimulus space co-occurred with changes in the neural geometry of colour responses. Specifically, in No Limit blocks colour response appeared to lie on a circle in two-dimensional principal component space found by principal component analysis (PCA), with more similar colours appearing next to each other. After introducing the Limit block, geometry changed to resemble a triangle missing one side, with the edges corresponding to edges of the colour semicircle and the corner corresponding to its mean (Uchimura et al., Unpublished manuscript). The aim of the current analysis was to understand whether induced centrality biases would carry over to the next testing session. If they did, adaptive context-dependent changes would be long-lasting, not explainable through serial biases, and potentially conducive to long-term learning of task structure. We expected frontal areas to be involved in this, as previously shown pivotal for learning-related changes during WM tasks.

6.3 Methods

Figure 6.1
Task schematic



(A) Task schematic. Every trial started with a fixation marker, before presenting one or two coloured squares for memorisation. After a delay, the direction of a retro-cue indicated which of the stimuli should be retrieved. After another short delay, a colour wheel was displayed and monkeys could make their report by saccading to the colour they believed the target to be. A full colour wheel was shown during retrieval in all conditions to allow for the choice of colours lying outside the stimulus space in Limit conditions. **(B)** Stimuli. In every session, monkeys completed two blocks: a No Limit block (250–300 trials) followed by a Limit block (∼450 trials). The colour space of the stimuli depended on the condition. In No Limit blocks colours were drawn from the full colour circle, whereas in Limit blocks they were drawn from half a colour circle. There were four different Limits, corresponding to the four possible semicircles, with only one of the Limits selected for each session.

6.3.1 Subjects

Two adult male rhesus macaques (Macaca mulatta, 10–12 years old, weighing ∼11kg and ∼8.5kg, respectively) completed this experiment. The number of subjects is consistent with previous work (Panichello & Buschman, 2021). All procedures were approved by the Princeton University Institutional Animal Care and Use Committee and were in accordance with the policies and procedures of the National Institutes of Health.

6.3.2 Experimental design

The experimental task was presented on a Dell U2413 LCD monitor. Subjects were seated at a viewing distance of 58cm. Tasks were programmed using MATLAB (The MathWorks Inc., 2021) and Psychtoolbox (Kleiner, 2007).

Design The task design is shown in Figure 6.1A. The monkeys initiated each trial by fixating their gaze on a point in the centre of the screen. Stimuli were then presented for 500ms. After a memory delay of 500ms, 1000ms, or 1500ms, a retro-cue was displayed for 300ms, indicating whether to report the upper or lower stimulus. After another memory delay of 400ms, 600ms, 900ms, 1100ms, 1400ms, or 1600ms, a colour wheel was displayed. The monkeys made their selection via an eye movement, saccading to the section of the colour wheel that corresponded to the colour of the cued stimulus in memory.

The crucial blockwise manipulation was the restriction of the colour wheel designed to induce behavioural biases. In *Limit* blocks, stimuli were drawn from only half of the colour wheel, whereas in *No Limit* blocks, they were drawn from the full colour wheel (6.1B, right panel). There were four different restrictions, corresponding to the four possible semicircles of the colour wheel (6.1B, left panel). Critically, the full colour wheel was displayed for response on every trial. This means that, responses lying outside of the possible stimulus colours could still be made in the Limit condition and the task would be equally hard across conditions. In other words, any performance benefits

between conditions relied on the monkeys learning the colours from the restricted circle throughout the experiment.

The particular Limit that was applied changed for every session, with no particular order. For example, session one might include a Limit block with stimuli based on Limit 1, and session two a Limit block based on Limit 3. Limits 1 through 4 were used in 8, 10, 12, and 13 sessions, respectively, of which 7, 6, 5, and 5 were recording sessions. At the start of every session, monkeys completed 250–300 trials in the No Limit block, followed by up to 450 trials in the Limit block.

Stimuli The stimuli were coloured squares that spanned 2 degrees of visual angle (DVA). Their colours were drawn at random from 64 (or 32 in Limit blocks) evenly spaced points on an isoluminant circle in CIELAB colour space ($L = 60, a = 6, b = 14$). They were displayed at two possible positions in each hemisphere: 45 degrees clockwise or anticlockwise from the horizontal midline and at 5 DVA distance from fixation. Durations for memory delay 1 (after Sample presentation) and delay 2 (after Cue presentation) were chosen at random. Retro-cues (2DVA extension) and the response colour wheel (2 DVA thick, extending 5 DVA at the outer radius) were both presented at fixation. To distinguish the visual appearance of the cue from its meaning, two types of retro-cues (circle with directed lines and a triangle cueing upper and lower visual field, respectively) were used. The response wheel was randomly rotated on each trial to prevent motor planning or spatial encoding of memories. A juice reward was administered, with the amount depending on accuracy. The number of drops awarded was determined according to a von Mises distribution ($M = 0, SD = 22°$) with the

peak amplitude scaled to 12. At a response error greater or equal to 60° there was no juice reward.

Procedure In every session, monkeys performed a minimum of 635 and a maximum of 965 trials. Responses had to be made within 2s, otherwise there was a timeout. Monkeys received a juice reward after every correctly completed trial. Eye position was monitored continuously at 1kHz using Eyelink 1000 Plus eye-tracking system (SR Research).

Monkey 1 and Monkey 2 performed 26 sessions with a total of 19,932 trials and 29 sessions with a total of 23,288 trials, respectively. They each completed six sessions with only a No Limit block and all other sessions contained a Limit block.

6.3.3 Recordings and signal processing

Monkey 1 and Monkey 2 underwent eleven and twelve recording sessions, respectively. Both were implanted with a headpost to immobilise their head during testing. Two titanium chambers were implanted to access the brain. They were positioned using 3D models of the brain from MRI scans to allow electrophysiological recordings from 5 regions of interest (ROIs): PFC, FEF, inferior parietal cortex (7ab), motor cortex, and an area comprising V4 in visual cortex and posterior inferotemporal cortex (V4pIT).

Epoxy-coated tungsten electrodes were used for recording, alongside a custom-built microdrive assembly to lower them. Recordings were acute: up to 80 electrodes were lowered at the start of each recording session and allowed to settle 2-3h before

recording. 30kHz broadband activity was recorded from each electrode (Blackrock Microsystems).

Signal preprocessing followed the same protocol as Panichello and Buschman (2021) and involved filtering signals using a 4-pole 300Hz high-pass Butterworth filter and re-referencing voltage to the median reference. Spike detection threshold for all recordings was $-4\sigma_n$, where

$$\sigma_n = median(\frac{|x|}{0.6745}) \tag{19}$$

with σ being an estimate of the standard deviation of the noise distribution.

Time points where this value was crossed with a negative slope were considered spiking events. Repeated spikes within 1.0667ms were excluded and waveforms around spike times were extracted. Spikes were manually sorted into single unit activity, multiunit activity, or noise using Plexon Offline Sorter.

Smoothed firing rates for analyses were calculated for each 10ms sampling interval. Each spike was represented as a delta function and convolved with a half-Gaussian (σ = 200ms).

6.3.4 Analysis

All complete trials where fixation was maintained and a saccade to the colour wheel was made were included in the analysis. Monkey data was pooled as it was similar regarding key effects S6A.

Behavioural analyses The full dataset (session with and without recordings) was used for behavioural analysis. Behavioural analyses were performed with R (version 4.1.2, R Core Team, 2019). Trials with a RT of $2.5SD$ above the monkey's mean RT were removed from analysis.

The analysis focussed on response error and bias. Mean absolute response error for colour memory was calculated as the absolute angular distance on the colour wheel

$$error = abs(\theta_{target} - \theta_{response}) \tag{20}$$

where θ is a colour value in degrees.

Bias was computed separately for each colour and for Limit and No Limit blocks as

$$bias = \theta_{response} - \theta_{target}. \tag{21}$$

Bias for each target colour θ in the Limit condition was then corrected for the expected bias for that colour calculated from No Limit blocks

$$bias_{\theta}^{corrected} = bias_{\theta}^{Limit} - bias_{\theta}^{NoLimit}. \tag{22}$$

To assess whether there was a bias towards the semicircle mean across Limit types in the Limit condition, we represented the target colour in terms of deviation from the semicircle mean.

Behavioural data were analysed using R (v4.1.2, R Core Team, 2019). Significance testing was performed with t-tests when there were only two factor levels. For tests with more than one factor, linear mixed-effects models (LLM) were fit using the

lme4 package (v1.1-28, Bates, 2010) and p-values were calculated with lmerTest (v3.1-3, Kuznetsova et al., 2017). We generally included interaction terms as fixed effects and the monkey as a random effect

$$y \sim x_1 * x_2 + (1|\text{monkey}). \tag{23}$$

When assessing accuracy differences between conditions across time, we tested for significance at every time point using rolling means with a window of 10 trials.

Neural analyses The first aim of the analysis was to determine whether, on a single-cell level, more neurons were selective for colours within than outside of the Previous Limit in the subsequent session. This analysis was limited to the No Limit block of the current session to avoid including trials where old and new limits overlapped. We calculated colour selectivity for every neuron separately for activity following the presentation of stimuli (Sample), retro-cue (Cue), and colour wheel (Wheel). The time window of interest was specified to be 50ms after stimulus onset until 50s after stimulus offset, which was 50ms to 550ms for Sample and Wheel events. For the activity following the Cue, we adopted a slightly delayed time window, since there has been shown to be a delay in the neural response reflecting the selection of an item from memory (Panichello & Buschman, 2021).

We discretised colours into 32 bins and averaged the firing rates separately for these time windows, for visual fields (upper and lower target presentation), and for Limit and No Limit blocks, resulting in eight groups. This provided a discrete 32-point

distribution of mean firing rates. The Rayleigh uniformity test was then performed to test for colour selectivity. If the test was significant, this indicated a non-uniform firing rate across colour bins, and would suggest that the neuron preferentially responded to a colour. We further performed a permutation test, comparing the observed z-value in the Rayleigh test to a null distribution generated by applying the same test using shuffled responses for each colour bin. The observed z-value to the null distribution was found by calculating the proportion of z-values that were higher than the observed one

$$p = \frac{\sum z_{obs} < z_{null}}{n}, \tag{24}$$

where z_{null} is a vector containing the observed z-value and null distribution z-values. To reduce computational cost, we initially generated a null distribution of $n = 500$ permutations. Then, if the p-value was between 0.025 and 0.1, we conducted a further 500 permutations to substantiate our result. The resultant vector (mean direction of the response), as well as its magnitude (strength of the response) and dispersion ($\frac{1}{\kappa}$, where κ is the centrality parameter of the von Mises distribution), were also calculated to compare between neurons with selectivity for colours within and outside of the previous Limit. All circular statistics were calculated using the Circular Statistics Toolbox in MATLAB (Berens, 2009).

Overall colour selectivities for different regions of interest (ROIs) and stimulus events are reported. The proportion of neurons selective for colours within the Previ-

ous Limit versus outside of the Previous Limit was computed. We then tested for a statistical difference between the proportions

$$z = \frac{p_{out} - p_{in}}{\sqrt{\hat{p}(1-\hat{p})\frac{1}{n_{out}} + \frac{1}{n_{in}}}} \tag{25}$$

where p_{out} and p_{in} are proportions of significantly selective neurons and n_{out} and n_{in} the total neuron count in the condition for colours outside and within of the previous Limit, respectively. MATLAB's normcdf function was used to find the associated p-value.

We followed this up with a population analysis, with which we aimed to assess whether the changes in representational geometry found by Uchimura et al. (Unpublished manuscript) would persist into the No Limit block of the next session. Since different colours were part of the Limit circle, we had a restricted number of trials available, but included both set size 1 and set size 2 trials. To increase the number of trials, the response space was rotated so that neuron responses were relative to the restricted colour space rather than colour-specific (illustrated in Fig. 6.8A). We averaged mean firing rates for 20 colour bins and organised them in a matrix $X_{m\times n}$, where m was the number of colour bins and n the number of neurons. PCA was conducted using the pca.m function in MATLAB (Fig. 6.8B). This function uses singular value decomposition to decompose the covariance matrix of X into its eigenvectors and eigenvalues. We retained the first two eigenvectors (principal components) to assess their

geometry. Using these, we constructed a 2-dimensional space onto which we projected the population response vector.

We first replicated the result of (Uchimura et al., Unpublished manuscript) by conducting this analysis for Limit and No Limit blocks, before repeating the analysis for trials within and outside of the Previous Limit in the No Limit block of the following session.

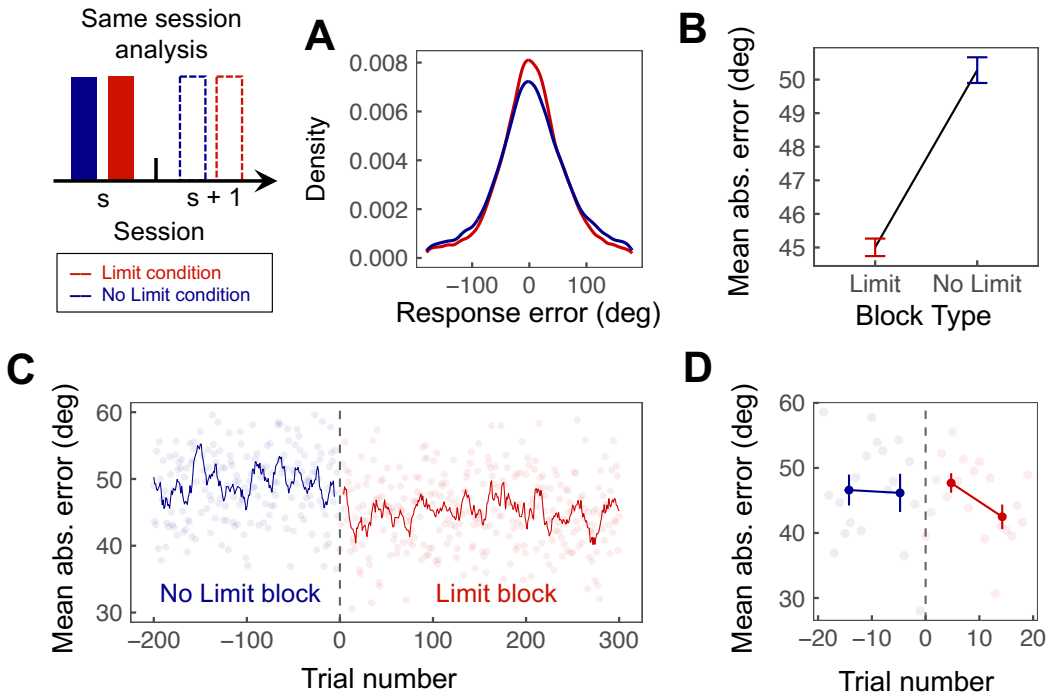
6.4 Results

6.4.1 Behavioural data

Effects of the restricted colour circle on the same session First, we analysed the effect of restricting the stimulus colour space on accuracy in the same session. Performance, as indicated by colour memory error, was better in the Limit condition than in the No Limit condition ($t(23190) = 11.39$, $p < 0.001$, Fig. 6.2A). This indicates a benefit for stimuli when the colour space was restricted.

To assess when this benefit became apparent, a rolling mean of 10 trials was used to visualise performance over time in Limit and No Limit conditions (Fig. 6.2C). Indeed, error decreased shortly after the block switch. To confirm this, we performed an ANOVA comparing groups from 20 trials before and after the onset of the Limit block in increments of 10 trials (Fig. 6.2D). This showed a trend towards a difference between trial bins ($F(3, 1673) = 2.45, p = 0.06$). Post-hoc t-tests showed a significant difference between conditions at 0-10 and 10–20 trials past the switch ($p < 0.05$), but

Figure 6.2
Better performance for targets within the colour Limit

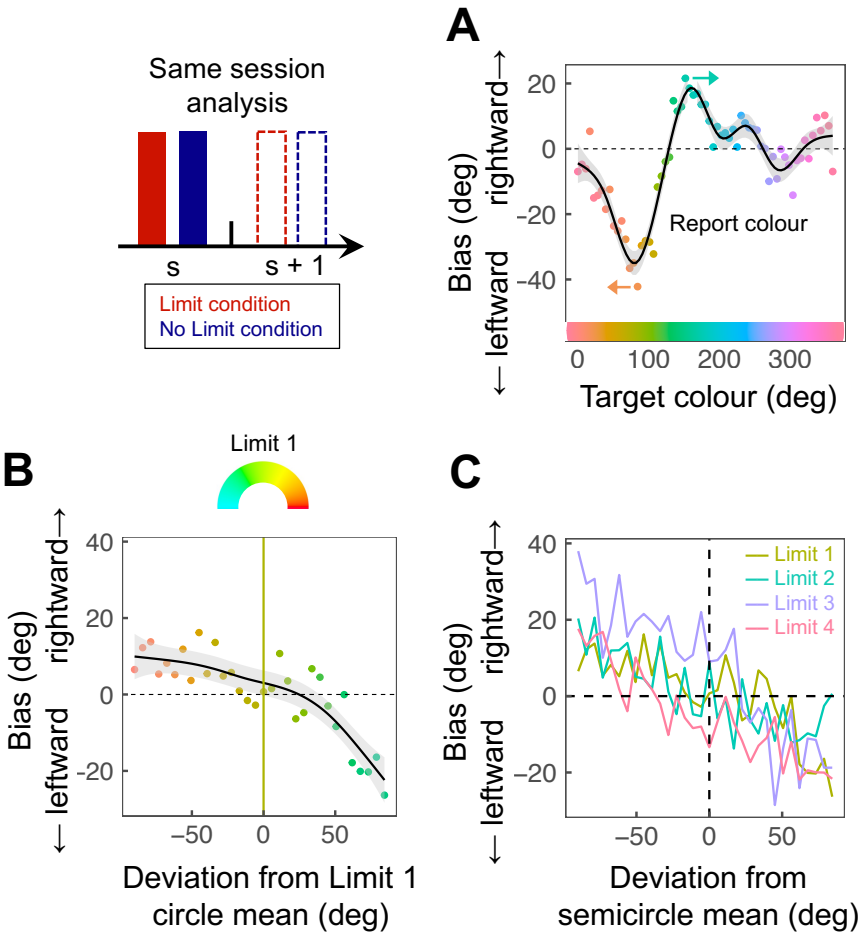


Results for colour error in the same session, comparing Limit blocks to No Limit blocks. **(A, B)** Colour error was lower in the Limited condition compared to the No Limit condition. **(C)** This benefit for the Limited condition arose shortly after the switch to the Limit block and was sustained over the course of the experiment. **(D)** Specifically, it arose in around 10–20 trials after the switch. Error bars are within-subject error bars, and points represent trial averages.

no others reached significance (all $p > 0.05$). This indicates that performance improved within 10–20 trials of the Limit block.

In addition, results showed typical set size effects, with lower error for set size one than two items (Fig. S6B). Crucially, the block effect seemed to only occur in the set size two conditions (Interaction effect of Block Type $\times$ set size $\beta = 5.45, p < 0.001$). While both the memory delays (after stimulus and after cue presentation) impacted error with increasing delay, this did not differ between Block Type conditions (all interaction effects $p > 0.05$, S6C & D).

Figure 6.3
Colour-specific and restriction-induced biases in the same session

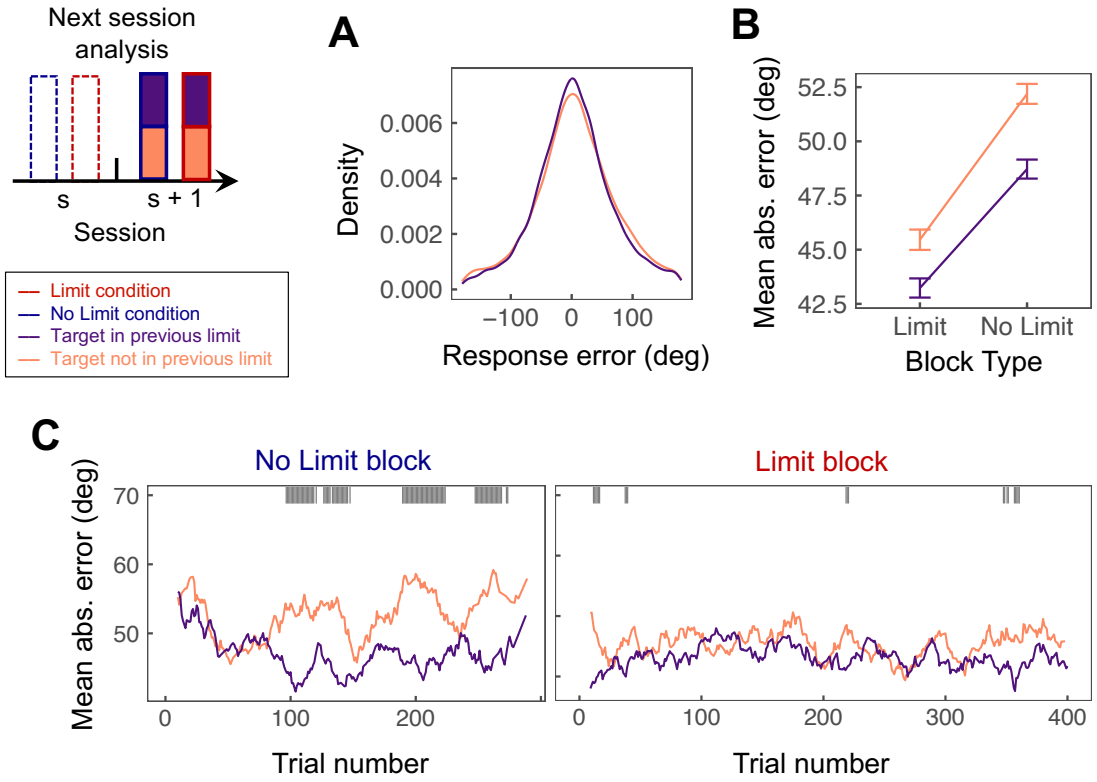


There were colour-specific biases in the Limit block. **(A)** Yellow colours were shifted leftwards to be represented more reddish (orange dot and arrow), and green colours were shifted to be more blueish (blue dot and arrow). Bias values were corrected by the expected bias for the colour computed across all No Limit blocks. Each dot represents the bias for one target colour. Colours represent reported colour, and the colour bar at the bottom represents target colours. **(B)** In addition, there were biases towards the centre of the semicircle mean of the Limit block. Limit 1 is visualised and shows that the further the colours were from the mean, the larger this bias. The vertical line represents the mean colour on the semicircle. Colours represent reported colour. **(C)** These biases were present in all Limit conditions. Colours represent the colour of the respective semicircle mean.

Results showed a behavioural bias for specific colours, specifically blue and red. For example, as shown in Figure 6.3A, veridically orange points were reported as more reddish, whereas greenish points were reported as more blueish. Further, there was

a colour bias towards the Limit mean, as illustrated for Limit 1 (Fig. 6.3B): colours leftward of the mean experienced a rightward bias and vice versa. A LLM showed that this bias was present in all different Limit conditions ($\beta = -0.17, p < 0.001$, Fig. 6.3C). It was particularly strong in Limit 3, which differed from Limit 1 (interaction effects respectively $\beta = -0.13, p < 0.001$). This suggests that the monkeys learned the Limits of the colour circle and adapted their responses to reflect this.

Figure 6.4
Benefit for colours contained within the previous semicircle Limits



(A, B) Colour error was lower for targets within the Previous Limit in both Limit and No Limit blocks. **(C)** This benefit appeared after around 100 trials into the No Limit block of the next session, and was sustained throughout this block (left panel). The plot shows the rolling mean for 10 trials, and the grey bars reflect a significant difference between conditions at a given time point. When the Limit block of the current session introduced a new Limit, the benefit for the old Limit shrank, however, did not disappear entirely (right panel).

Effects of the restricted colour circle on the following session Next, we assessed to what extent performance benefit and biases found in the same session would persist on to the next session. In other words, would memory traces enable monkeys to sustain the improved performance for stimulus colours falling within the previous item Limit compared to "new" colours falling outside of it?

Overall, colour error was lower for the Limit block ($\beta = 6.63, p < 0.001$). Monkeys also generally performed better when probed on items within the Previous Limit compared to items outside of it ($\beta = -2.18, p < 0.001$, Fig. 6.4A). There was no interaction effect ($\beta = -1.34, p = 0.14$), indicating that there was no difference between Block Types, and the benefit for items within the Previous Limit was therefore maintained throughout both No Limit and Limit blocks of the next session.

The difference between Previous Limit conditions was compared over time using rolling means with a window of 10 trials. It was apparent that there was a benefit for the old item colours after around 100 trials into the No Limit block of the new session. A linear regression indicated that this benefit for targets in the Previous Limit arose over time ($\beta = -0.03, p < 0.01$), but the main effect of Previous Limit stayed nonsignificant ($\beta = 0.01, p = 0.22$). Rolling t-tests indicated fewer significant differences for Limit blocks, however, an LMM showed that a benefit of about 3 degrees of absolute error was sustained ($\beta = -2.96, p < 0.05$), and did not change over time ($\beta = -0.01, p = 0.82$).

While both monkeys showed better performance for items in the Previous Limit block in the No Limit condition (both $p < 0.05$), this diverged in the Limit block, with

Monkey 1 not showing a difference ($\beta = -0.17, p = 0.85$), whereas Monkey 2 did (IE Monkey $\times$ Target in prev. Limit: $\beta = -3.66, p < 0.01$). This could possibly be due to a ceiling effect of Monkey 1 in the Limit condition (Fig. S8B).

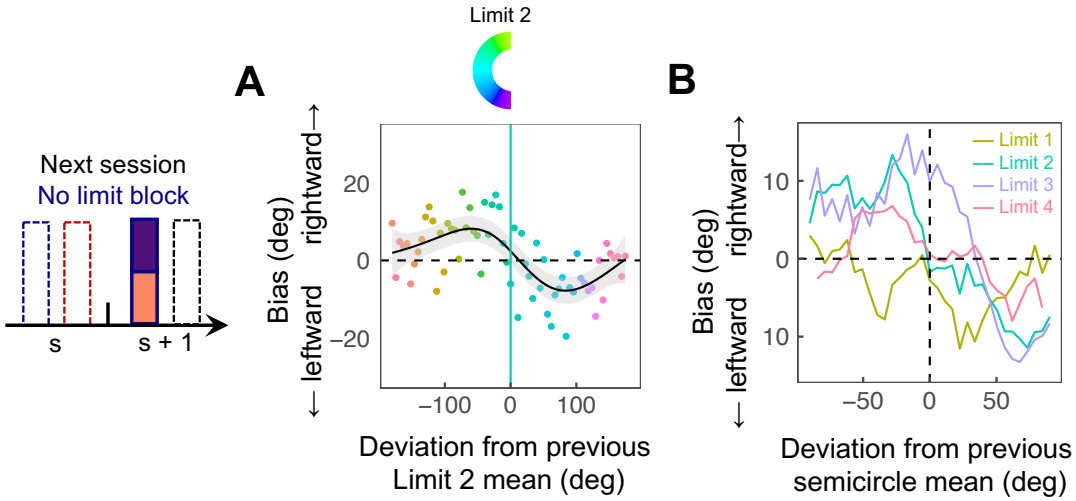
Performance was again better for the lower set size compared to the higher set size (Fig. S8B). Delay impacted performance non-linearly, but did not change the effect of the Previous Limit qualitatively (Fig. S8C&D). This might be explained due to the Previous Limit sometimes overlapping with the new Limit. For example, when Limit 2 followed Limit 3, colours in the lower-left quadrant (180–270° of the colour wheel) would occur in Limit blocks of the previous and current session.

Taken together, these findings indicate that colours from the Previous Limit condition exhibit a performance benefit beyond the duration of the Limit condition itself.

Next, we investigated whether the biases towards the mean of the Previous Limit circle were sustained in the new session. Analyses showed that this was dependent on the specific colour Limit. For example, Limit 2 responses showed the typical bias towards its mean (illustrated in Fig. 6.5A, $\beta = -0.03p < 0.05$, other biases are shown in Fig. S9). Similarly, Limits 3 and 4 showed the bias (Limit 3: $\beta = -0.03, p = 0.05$; Limit 4: $\beta = -0.03, p < 0.05$), whereas Limit 1 did not ($\beta = -0.009, p = 0.34$, 6.5B).

Overall, these results provide strong evidence for a sustained signature of performance benefits for stimuli within the previous semicircle and some evidence for carry-over of bias. However, the latter seems strongest in Limit conditions where the Limit

Figure 6.5
Bias for colours from the previous limit



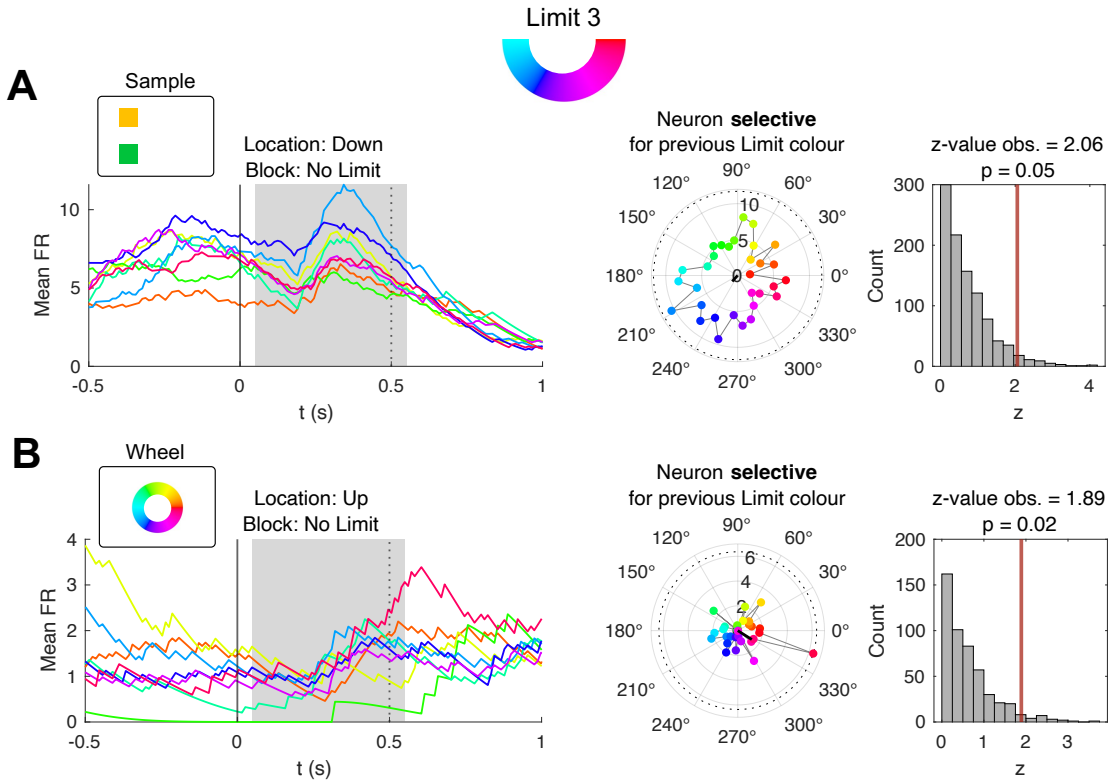
(A) In some Limits there were biases towards the Previous Limit mean. Responses during the Limit 2 condition displayed the typical pattern with rightward biases for values lower than the mean and leftward biases for values above the mean. Colours represent reported colours. Each point represents one colour increment. **(B)** Specifically, Limits 2 and 4 showed the biases, but not Limit 1 and Limit 3. Line colours are the respective colour for the previous semicircle mean.

mean coincided with the a priori preference of the monkeys: red and blue (Limits 2 & 4).

6.4.2 Neural data

Single-neuron selectivity analyses The single-neuron analysis approach is shown for two example neurons in Figure 6.6A&B. Both neurons were PFC neurons and were recorded in a session where the Previous Limit included colours from the Limit 3 colour circle. For visualisation, the average neuron response was averaged for eight colour bins around the respective event onset (Sample for 6.6A, Wheel for 6.6B), in the respective left panels of the figure. Neuron selectivities were calculated for the grey-shaded timeframes, but for 32 colour bins.

Figure 6.6
Selectivity analysis example neurons

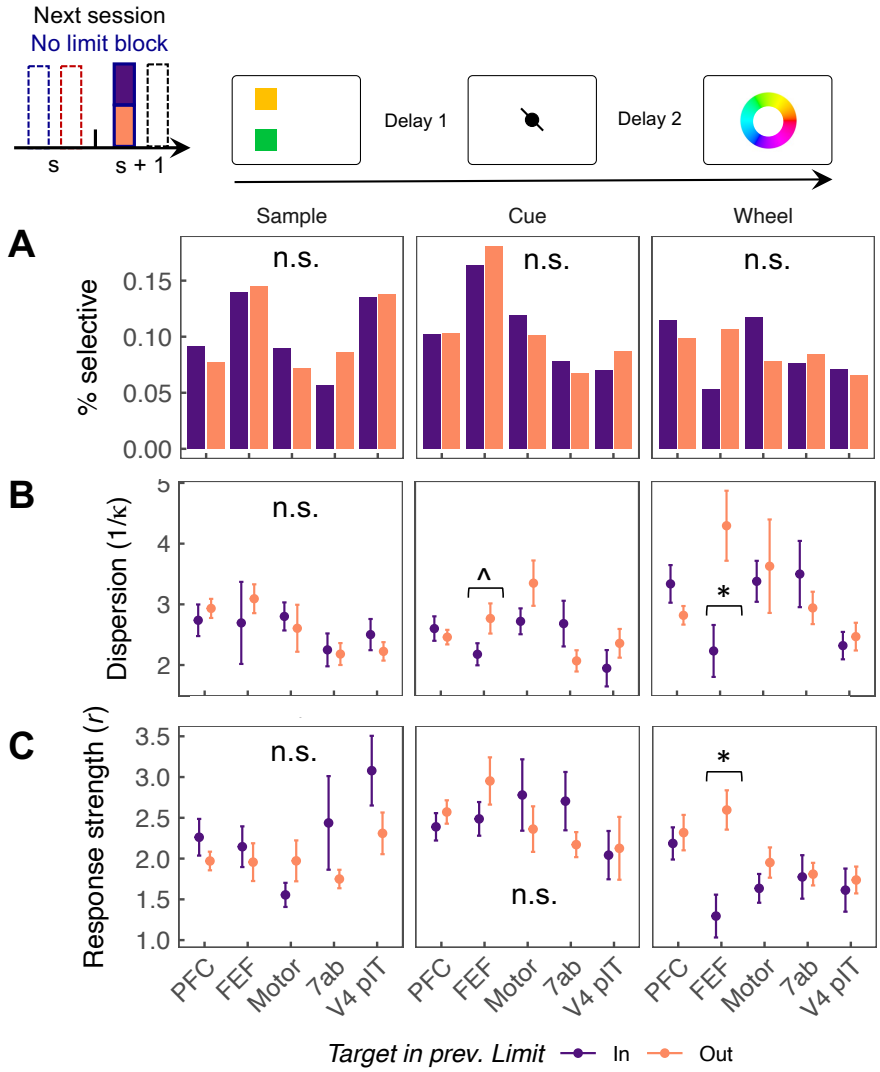


Firing rates **(left-hand side)** and colour selectivity **(right-hand side)** for example neurons in PFC that were selective for a colour included in the previous day's Limit colours (Limit 3) following the presentation of the stimulus samples **(A)** and the response wheel **(B)**. The right panels show the observed z-values and the null distributions they were compared against to determine whether the neuron had a preferred colour. Grey-shaded area indicates the time window over which selectivity was computed.

We found colour-selective neurons in areas PFC, FEF, Motor, 7ab, V4pIT, with neuron counts N = 500 (9.6% of recorded neurons), N = 177 (13.7%), N = 144 (9.5%), N = 307 (7.6%), and N = 226 (9.6%), respectively.

Next, the proportions of neurons selective for colours within the Previous Limit versus outside of the Previous Limit were compared for trials in the No Limit block of the session following the Limit (Fig. 6.7A). This was done separately for each ROI and stimulus event. There were no significant differences between proportions of selective

Figure 6.7
Results of single-neuron analyses



Single-neuron responses for colours within versus outside of the Previous Limit in different brain areas and time windows across the task. **(A)** The proportion of neurons selective colours in each group. **(B)** Mean dispersion of the neurons tuning curve, corresponding to representational precision. **(C)** Mean strength of the response. Error bars indicate S.E. $^{\wedge}p < 0.1, ^{*}p < 0.05$.

neurons in any ROI for any event (all $p > 0.5$). This suggests that the number of selective neurons did not increase as a result of the colour restriction, and cannot account for performance benefits seen for Previous Limit items.

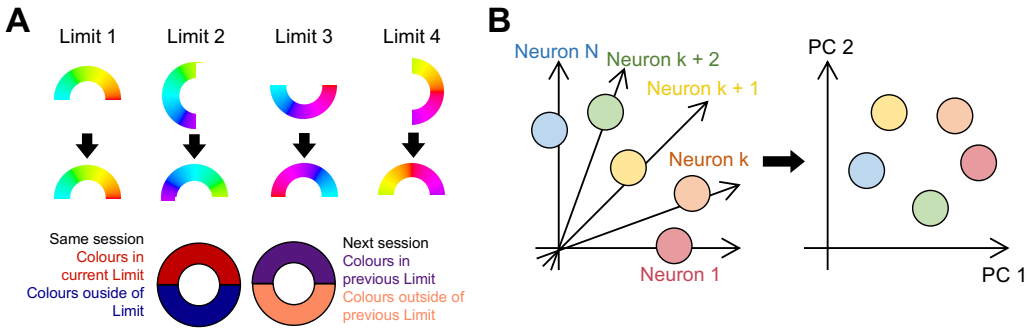
We therefore assessed whether, if not the count, the response precision of single neurons' responses might have changed. First, we assessed the spread around the resultant vector given by the dispersion parameter $1/\kappa$ from the von Mises distribution (Fig. 6.7B). There were no significant differences for the Sample period (all $p > 0.05$). For the period following the Cue, we found a trend towards decreased dispersion for the Previous Limit items in FEF ($t(69.45) = 1.915, p = 0.06$). Similarly, only the effect for FEF became significant in the time window after the response wheel ($t(28.679) = 2.872, p < 0.01$). This indicates that the neural response was more precise in FEF for colours in the Previous Limit condition that colours outside of it.

Finally, we investigated whether the average response strength, i.e. the resultant vector r was different between Previous Limit groups (Fig. 6.7C), and found a significant effect for FEF in the period after the Wheel ($t(17.826) = 3.65, p = 0.002$). However, this was in the opposite direction of what we expected, with Previous Limit items showing a weaker response.

To summarise, a more precise representation was accompanied by a simultaneously weaker representation of the Previous Limit colours in FEF for stages in the trial past the stimulus selection. These results speak for the importance of FEF in representing the Previous Limit context on a single-neuron level, possibly contributing to behavioural effects.

Exploratory population analyses To investigate whether there was structure in the population response we conducted a PCA on mean firing rates in rotated colour space (Fig. 6.8).

Figure 6.8
Schematic of exploratory population analysis



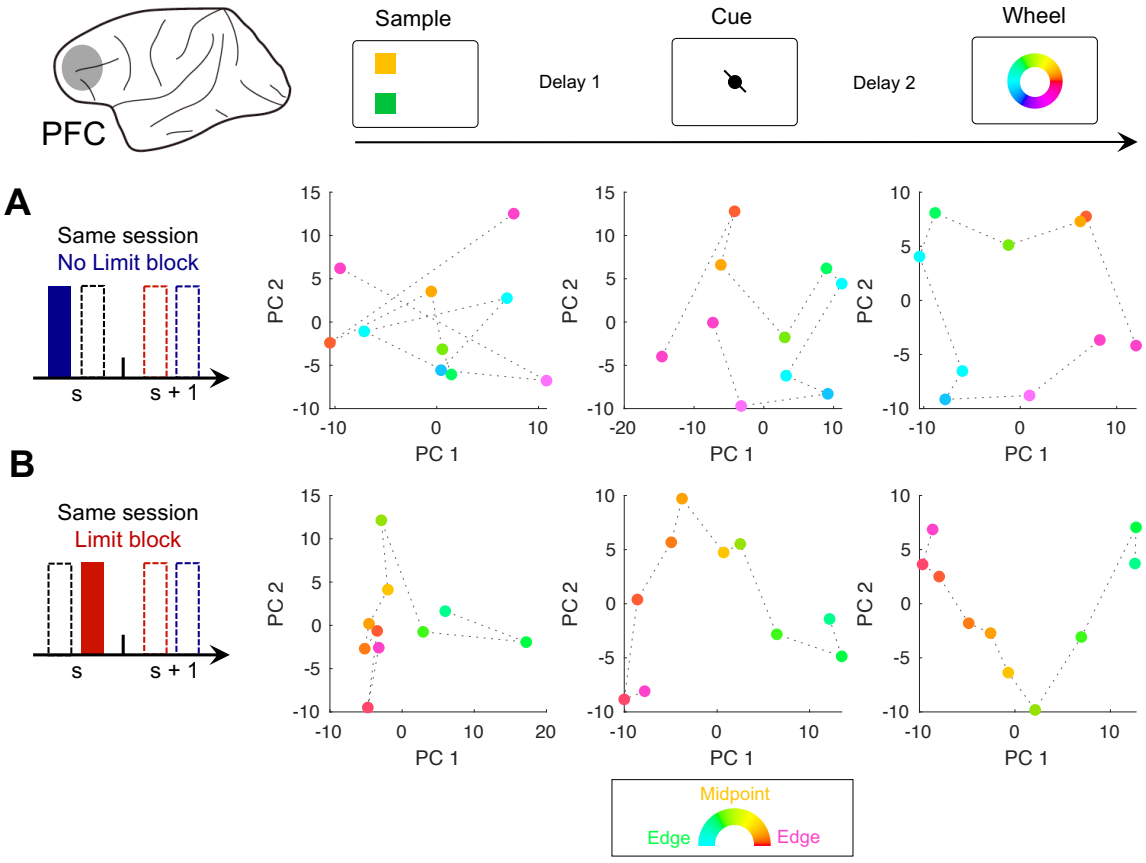
(A) We rotated the colour space to obtain a new, task-based space in which the upper half of colours represented items in the restricted colour space, according to either the current or previous Limit. The upper half thus comprised colours within the current or Previous Limit on the next session analysis, whereas the lower half included colours outside the current or previous Limit. **(B)** PCA was used to reduce the neural population response to two dimensions and project it into the new PC-space.

We report results from PFC and FEF as two exemplary ROIs due to them showing the most pronounced effects in the single-neuron analyses (Figs. 6.9–6.12). The remaining results are included in the Appendix (7.3, Figs. S10–S15).

As a first step, we replicated existing results (Uchimura et al., Unpublished manuscript). They showed that the population response in the No Limit condition, when analysed in 2-dimensional PC-space, was best described as a circle with similar colours adjacent. After the switch to a Limit block, this changed to resemble an open triangle.

From the period after the cue onwards, we observed the same pattern in PFC: A circle-like shape in the No Limit block (Fig. 6.9A, middle and right panels) and a hat-like shape in the Limit block of the same session (Fig. 6.9B, middle and right panels). The triangle was characterised by the centre of the Limit semicircle being represented by the middle corner of the triangle, and Limit edges by the ends of the

Figure 6.9
Same session neural geometry in the PFC

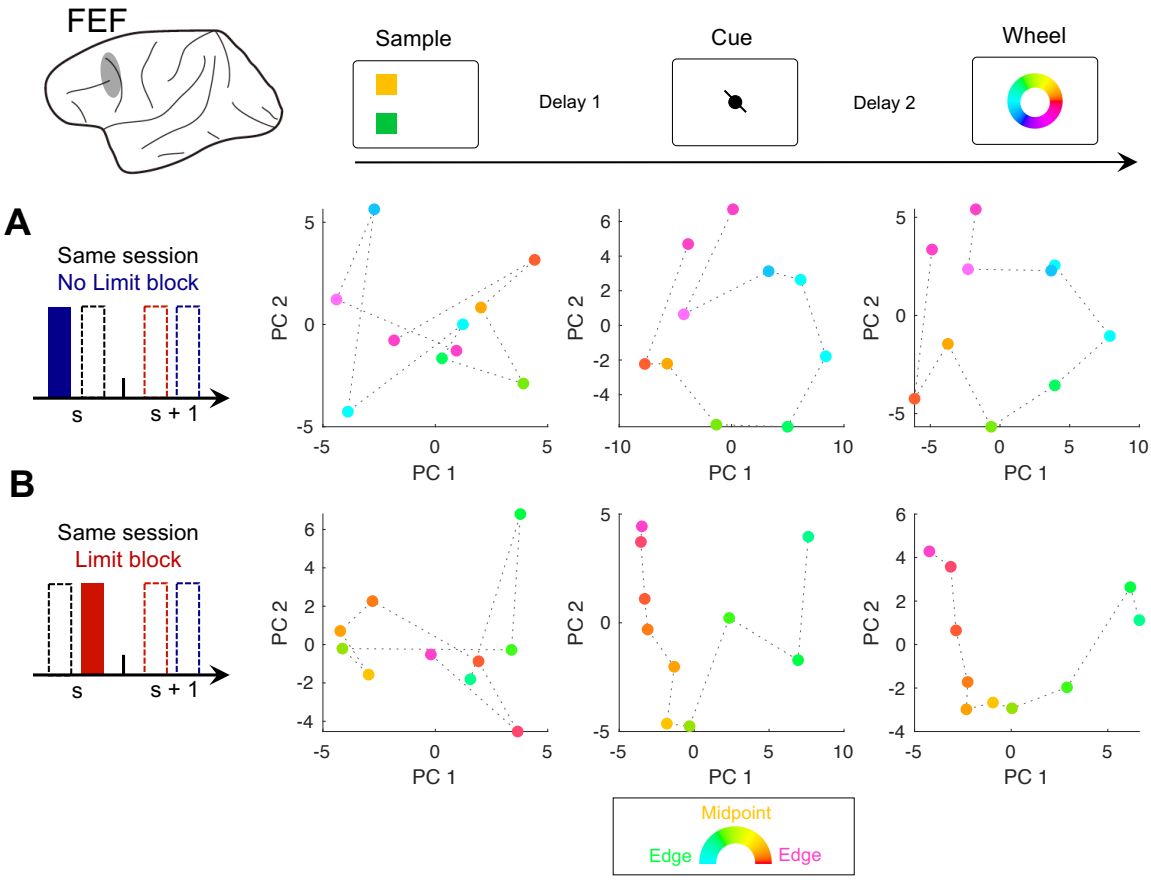


Population analysis results for the effect of stimulus space restriction on the same session in PFC. In Limit blocks **(A)**, a circular pattern emerged after the cue, whereas in No Limit blocks **(B)**, this has shifted to a triangular pattern. Colours are just for illustration, as different colours made up the limit and no limit colours in different sessions.

triangle. The distinction was more pronounced in the period following the Wheel than the one following the Cue. We found the same pattern of results in FEF, with the condition-distinct shapes emerging after the Cue, and growing stronger in the period following the wheel (Fig. 6.10A&B, middle and right panels).

The goal of the next analysis was to find out if the observed neural geometry effects were sustained into the next session (the next day). Specifically, we were looking

Figure 6.10
Same session neural geometry in FEF

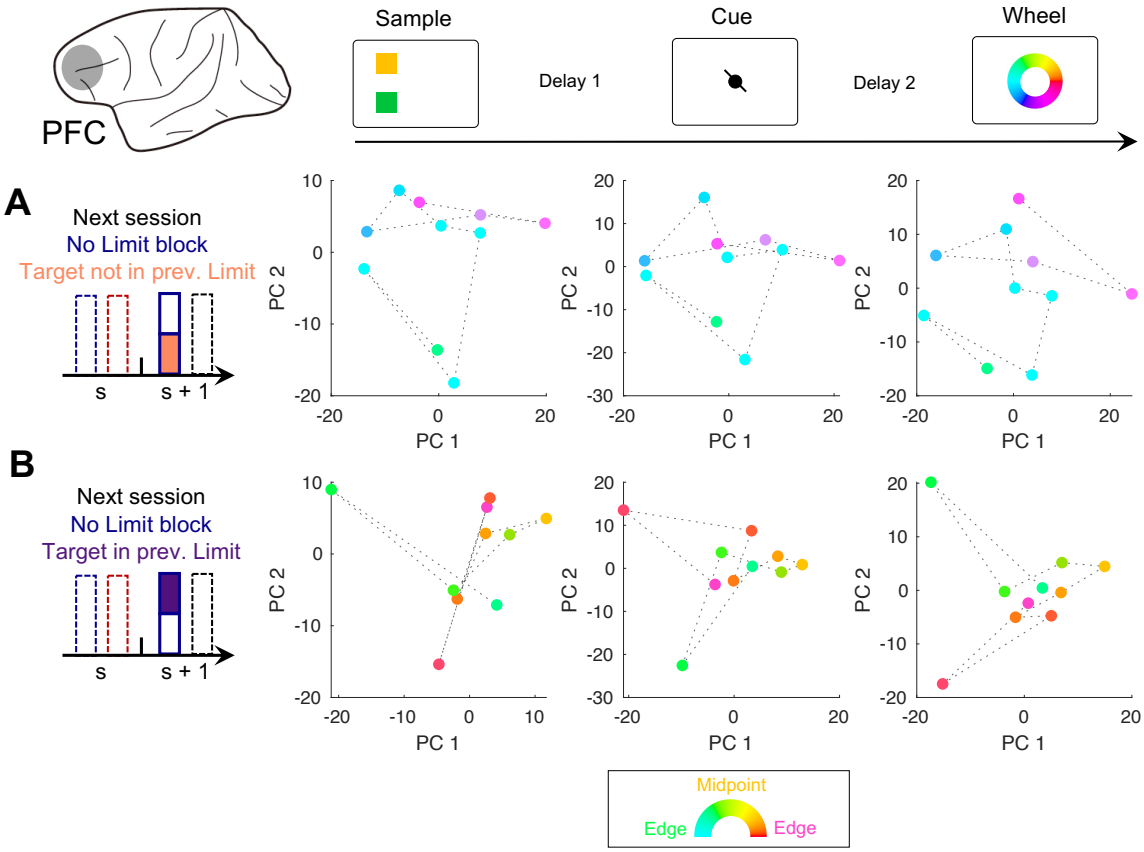


Population analysis results for the effect of stimulus space restriction on the same session in FEF. In Limit blocks **(A)**, a circular pattern emerged after the cue, whereas in No Limit blocks **(B)**, this has shifted to a triangular pattern. Colours are just for illustration, as different colours made up the limit and no limit colours in different sessions.

for a triangle geometry underlying items within the previous Limit. This would indicate retention of the Limit context beyond serial biases at a population level.

In the PFC, such a pattern was largely absent (Fig. 6.11). Responses for items outside the Previous Limit vaguely resembled a circle, however, similar colours were intermixed rather than adjacent. Responses for Previous Limit items tended to fall onto a line or a narrow "V" rather than a triangle.

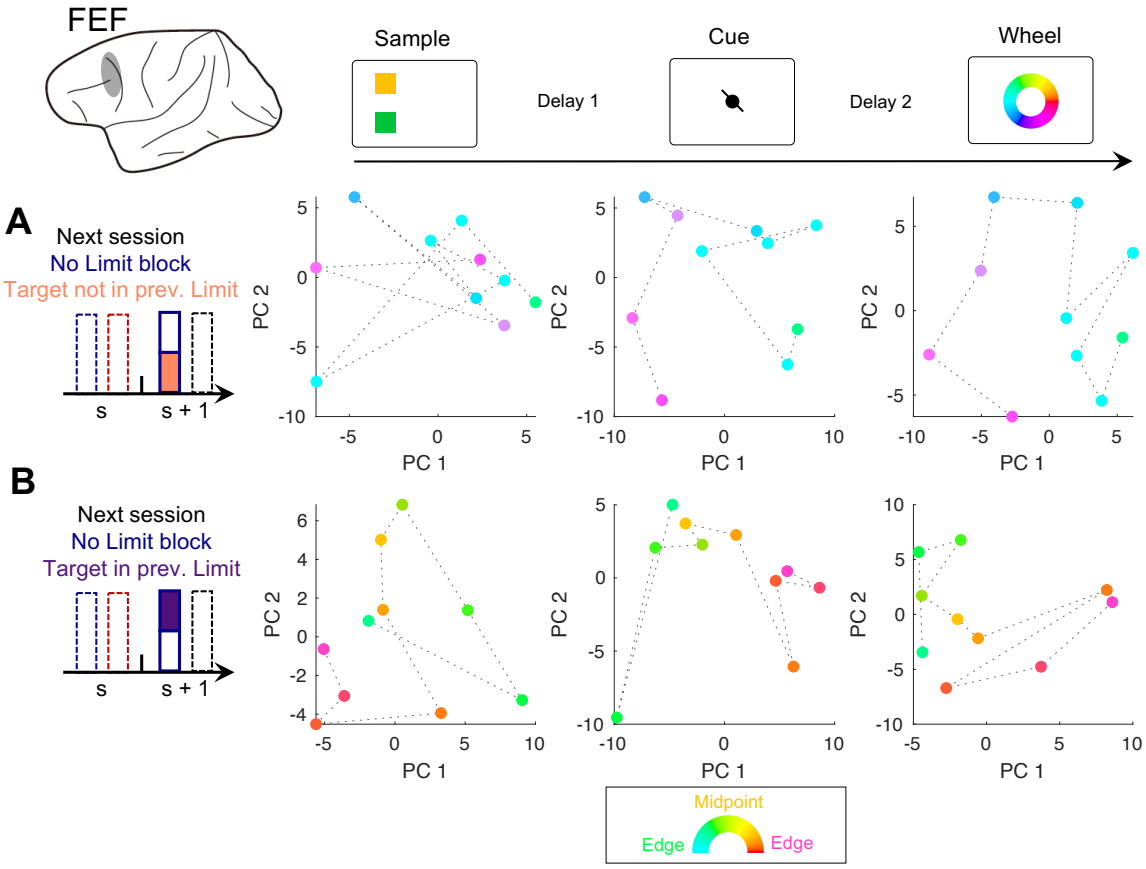
Figure 6.11
Next session neural geometry in PFC



Population analysis results for the effect of stimulus space restriction on the No Limit block in the next session in PFC. In trials where the target was outside of the Previous Limit **(A)**, a circular pattern emerged after the cue, akin to the circle in the same session results. In trials where it was within the Previous Limit **(B)**, no clear pattern was visible. Colours are just for illustration, as different colours made up the limit and no limit colours in different sessions.

In FEF, however, we found a pattern resembling the one we hypothesised. Starting in the period after the cue, a triangle is visible for items inside the Previous Limit (Fig. 6.12B, middle panel), however pointing in the opposite direction of the triangle in the same session. Unexpectedly, the pattern associated with the items outside of the Previous Limit resembled a circle (Fig. 6.12A, middle panel), with edges of the semicircle joining, despite the activity representing only half the colour wheel. FEF re-

Figure 6.12
Next session neural geometry in FEF



Population analysis results for the effect of stimulus space restriction on the No Limit block in the next session in FEF. In trials where the target was outside of the Previous Limit **(A)**, a circular pattern emerged after the cue, akin to the circle in the same session results. In trials where it was within the Previous Limit **(B)**, it resembled a triangular formation as in the same session results. Colours are just for illustration, as different colours made up the limit and no limit colours in different sessions.

sponses after the Cue strongly resembled the Previous Limit condition's response, with a downward pointing triangle with similar colours adjacent to one another. Despite the original colour space for items outside the Previous Limit being a semicircle just like the items within, the neuronal response resembled a circle.

Taken together, we replicated the result by Uchimura et al. (Unpublished manuscript), and in addition showed that the neural geometry underpinning induced colour biases carried over to the next testing day.

6.5 Discussion

In this study we reanalysed an NHP dataset collected from monkeys performing a delayed colour estimation task where the colour space was restricted to induce a centrality bias. It had previously been reported on the same dataset that this led to performance increases and indeed produced a bias towards the centre of the colour distribution. This was underpinned by the geometry of population neural responses morphing from a circle in unrestricted colour space to a triangle in restricted colour space. Here, we extended these results by analysing the effect that the colour limit had on stimuli within it in the next session. Behavioural results showed that performance benefits as well as biases to the centre of the colour distribution persisted into the first (No Limit) block of the next session, and, to a weaker extent, to the second (Limit) block. In our single-cell analyses, we found that while overall numbers of selective neurons did not differ between items within and outside the previous Limit, neurons in PFC and FEF tended to represent the Previous Limit colour at higher fidelity during periods following Cue and Wheel, respectively. Paradoxically, this was accompanied by lower encoding strength in FEF neurons after the response wheel was presented. Further, exploratory population analyses demonstrated that geometry induced in the

previous session was clearly maintained in FEF, again after response wheel presentation and hints of this being the case after cue presentation.

Taken together, we interpret this result to indicate that monkeys were able to create long-lasting adaptations to aid WM of colour features. Crucially, while the effect extended to the next session, relying on long-term retention, it was also flexible: upon the introduction of a new limit, monkeys changed their responses to it within 10 trials. This suggests WM supports adaptive both in being susceptible to biases that lead to relative performance benefits (Panichello et al., 2019) and being flexible enough to enable shifts in behaviour when faced with new environmental statistics.

In the same session, behavioural results showed a response benefit in the Limit blocks compared to the No Limit block. This benefit arose in the 10 to 20 trials after the block switch. Since the full colour wheel was displayed at response, however, the task was not inherently easier unless there was memory of the stimulus history. One reason for this result might thus be that the monkeys learned that the task context had changed through learning that the stimuli were coming from one side of the semicircle. This is plausible insofar as sampling 10 trials at random from a full colour circle would be very unlikely to lead to the same outcome ($p = 0.5^{10}$), and enough experience to learn that the stimulus generation had changed. This would suggest that the monkeys acquired abstract knowledge about the task. Another possible explanation that the colours became easier to chunk: due to the stimuli in the Limit block being more similar to one another, they may have experienced an attraction bias, which would bring responses closer to the semicircle mean. This may be supported by our results

since the performance benefit only appeared for trials with two samples. Alternatively, this could be a ceiling effect, as performance was generally much better with only one sample. More fine-grained analyses could tease apart centrality and attraction biases.

We also found that the performance benefits and biases carried over to the next session, despite this being on the next day. They appeared after about 90 trials into the No Limit block of the next session. It is unclear why the benefit arose this late into the block. If the monkey had an abstract representation of the task, it might notice that colours are drawn from a full circle again sooner and readjust priorisation to be equal for all colours. The results seem more in line with a slow re-learning or remembering of the old colours. Like the initial learning process, this could rely on attraction biases and a higher propensity to associate colours from the old stimulus space. Since in a No Limit block trials with both stimuli drawn from the Previous Limit this would explain the slower appearance of the benefit. However, more fine-grained analyses are needed to confirm this.

Our findings point to a special role of PFC and FEF. This was apparent throughout single-neuron and exploratory population analyses. These results are in line with prior research showing that activity in frontal areas, in particularly the PFC, adapts to task demands and shows plasticity in learning tasks in humans and NHPs (Miller & Constantinidis, 2024; Riley et al., 2018). While same-session analyses showed effects in PFC and FEF, single-neuron analyses for the benefit of Previous Limit colours in the next session only found FEF to be implicated. Dispersion was lower for old colours, which may suggest that this narrower selectivity distribution underlies the benefit we

saw in behaviour. Concurrently, responses of selective neurons were weaker for old colours, perhaps suggesting that old and new colours are distinguished on the single-neuron level. At the population level, distinct effects of the Previous Limit were much less pronounced and need further evaluation. Nevertheless, this suggests a role for FEF in context-dependent WM.

One interpretation of this is that it might represent the Previous Limit via supporting memory-guided saccade tasks through persistent firing (Balan & Ferrera, 2003; Clark et al., 2012; Funahashi et al., 1989; Hart & Huk, 2020). We mainly observed its responses to be reflecting the previous item limit after the retro-cue that indicated which item to report. In the No Limit block, it is plausible that the monkey would already prepare to report a colour from the limit block. Moreover, FEF showing increased activity in trial stages after the retro-cue is consistent with literature in that the centrality bias has been shown to occur during retrieval and judgment processes as opposed to encoding (Ashourian & Loewenstein, 2011; Jou et al., 2004).

A limitation of the study is our approach to computing colour selectivity. Neurons may be bimodally tuned to different task properties (Rigotti et al., 2013), and therefore could be selective for two colours at the same time. However, the uniformity tests used to ascertain selectivity here assume unimodality and return an inaccurate fit with overly strong dispersion when this is not met. Moreover, the same analysis could be done in a more fine-grained way, where metrics are assessed at every time point in a trial. Thus, our results on the relative proportion of selective neurons retaining items within or outside the Previous Limit should not be overinterpreted. Extensions to the

analysis should include different measures such as entropy-based ones that do not make assumptions on modality (Panichello & Buschman, 2021). Another limitation is the exploratory nature of the population analyses, preventing firm conclusions. Future work should supplement this with metrics that lend themselves to statistical comparison.

Nevertheless, our results suggest that biases were carried over across sessions behaviourally and at the single-neuron level. We also found some evidence that the effect might be detectable at the population level. We therefore propose that biases in WM enable adaptive behaviour in changing stimulus spaces across days, and potentially conducive to long-term learning of task structure.

Apart from refining the analysis, several questions remain for future research to answer. For example, it is unclear how neural activity develops after a block switch: do the the same neurons that represented the old colours switch to representing the newly introduced limit colours? How are colours represented differently when they are included in overlapping (the old and new) limit? Moreover, this task is one example of introducing a context and may depend on attraction biases between stimuli. A related question is whether these findings would generalise to other kinds of biases.

7 Discussion

7.1 Summary of Results

Using different experimental paradigms and methods, all chapters in this thesis pursued the common goal of extending our understanding of the relationship between working memory (WM) and long-term memory (LTM). The first line of research focused on the role of WM contributions to learning and LTM formation, including statistical learning, reward learning, and the interaction with episodic memory formation. The second aim was the role of prior knowledge stored in LTM and how WM can draw upon these existing representations.

Chapter 2 attempted a replication of the WM compression effect (Brady et al., 2009) in a continuous delayed estimation paradigm. To investigate whether feature binding contributes to the capacity savings in the compression effect, stimuli were conjunctions of arrow shape and colour. We hypothesised that bindings between them would, through repetition, be maintained as an integrated object and improve performance. We also tested whether participants could reproduce the learned associations after the experiment. Across three experiments, participants either maintained repeated arrows or random stimuli in memory.

Results showed a tentative benefit for the learned item, however, there was no performance benefit for the non-learned items when shown with learned items. Since a true compression effect is thought to recode WM to free up space for all items, this suggests that there was a different mechanism at play. In addition, most participants

failed to notice the repeating item, even under explicit instruction. Our results are consistent with the literature only in part. We encountered difficulties creating robust memory traces. This is consistent with other studies which have since shown that long-term traces can only be acquired under specific circumstances, such as reporting every item on every trial (Musfeld, Souza, & Oberauer, 2024), and findings showing restrictions of LTM encoding at high set sizes (Forsberg, Guitard, & Cowan, 2021). Our observation of a learning benefit with participants lacking explicit awareness of the learned stimulus is inconsistent with literature showing that explicit knowledge is needed to create a compression effect (Ngiam, Khaw, et al., 2019), unless a simple benefit is different from the compression effect and can operate implicitly. While our initial hypothesis could not be addressed, this strengthens the notion of WM as a system that is specialised for flexible encodings and re-encodings rather than the accumulation of long-term traces for learning.

Chapter 3 introduced the Any-back task, a novel experimental task devised to help further delineate signatures of LTM and WM in tasks where they operate jointly. Participants switched between different trial types, encoding and retrieving a continuous sequence of unique images in different colour hues, while protecting these memories from interference through distractor items. Performance was measured by assessing object and colour performance.

A major finding of this chapter is a striking decrease in colour performance after only one interfering item, while further increasing interference thereafter left performance intact. Performance after the decrement was consistent with a strategy of using

colour categories rather than fine-grained colour information to generate responses. We suspect that this might indicate a "hand-off" of items from WM into LTM. Moreover, WM was sustained above chance-level even throughout longer retention intervals of up to (90s). Object recognition decreased more gradually, and was sustained at high performance throughout the task. We further found interference effects and the number of items held in memory to be differentiating WM and LTM. This study contributes to a line of literature investigating the benefits of real-world objects compared to abstract objects in WM (Brady & Störmer, 2020b; Brady et al., 2016). Specifically, the paradigm can further be used to study how continuous feature information is bound to real-world objects (Chung et al., 2023), and how it develops under conditions of interference.

Chapter 4 showed that holding WM of the stimulus (state) component of RL can result in better learning, even when this maintenance is not necessary for the task. Across four experiments, participants performed a combined probabilistic RL and WM task. Two experiments designed to enhance RL through WM maintenance produced the expected effect. However, this improvement was not driven by a win-stay lose-switch strategy, contrary to what would be expected if participants were simply using WM to solve the task. Unexpectedly, we observed no significant effect of WM suppression on RL performance, neither with probabilistic nor deterministic reward. Additionally, when including suppression blocks, the enhancement of WM effect from previous experiments was not replicated. Overall, the findings provide limited evidence for a WM contribution to RL performance, suggesting that this only occurs under specific con-

ditions. Nevertheless, this study sheds doubt on accounts that frame WM and RL as competitive systems (Collins, 2018; Collins, Ciullo, et al., 2017a) by demonstrating a case in which they collaborate.

Chapter 5 presented a neural network model integrating components for WM and LTM, implemented through a recurrent Buffer and a Hopfield model, respectively. The theoretical question of how neurons should be distributed between these two memory systems in light of varying memory systems was explored. The memory stores used separately demonstrated the expected hallmarks of WM and LTM. In the Hopfield model, memories were stored reliably, but its performance deteriorated once its storage capacity was exceeded, as expected (Hopfield, 1982). In contrast, memories in the Buffer were vulnerable to interference from new inputs and noise. Using them in conjunction and dividing up short and long memories between them, we confirmed that there was a trade-off between the two. The longer a memory was needed, the more beneficial it was to store it in the Hopfield model, due to the Buffer's susceptibility to memory degradation with time. Finally, we compared the combined model's performance between environments, finding that the combined model was beneficial compared to single stores in all environments. Overall, the model demonstrates a normative account of specialisation in memory systems.

In **Chapter 6** we reanalysed an NHP dataset from monkeys performing a delayed colour estimation task with stimuli drawn from restricted versus unrestricted colour space. This induced a centrality bias (Hollingworth, 1910). Previous analyses of the same dataset showed improved performance and a bias toward the colour dis-

tribution mean, with a shift in neural geometry from a circle to a triangle underlying this benefit. Our analysis extended this to show that performance benefits and centrality bias persisted into the next session. While single-cell analyses showed no overall change in selective neuron percentages, they revealed a higher precision representation of colours from the previous restriction in FEF neurons in the later trial stages. Exploratory population analyses suggested that the geometry induced in the prior session may in part persist in FEF. These results are in line with plasticity in frontal areas previously observed in both humans and NHPs (Miller & Constantinidis, 2024; Riley et al., 2018), but demonstrate that biases in WM persist over longer time periods than previously thought.

7.2 Implications and Future Directions

The empirical findings presented across the five chapters directly contribute to the aims of this thesis. The results add to our understanding of how WM maintenance influences LTM formation, its role in statistical and reward learning, and how prior knowledge shapes the content and precision of WM representations. Together, they give rise to broader implications about the interplay between WM and LTM and prompt questions that point to promising future research directions.

WM as a capacity for learning and formation of LTM Across several empirical chapters, WM emerges as a central capacity involved in learning and the formation of LTM. The findings suggest that WM is not merely a temporary buffer but may actively shape what is stored in LTM. According to our findings, WM seems

to support the gradual accumulation of statistical regularities over time under specific circumstances, and a "hand-off" of specific representations into LTM. These distinct patterns point to multiple mechanisms through which WM can contribute to long-term retention.

The findings suggest that WM serves as a tool for learning only under specific conditions. In **Chapter 2**, we explored statistical learning within a WM paradigm but found it challenging to establish a consistent learning effect. The repeated pattern was relatively weak—limited to a single repeating item—and overall, the evidence for robust learning was limited. This aligns with prior work suggesting that statistical learning, particularly of the Hebbian type (Hebb, 1961), is difficult to elicit using traditional WM tasks (Musfeld, Souza, & Oberauer, 2024).

By contrast, in **Chapter 6**, we examined learning through the formation of perceptual biases and observed a much more consistent effect. Here, the task involved a stronger structure and lower set sizes. While no correlation between stimulus features was present, the context (block) monkeys found themselves in was predictive of the range of colours stimuli would take on. This complete shift in the stimulus distribution led to consistent improved performance and centrality bias. Importantly, these effects persisted beyond the WM task into the next day.

In **Chapter 4**, we also found that WM can support reward-based learning. Unlike the more gradual accumulation seen in statistical learning, reward learning may involve WM when it is beneficial. This raises the possibility that WM is not uniformly engaged across all learning contexts, but rather acts as a tool that is recruited strate-

gically when it enhances performance. When WM is unreliable or overloaded, other learning systems may take precedence, as previously suggested (Yoo & Collins, 2022). This suggests an interesting parallel to the statistical learning case, in which learning only occurs if the structure is noticeable enough.

In contrast to Chapters 2, 6, and 4, in which learning happened gradually, in **Chapter 3**, we observed what appears to be a "hand-off" from WM to, possibly episodic, LTM. We found a drastic shift in colour memory precision to suggest that the representation may no longer be held in WM, but has been relayed into LTM. If this is the case, it suggests a rapid, one-shot transition, in which information maintained in WM is transferred into a more durable memory trace. This contrasts with the more gradual accumulation seen in statistical learning. Notably, this memory remained stable over a prolonged delay, consistent with long-term retention.

Our findings from simulations **Chapter 5** further suggest that such a shift might serve a protective function: moving information out of WM could help maintain balance in memory resources, especially when interference threatens its stability. What prompts this hand-off? In our simulations, one reason for this was a highly prolific environment with rapid-succession memory arrivals. The any-back paradigm used in **Chapter 3** confronted participants with a continuous stream of memoranda, potentially prompting this protective hand-off to LTM after just one item.

However, the exact mechanism of this phenomenon, and whether it constitutes a pure transfer is unknown. Is the shift into LTM driven by interference—effectively "pushing" the memory out of WM – or is the representation grounded in LTM from

the outset? One possibility is that binding processes are involved in this transition, supporting the integration of WM content into episodic LTM. Such binding may be essential not only for the formation of lasting memories but also for the selection of what gets stored and when. The hippocampus is important for binding in WM as well as episodic memory (Yonelinas, 2013), offering a potential neural correlate for this mechanism. If this is the case, an important open question is whether transitions are hippocampal-dependent from the start, or whether the hippocampus becomes involved only after a critical threshold, such as uniqueness, salience, or interference with the WM representation, is crossed. One theory proposes that LTM traces stem from activity-silent memory, and predict that the hippocampus should be involved as soon as information is represented through short-term synaptic plasticity (Beukers et al., 2021). Future work could directly test this using human iEEG to examine whether and when hippocampal involvement emerges during the Any-back task.

Taken together, these findings open important avenues for integrating different forms of WM-guided learning, such as gradual statistical learning and rapid episodic "hand-off". A central question moving forward is to clarify how these learning processes relate to distinct but interconnected memory systems, particularly the role of plasticity in PFC (see also next section) alongside hippocampal and ATL contributions to LTM. Beyond episodic memories, the hippocampus is known to support memory decontextualisation and consolidation through statistical learning (Schapiro et al., 2014, 2016), raising the question of whether it is involved in statistical learning in WM (**Chapter**

2 & 6) as well as receiving one-shot information from WM for encoding in episodic memory (**Chapter 3**).

These insights also motivate the development of computational models that incorporate WM as a component of learning. Such models could help delineate the conditions under which WM supports learning, the nature of its interaction with long-term systems, and how these interactions evolve over time.

WM and prior knowledge from LTM In **Chapters 2 and 6** we explored how prior knowledge acquired through statistical learning or biases can alter representations in WM. In **Chapter 2**, despite few participants reporting conscious recollection, results showed performance benefits for repeated features, suggesting that prior learning may have influenced WM implicitly. However, this effect did not reflect a compression mechanism, as no benefit was observed for non-learned items within the same display, instead suggesting a specific facilitation for previously encountered feature conjunctions.

The task in **Chapter 6**, on the other hand, was successful in establishing a robust learning effect, in that behaviour was biased by a restricted stimulus range, possibly through a mechanism of attraction biases. This allowed for more detailed examination of how prior knowledge modulates WM representations at both behavioural and neural levels. Repeated items were represented more precisely in WM, as shown at the behavioural and neural level. However, behaviourally, this came at the cost of reduced precision for non-repeated items, suggesting information was selectively forgotten in the learning process. Neurally, old colours, more than new ones, were represented

more precisely by the neurons selective for them, whereas new colours were associated with stronger neuronal responses, indicating a dissociation at the neural level. Importantly, the results suggest that statistical learning can modulate activity in frontal regions, and that this modulation exceeds the duration of the learning task. This suggests that intermediate duration memories may be retained "within WM", in frontal areas typically associated with WM. This is in contrast to areas canonically associated with LTM (MTL and ATL areas), and points to a wider range of function for the frontal lobe. An important question to be addressed by future research is how these formats may interact or coexist when both WM and LTM contribute to behaviour.

The findings also raise questions about the temporal limits of WM plasticity. For how long does the WM attunement to environmental statistics last, and when is it forgotten to adapt to new ones? While we observed durable effects over delays up to a day, memory over the course of periods longer than a day likely involves additional consolidation processes. Future research could compare WM representations between this intermediate stage and even longer durations.

7.3 Conclusion

Returning to the traditional question of whether WM and LTM are separate systems, while the aim of this thesis was not to isolate systems, the evidence presented suggests possible dissociation and integration. The episodic "hand-off" phenomenon implies a form of separation: WM encoding initially represents a precise memory trace which, after interference, initiates a less precise LTM trace. However, we cannot con-

clude that WM is necessary for LTM. Moreover, effects of statistical learning within WM suggest a form of learning that remains embedded within the WM system itself, reframing WM as exhibiting a form of LTM the temporal frame of which is yet unclear. Further research may determine whether this is in fact distinct from other forms of learning, of if it can lead to durable memories, perhaps in conjunction with other memory systems. The directions laid out in this thesis—linking behavioural, computational, and neural evidence—provide a framework for resolving these questions.

More broadly, understanding how WM and LTM interact is not only crucial for cognitive neuroscience, but also for understanding the development of human cognition more generally. The interplay between prior knowledge and WM can create a virtuous cycle as laid out by Jones et al. (2007) for verbal learning. By drawing on prior information to compress or structure information in WM (Brady et al., 2009), individuals may free up capacity, which in turn can facilitate further learning or encoding of episodic memories (Forsberg et al., 2023). Applications to these findings include educational practice, such as designing curricula that take into account prior knowledge to enhance learning as well as keeping learning material within WM capacity. From a neuropsychological perspective, researching the interplay between WM and LTM may help more specifically characterise cognitive deficits in memory across neurological conditions. Ultimately, this work contributes to a growing understanding of memory as interacting processes that support flexible and adaptive cognition.

References

Adam, K. C. S., Rademaker, R. L., & Serences, J. T. (2022). Evidence for, and challenges to, sensory recruitment models of visual working memory. In *Visual Memory*. Routledge.

Adam, K. C. S., Robison, M. K., & Vogel, E. K. (2018). Contralateral Delay Activity Tracks Fluctuations in Working Memory Performance. *Journal of Cognitive Neuroscience*, *30*(9), 1229–1240. https://doi.org/10.1162/jocn_a_01233

Adam, K. C. S., Zhao, C., & Vogel, E. K. (2023). Behavioral signatures of the rapid recruitment of long-term memory to overcome working memory capacity limits. https://doi.org/10.31234/osf.io/2h5pf

Alberini, C. M., & LeDoux, J. E. (2013). Memory reconsolidation. *Current Biology*, *23*(17), R746–R750. https://doi.org/10.1016/j.cub.2013.06.046

Allen, M. G., Destefano, I., & Brady, T. F. (2021). Chunks are not "Content-Free": Hierarchical Representations Preserve Perceptual Detail within Chunks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/2b0846p3

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2006). Is the binding of visual features in working memory resource-demanding? *Journal of Experimental Psychology: General*, *135*(2), 298. https://doi.org/10.1037/0096-3445.135.2.298

Allen, R. J., Vargha-Khadem, F., & Baddeley, A. D. (2014). Item-location binding in working memory: Is it hippocampus-dependent? *Neuropsychologia*, *59*, 74–84. https://doi.org/10.1016/j.neuropsychologia.2014.04.013

Allon, A. S., Vixman, G., & Luria, R. (2019). Gestalt grouping cues can improve filtering performance in visual working memory. *Psychological Research*, *83*(8), 1656–1672. https://doi.org/10.1007/s00426-018-1032-5

Alvarez, G., & Cavanagh, P. (2004). The Capacity of Visual Short-Term Memory is Set Both by Visual Information Load and by Number of Objects. *Psychological Science*, *15*(2), 106–111. https://doi.org/10.1111/j.0963-7214.2004.01502006.x

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annual Review of Psychology*, *56*(1), 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Ashourian, P., & Loewenstein, Y. (2011). Bayesian Inference Underlies the Contraction Bias in Delayed Comparison Tasks (A. G. Dyer, Ed.). *PLoS ONE*, *6*(5), e19551. https://doi.org/10.1371/journal.pone.0019551

Asp, I., Störmer, V. S., & Brady, T. (2019). Greater visual working memory capacity for visually-matched stimuli when they are recognized as meaningful. https://doi.org/10.1162/jocn_a_01693

Atkinson, R. C., & Shiffrin, R. M. (1971). The Control of Short-Term Memory. *Scientific American*, *225*(2), 82–91. https://www.jstor.org/stable/24922803

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological science*, *18*(7), 622–628. https://doi.org/10.1111/j.1467-9280.2007.01949.x

Axmacher, N., Henseler, M. M., Jensen, O., Weinreich, I., Elger, C. E., & Fell, J. (2010). Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proceedings of the National Academy of Sciences*, *107*(7), 3228–3233. https://doi.org/10.1073/pnas.0911531107

Axmacher, N., Schmitz, D. P., Weinreich, I., Elger, C. E., & Fell, J. (2008). Interaction of Working Memory and Long-Term Memory in the Medial Temporal Lobe. *Cerebral Cortex*, *18*(12), 2868–2878. https://doi.org/10.1093/cercor/bhn045

Backus, A. R., Schoffelen, J.-M., Szebényi, S., Hanslmayr, S., & Doeller, C. F. (2016). Hippocampal-Prefrontal Theta Oscillations Support Memory Integration. *Current Biology*, *26*(4), 450–457. https://doi.org/10.1016/j.cub.2015.12.048

Baddeley, A., Conway, M., Aggleton, J., & Baddeley, A. (2001). The concept of episodic memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1345–1350. https://doi.org/10.1098/rstb.2001.0957

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning & Verbal Behavior*, *9*(2), 176–189. https://doi.org/10.1016/S0022-5371(70)80048-2

Baddeley, A. D., Allen, R. J., & Hitch, G. J. (2011). Binding in visual working memory: The role of the episodic buffer. *Neuropsychologia*, *49*(6), 1393–1400. https://doi.org/10.1016/j.neuropsychologia.2010.12.042

Baddeley, A. D., & Hitch, G. (1974, January 1). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (pp. 47–89, Vol. 8). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Bae, G.-Y. (2021). Neural evidence for categorical biases in location and orientation representations in a working memory task. *NeuroImage*, *240*, 118366. https://doi.org/10.1016/j.neuroimage.2021.118366

Bae, G.-Y., & Luck, S. J. (2017). Interactions between visual working memory representations. *Attention, Perception, & Psychophysics*, *79*(8), 2376–2395. https://doi.org/10.3758/s13414-017-1404-8

Bae, G.-Y., & Luck, S. J. (2020). Serial dependence in vision: Merely encoding the previous-trial target is not enough. *Psychonomic Bulletin & Review*, *27*(2), 293–300. https://doi.org/10.3758/s13423-019-01678-7

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744–763. https://doi.org/10.1037/xge0000076

Balan, P. F., & Ferrera, V. P. (2003). Effects of Gaze Shifts on Maintenance of Spatial Memory in Macaque Frontal Eye Field. *Journal of Neuroscience*, *23*(13), 5446–5454. https://doi.org/10.1523/JNEUROSCI.23-13-05446.2003

Balkenius, C., Tjøstheim, T. A., Johansson, B., Wallin, A., & Gärdenfors, P. (2020). The Missing Link Between Memory and Reinforcement Learning. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.560080

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bartsch, L. M., Frischkorn, G. T., & Shepherdson, P. (2024). When Load is Low, Working Memory is Shielded From Long-Term Memory's Influence. *Journal of Cognition*, *7*(1), 44. https://doi.org/10.5334/joc.368

Bartsch, L. M., & Oberauer, K. (2023). The contribution of episodic long-term memory to working memory for bindings. *Cognition*, *231*, 105330. https://doi.org/10.1016/j.cognition.2022.105330

Bartsch, L. M., & Shepherdson, P. (2022). Freeing capacity in working memory (WM) through the use of long-term memory (LTM) representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(4), 465–482. https://doi.org/10.1037/xlm0001024

Bates, C. J., Alvarez, G. A., & Gershman, S. J. (2024). Scaling models of visual working memory to natural images. *Communications Psychology*, *2*(1), 1–8. https://doi.org/10.1038/s44271-023-00048-3

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 11. https://doi.org/10.1167/19.2.11

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *arXiv: Methodology*. https://arxiv.org/pdf/1506.04967

Bates, D. M. (2010). Lme4: Mixed-effects modeling with R.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7–7. https://doi.org/10.1167/9.10.7

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, *11*(10), 6–6. https://doi.org/10.1167/11.10.6

Bays, P. M., & Husain, M. (2008). Dynamic Shifts of Limited Working Memory Resources in Human Vision. *Science*, *321*(5890), 851–854. https://doi.org/10.1126/science.1158023

Bays, P. M. (2014). Noise in Neural Populations Accounts for Errors in Working Memory. *Journal of Neuroscience*, *34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431–438. https://doi.org/10.1016/j.tics.2015.06.004

Bays, P. M., Schneegans, S., Ma, W. J., & Brady, T. F. (2024). Representation and computation in visual working memory. *Nature Human Behaviour*, *8*(6), 1016–1034. https://doi.org/10.1038/s41562-024-01871-2

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622–1631. https://doi.org/10.1016/j.neuropsychologia.2010.12.023

Ben-Artzi, I., Luria, R., & Shahar, N. (2022). Working memory capacity estimates moderate value learning for outcome-irrelevant features. *Scientific Reports*, *12*(1), 19677. https://doi.org/10.1038/s41598-022-21832-x

Benchenane, K., Tiesinga, P. H., & Battaglia, F. P. (2011). Oscillations in the prefrontal cortex: A gateway to memory and attention. *Current Opinion in Neurobiology*, *21*(3), 475–485. https://doi.org/10.1016/j.conb.2011.01.004

Berens, P. (2009). **CircStat** : A *MATLAB* Toolbox for Circular Statistics. *Journal of Statistical Software*, *31*(10). https://doi.org/10.18637/jss.v031.i10

Bergmann, H. C., Kiemeneij, A., Fernández, G., & Kessels, R. P. (2013). Early and late stages of working-memory maintenance contribute differentially to long-term memory formation. *Acta Psychologica*, *143*(2), 181–190. https://doi.org/10.1016/j.actpsy.2013.02.009

Bettencourt, K. C., & Xu, Y. (2016). Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience*, *19*(1), 150–157. https://doi.org/10.1038/nn.4174

Beukers, A. O., Buschman, T. J., Cohen, J. D., & Norman, K. A. (2021). Is Activity Silent Working Memory Simply Episodic Memory? *Trends in Cognitive Sciences*, *25*(4), 284–293. https://doi.org/10.1016/j.tics.2021.01.003

Biderman, N., Gershman, S. J., & Shohamy, D. (2023). The role of memory in counterfactual valuation. *Journal of Experimental Psychology: General*, *152*(6), 1754–1767. https://doi.org/10.1037/xge0001364

Biderman, N., Luria, R., Teodorescu, A. R., Hajaj, R., & Goshen-Gottstein, Y. (2019). Working Memory Has Better Fidelity Than Long-Term Memory: The Fidelity Constraint Is Not a General Property of Memory After All. *Psychological Science*, *30*(2), 223–237. https://doi.org/10.1177/0956797618813538

Biderman, N., & Shohamy, D. (2021). Memory and decision making interact to shape the value of unchosen options. *Nature Communications*, *12*(1), 4648. https://doi.org/10.1038/s41467-021-24907-x

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Binney, R. J., Embleton, K. V., Jefferies, E., Parker, G. J. M., & Lambon Ralph, M. A. (2010). The Ventral and Inferolateral Aspects of the Anterior Temporal Lobe Are Crucial in Semantic Memory: Evidence from a Novel Direct Comparison of Distortion-Corrected fMRI, rTMS, and Semantic Dementia. *Cerebral Cortex*, *20*(11), 2728–2738. https://doi.org/10.1093/cercor/bhq019

Blalock, L. D. (2015). Stimulus familiarity improves consolidation of visual working memory representations. *Attention, Perception, & Psychophysics*, *77*(4), 1143–1158. https://doi.org/10.3758/s13414-014-0823-z

Blumenfeld, R. S., & Ranganath, C. (2006). Dorsolateral prefrontal cortex promotes long-term memory formation through its role in working memory organization.

Journal of Neuroscience, *26*(3), 916–925. https://doi.org/10.1523/JNEUROSCI.2353-05.2006

Boboeva, V., Pezzotta, A., Clopath, C., & Akrami, A. (2023, September 26). *From recency to central tendency biases in working memory: A unifying network model*. https://doi.org/10.1101/2022.05.16.491352

Borders, A. A., Ranganath, C., & Yonelinas, A. P. (2021). The hippocampus supports high-precision binding in visual working memory. *Hippocampus*, *32*(3). https://doi.org/10.1002/hipo.23401

Bouchacourt, F., & Buschman, T. J. (2019). A Flexible Model of Working Memory. *Neuron*, *103*(1), 147–160.e8. https://doi.org/10.1016/j.neuron.2019.04.020

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4–4. https://doi.org/10.1167/11.5.4

Brady, T., & Störmer, V. S. (2020a). *Greater capacity for objects than colors in visual working memory: Comparing memory across stimulus spaces requires maximally dissimilar foils*.

Brady, T., & Störmer, V. S. (2020b). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Brady, T. F., & Alvarez, G. A. (2015a). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of vision*, *15*(15), 6–6. https://doi.org/10.1167/15.15.6

Brady, T. F., & Alvarez, G. A. (2015b). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 921. https://doi.org/10.1037/xlm0000075

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502. https://doi.org/10.1037/a0016797

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual Long-Term Memory Has the Same Limit on Fidelity as Visual Working Memory. *Psychological Science*, *24*(6), 981–990. https://doi.org/10.1177/0956797612465439

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. https://doi.org/10.1073/pnas.1520027113

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109. https://doi.org/10.1037/a0030779

Bruce, A. D., Gardner, E. J., & Wallace, D. J. (1987). Dynamics and statistical mechanics of the Hopfield model. *Journal of Physics A: Mathematical and General*, *20*(10), 2909–2934. https://doi.org/10.1088/0305-4470/20/10/035

Bruning, A. L., & Lewis-Peacock, J. A. (2020). Long-term memory guides resource allocation in working memory. *Scientific Reports*, *10*(1), 22161. https://doi.org/10.1038/s41598-020-79108-1

Burak, Y., & Fiete, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences*, *109*(43), 17645–17650. https://doi.org/10.1073/pnas.1117386109

Chekaf, M., Cowan, N., & Mathy, F. (2016a). Chunk formation in immediate memory and how it relates to data compression. *Cognition*, *155*, 96–107. https://doi.org/10.1016/j.cognition.2016.05.024

Chekaf, M., Cowan, N., & Mathy, F. (2016b). Chunk formation in immediate memory and how it relates to data compression. *Cognition*, *155*, 96–107. https://doi.org/10.1016/j.cognition.2016.05.024

Chen, Y.-T., van Ede, F., & Kuo, B.-C. (2022). Alpha oscillations track content-specific working memory capacity. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2296-21.2022

Christophel, T. B., Hebart, M. N., & Haynes, J.-D. (2012). Decoding the Contents of Visual Short-Term Memory from Human Visual and Parietal Cortex. *Journal of Neuroscience*, *32*(38), 12983–12989. https://doi.org/10.1523/JNEUROSCI.0184-12.2012

Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C., & Haynes, J.-D. (2018). Cortical specialization for attended versus unattended working memory. *Nature Neuroscience*, *21*(4), 494–496. https://doi.org/10.1038/s41593-018-0094-4

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, *21*(2), 111–124. https://doi.org/10.1016/j.tics.2016.12.007

Chung, Y. H., Brady, T. F., & Störmer, V. S. (2023). No Fixed Limit for Storing Simple Visual Features: Realistic Objects Provide an Efficient Scaffold for Holding Features in Mind. *Psychological Science*, *34*(7), 784–793. https://doi.org/10.1177/09567976231171339

Chunharas, C., Rademaker, R. L., Brady, T., & Serences, J. (2019). Adaptive memory distortion in visual working memory. https://doi.org/10.31234/osf.io/e3m5a

Clark, K. L., Noudoost, B., & Moore, T. (2012). Persistent Spatial Information in the Frontal Eye Field during Object-Based Short-Term Memory. *Journal of Neuroscience*, *32*(32), 10907–10914. https://doi.org/10.1523/JNEUROSCI.1450-12.2012

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. https://doi.org/10.1162/jocn_a_01238

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. *Biological Psychiatry*, *82*(6), 431–439. https://doi.org/10.1016/j.biopsych.2017.05.017

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017a). Working Memory Load Strengthens Reward Prediction Errors. *Journal of Neuroscience*, *37*(16), 4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017b). Working Memory Load Strengthens Reward Prediction Errors. *Journal of Neuroscience*, *37*(16), 4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, *115*(10), 2502–2507. https://doi.org/10.1073/pnas.1720963115

Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. *The Journal of Neuroscience*, *34*(41), 13747–13756. https://doi.org/10.1523/JNEUROSCI.0989-14.2014

Colzato, L. S., Raffone, A., & Hommel, B. (2006). What do we learn from binding features? Evidence for multilevel feature integration. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 705. https://doi.org/10.1037/0096-1523.32.3.705

Compte, A. (2000). Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cerebral Cortex*, *10*(9), 910–923. https://doi.org/10.1093/cercor/10.9.910

Constantinidis, C., & Steinmetz, M. A. (1996). Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task. *Journal of Neurophysiology*, *76*(2), 1352–1355. https://doi.org/10.1152/jn.1996.76.2.1352

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X.-L., Wang, M., & Arnsten, A. F. (2018). Persistent Spiking Activity Underlies Working Memory. *The Journal of Neuroscience*, *38*(32), 7020–7028. https://doi.org/10.1523/JNEUROSCI.2486-17.2018

Cooper, R. A., Kensinger, E. A., & Ritchey, M. (2019). Memories Fade: The Relationship Between Memory Vividness and Remembered Visual Salience. *Psychological Science*, *30*(5), 657–668. https://doi.org/10.1177/0956797619836093

Cooper, R. A., & Ritchey, M. (2019). Cortico-hippocampal network connections support the multidimensional quality of episodic memory. *eLife*, *8*, e45591. https://doi.org/10.7554/eLife.45591

Cowan, N. (1999, April 13). An Embedded-Processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (1st ed., pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Curby, K. M., & Gauthier, I. (2007). A visual short-term memory advantage for faces. *Psychonomic bulletin & review*, *14*(4), 620–628. https://doi.org/10.3758/bf03196811

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, *28*(6), 923–938. https://doi.org/10.3758/bf03209340

Danker, J. F., & Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, *136*(1), 87–102. https://doi.org/10.1037/a0017937

Davachi, L., & Wagner, A. D. (2002). Hippocampal Contributions to Episodic Encoding: Insights From Relational and Item-Based Learning. *Journal of Neurophysiology*, *88*(2), 982–990. https://doi.org/10.1152/jn.2002.88.2.982

Davoudi, S., Parto Dezfouli, M., Knight, R. T., Daliri, M. R., & Johnson, E. L. (2021). Prefrontal Lesions Disrupt Posterior Alpha–Gamma Coordination of Visual Working Memory Representations. *Journal of Cognitive Neuroscience*, 1–13. https://doi.org/10.1162/jocn_a_01715

de Vries, I. E., Slagter, H. A., & Olivers, C. N. (2020). Oscillatory control over representational states in working memory. *Trends in Cognitive Sciences*, *24*(2), 150–162. https://doi.org/10.1016/j.tics.2019.11.006

Debnath, T., & Song, M. (2021). Fast Optimal Circular Clustering and Applications on Round Genomes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(6), 2061–2071. https://doi.org/10.1109/TCBB.2021.3077573

D'Esposito, M., & Postle, B. R. (2015). The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Doody, M., Van Swieten, M. M. H., & Manohar, S. G. (2022). Model-based learning retrospectively updates model-free values. *Scientific Reports*, *12*(1), 2358. https://doi.org/10.1038/s41598-022-05567-3

Duffy, S., Huttenlocher, J., Hedges, L. V., & Elizabeth Crawford, L. (2010). Category effects on stimulus estimation: Shifting and skewed frequency distributions. *Psy-*

chonomic Bulletin & Review, *17*(2), 224–230. https://doi.org/10.3758/PBR.17.2.224

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, *3*(11), 1184–1191. https://doi.org/10.1038/81460

Eichenbaum, H. (2013). Memory on time. *Trends in cognitive sciences*, *17*(2), 81–88. https://doi.org/10.1016/j.tics.2012.12.007

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron*, *87*(4), 893–905. https://doi.org/10.1016/j.neuron.2015.07.013

Ezzyat, Y., & Olson, I. R. (2008). The medial temporal lobe and visual working memory: Comparisons across tasks, delays, and visual similarity. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(1), 32–40. https://doi.org/10.3758/CABN.8.1.32

Fairhall, S. L., & Caramazza, A. (2013). Brain Regions That Represent Amodal Conceptual Knowledge. *Journal of Neuroscience*, *33*(25), 10552–10558. https://doi.org/10.1523/JNEUROSCI.0051-13.2013

Fiebig, F., & Lansner, A. (2017). A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. *Journal of Neuroscience*, *37*(1), 83–96. https://doi.org/10.1523/JNEUROSCI.1989-16.2016

Fischer, C., Czoschke, S., Peters, B., Rahm, B., Kaiser, J., & Bledowski, C. (2020). Context information supports serial dependence of multiple visual objects across

memory episodes. *Nature Communications*, *11*(1), 1932. https://doi.org/10.1038/s41467-020-15874-w

Forsberg, A., Adams, E. J., & Cowan, N. (2021, January 1). Chapter One - The role of working memory in long-term learning: Implications for childhood development. In K. D. Federmeier (Ed.), *Psychology of Learning and Motivation* (pp. 1–45, Vol. 74). Academic Press. https://doi.org/10.1016/bs.plm.2021.02.001

Forsberg, A., Guitard, D., Adams, E. J., Pattanakul, D., & Cowan, N. (2022). Children's long-term retention is directly constrained by their working memory capacity limitations. *Developmental Science*, *25*(2), e13164. https://doi.org/10.1111/desc.13164

Forsberg, A., Guitard, D., Adams, E. J., Pattanakul, D., & Cowan, N. (2023). Working Memory Constrains Long-Term Memory in Children and Adults: Memory of Objects and Bindings. *Journal of Intelligence*, *11*(5), 94. https://doi.org/10.3390/jintelligence11050094

Forsberg, A., Guitard, D., & Cowan, N. (2021). Working memory limits severely constrain long-term retention. *Psychonomic Bulletin & Review*, *28*(2), 537–547. https://doi.org/10.3758/s13423-020-01847-z

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12), 27–27. https://doi.org/10.1167/10.12.27

Fougnie, D., Cormiea, S. M., & Alvarez, G. A. (2013). Object-based benefits without object-based representations. *Journal of Experimental Psychology: General*, *142*(3), 621–626. https://doi.org/10.1037/a0030300

Fougnie, D., Cormiea, S. M., Kanabar, A., & Alvarez, G. A. (2016). Strategic trade-offs between quantity and quality in working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1231.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. *Science*, *291*(5502), 312–316. https://doi.org/10.1126/science.291.5502.312

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A Comparison of Primate Prefrontal and Inferior Temporal Cortices during Visual Categorization. *Journal of Neuroscience*, *23*(12), 5235–5246. https://doi.org/10.1523/JNEUROSCI.23-12-05235.2003

Fuentemilla, L., Penny, W. D., Cashdollar, N., Bunzeck, N., & Düzel, E. (2010). Theta-Coupled Periodic Replay in Working Memory. *Current Biology*, *20*(7), 606–612. https://doi.org/10.1016/j.cub.2010.01.057

Fukuda, K., Awh, E., & Vogel, E. K. (2010). Discrete capacity limits in visual working memory. *Current opinion in neurobiology*, *20*(2), 177–182. https://doi.org/10.1016/j.conb.2010.03.005

Fukuda, K., Mance, I., & Vogel, E. K. (2015). Power Modulation and Event-Related Slow Wave Provide Dissociable Correlates of Visual Working Memory. *Journal*

of Neuroscience, *35*(41), 14009–14016. https://doi.org/10.1523/JNEUROSCI.5003-14.2015

Fukuda, K., & Woodman, G. F. (2017). Visual working memory buffers information retrieved from visual long-term memory. *Proceedings of the National Academy of Sciences*, *114*(20), 5306–5311. https://doi.org/10.1073/pnas.1617874114

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.1989.61.2.331

Funahashi, S. (2006). Prefrontal cortex and working memory processes. *Neuroscience*, *139*(1), 251–261. https://doi.org/10.1016/j.neuroscience.2005.07.003

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654. https://doi.org/10.1126/science.173.3997.652

Ganguli, S., & Sompolinsky, H. (2012). Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis. *Annual Review of Neuroscience*, *35*(1), 485–508. https://doi.org/10.1146/annurev-neuro-062111-150410

Gao, Z., Yin, J., Xu, H., Shui, R., & Shen, M. (2011). Tracking object number or information load in visual working memory: Revisiting the cognitive implication of contralateral delay activity. *Biological Psychology*, *87*(2), 296–302. https://doi.org/10.1016/j.biopsycho.2011.03.013

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135. https://doi.org/10.1016/j.tics.2011.11.014

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*, *68*, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Gluck, M. A., Meeter, M., & Myers, C. E. (2003). Computational models of the hippocampal region: Linking incremental learning and episodic memory. *Trends in Cognitive Sciences*, *7*(6), 269–276. https://doi.org/10.1016/S1364-6613(03)00105-0

Goldman-Rakic, P. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485. https://doi.org/10.1016/0896-6273(95)90304-6

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic Updating of Working Memory Resources for Visual Objects. *Journal of Neuroscience*, *31*(23), 8502–8511. https://doi.org/10.1523/JNEUROSCI.0208-11.2011

Greene, N. R., Guitard, D., Forsberg, A., Cowan, N., & Naveh-Benjamin, M. (2024). Working memory limitations constrain visual episodic long-term memory at both specific and gist levels of representation. *Memory & Cognition*, *52*(8), 1958–1982. https://doi.org/10.3758/s13421-024-01593-w

Griffin, I. C., & Nobre, A. C. (2003). Orienting Attention to Locations in Internal Representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

Grogan, J. P., Randhawa, G., Kim, M., & Manohar, S. G. (2022). Motivation improves working memory by two processes: Prioritisation and retrieval thresholds. *Cognitive Psychology*, *135*, 101472. https://doi.org/10.1016/j.cogpsych.2022.101472

Hakim, N., Awh, E., & Vogel, E. K. (2020, November 5). Manifold Visual Working Memory. In *Working Memory* (pp. 311–332). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0011

Hannula, D. E., & Ranganath, C. (2008). Medial Temporal Lobe Activity Predicts Successful Relational Memory Binding. *The Journal of Neuroscience*, *28*(1), 116–124. https://doi.org/10.1523/JNEUROSCI.3086-07.2008

Hardman, K. O., & Cowan, N. (2015). Remembering complex objects in visual working memory: Do capacity limits restrict objects or features? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 325–347. https://doi.org/10.1037/xlm0000031

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 30–54. https://doi.org/10.1037/xhp0000290

Harris, J. J., Jolivet, R., & Attwell, D. (2012). Synaptic Energy Use and Supply. *Neuron*, *75*(5), 762–777. https://doi.org/10.1016/j.neuron.2012.08.019

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Hart, E., & Huk, A. C. (2020). Recurrent circuit dynamics underlie persistent activity in the macaque frontoparietal network (E. Salinas, T. E. Behrens, E. Salinas, & A. Compte, Eds.). *eLife*, *9*, e52460. https://doi.org/10.7554/eLife.52460

Hasselmo, M. E., & Stern, C. E. (2006). Mechanisms underlying working memory for novel information. *Trends in cognitive sciences*, *10*(11), 487–493. https://doi.org/10.1016/j.tics.2006.09.005

Haxby, J. V., Petit, L., Ungerleider, L. G., & Courtney, S. M. (2000). Distinguishing the functional roles of multiple regions in distributed neural systems for visual working memory. *Neuroimage*, *11*(2), 145–156. https://doi.org/10.1006/nimg.2000.0592

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: Making working memory work. *Neuroscience*, *139*(1), 105–118. https://doi.org/10.1016/j.neuroscience.2005.04.067

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1601–1613. https://doi.org/10.1098/rstb.2007.2055

Hebb, D. (1961). Distinctive features of learning in the higher animal. In B. F. Delafresnaye (Ed.), *Brain mechanisms and learning*. Blackwell.

Hedayati, S., O'Donnell, R. E., & Wyble, B. (2022). A model of working memory for latent representations. *Nature Human Behaviour*, 1–11. https://doi.org/10.1038/s41562-021-01264-9

Helfer, P., & Shultz, T. R. (2020). A computational model of systems memory consolidation and reconsolidation. *Hippocampus*, *30*(7), 659–677. https://doi.org/10.1002/hipo.23187

Henson, R. N. A., Rugg, M. D., Shallice, T., Josephs, O., & Dolan, R. J. (1999). Recollection and Familiarity in Recognition Memory: An Event-Related Functional Magnetic Resonance Imaging Study. *Journal of Neuroscience*, *19*(10), 3962–3972. https://doi.org/10.1523/JNEUROSCI.19-10-03962.1999

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*, *3*(3–4), 463–495. https://doi.org/10.1080/09658219508253161

Hoffman, P., Binney, R. J., & Lambon Ralph, M. A. (2015). Differing contributions of inferior prefrontal and anterior temporal cortex to concrete and abstract conceptual knowledge. *Cortex*, *63*, 250–266. https://doi.org/10.1016/j.cortex.2014.09.001

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(3), 293–328. https://doi.org/10.1037/rev0000094

Hojjati, F., Motahharynia, A., Adibi, A., Adibi, I., & Sanayei, M. (2024, March 25). *Resource or slot model in visual working memory: Are they different?* https://doi.org/10.1101/2024.02.08.579494

Hollingworth, H. L. (1910). The Central Tendency of Judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*(17), 461–469. https://doi.org/10.2307/2012819

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Hoskin, A. N., Bornstein, A. M., Norman, K. A., & Cohen, J. D. (2019). Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(2), 338–354. https://doi.org/10.3758/s13415-018-00674-z

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Huang, L. (2011). Familiarity does not aid access to features. *Psychonomic Bulletin & Review*, *18*(2), 278–286. https://doi.org/10.3758/s13423-011-0052-6

Huang, L. (2020). Distinguishing target biases and strategic guesses in visual working memory. *Attention, Perception, & Psychophysics*, *82*(3), 1258–1270. https://doi.org/10.3758/s13414-019-01913-2

Huang, L., & Awh, E. (2018). Chunking in working memory via content-free labels. *Scientific Reports*, *8*(1), 23. https://doi.org/10.1038/s41598-017-18157-5

Ikkai, A., McCollough, A. W., & Vogel, E. K. (2010). Contralateral Delay Activity Provides a Neural Measure of the Number of Representations in Visual Working

Memory. *Journal of Neurophysiology*, *103*(4), 1963–1968. https://doi.org/10.1152/jn.00978.2009

Infanti, E., Hickey, C., & Turatto, M. (2015). Reward associations impact both iconic and visual working memory. *Vision Research*, *107*, 22–29. https://doi.org/10.1016/j.visres.2014.11.008

Jackson, M. C., & Raymond, J. E. (2008). Familiarity enhances visual working memory for faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 556. https://doi.org/10.1037/0096-1523.34.3.556

Jeneson, A., & Squire, L. R. (2012). Working memory, long-term memory, and medial temporal lobe function. *Learning & Memory*, *19*(1), 15–25. https://doi.org/10.1101/lm.024018.111

Jeneson, A., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2012). Visual Working Memory Capacity and the Medial Temporal Lobe. *Journal of Neuroscience*, *32*(10), 3584–3589. https://doi.org/10.1523/JNEUROSCI.6444-11.2012

Jiang, Y. V., Lee, H. J., Asaad, A., & Remington, R. (2016). Similarity effects in visual working memory. *Psychonomic bulletin & review*, *23*(2), 476–482. https://doi.org/10.3758/s13423-015-0905-5

Jones, G., Gobet, F., & Pine, J. M. (2007). Linking working memory and long-term memory: A computational model of the learning of new words. *Developmental Science*, *10*(6), 853–873. https://doi.org/10.1111/j.1467-7687.2007.00638.x

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. *Annual Review of Psychology*, *59*, 193–224. https://doi.org/10.1146/annurev.psych.59.103006.093615

Jou, J., Leka, G. E., Rogers, D. M., & Matus, Y. E. (2004). Contraction Bias in Memorial Quantifying Judgment: Does It Come from a Stable Compressed Memory Representation or a Dynamic Adaptation Process? *The American Journal of Psychology*, *117*(4), 543–564. https://doi.org/10.2307/4148991

Kaiser, D., Stein, T., & Peelen, M. V. (2015). Real-world spatial regularities affect visual working memory for objects. *Psychonomic bulletin & review*, *22*(6), 1784–1790. https://doi.org/10.3758/s13423-015-0833-4

Kamiński, J., Sullivan, S., Chung, J. M., Ross, I. B., Mamelak, A. N., & Rutishauser, U. (2017). Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nature Neuroscience*, *20*(4), 590–601. https://doi.org/10.1038/nn.4509

Kandel, E. R., Dudai, Y., & Mayford, M. R. (2014). The Molecular and Systems Biology of Memory. *Cell*, *157*(1), 163–186. https://doi.org/10.1016/j.cell.2014.03.001

Kleiner, M. (2007). What's new in Psychtoolbox-3? *Perception*, *36*(14), 1–16.

Koyluoglu, O. O., Pertzov, Y., Manohar, S., Husain, M., & Fiete, I. R. (2017). Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity (L. Davachi, Ed.). *eLife*, *6*, e22225. https://doi.org/10.7554/eLife.22225

Kriegeskorte, N., & Wei, X.-X. (2021). Neural tuning and representational geometry. *Nature Reviews Neuroscience*, *22*(11), 703–718. https://doi.org/10.1038/s41583-021-00502-3

Kumar, A. A. (2020). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01792-x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences*, *107*(6), 2717–2722. https://doi.org/10.1073/pnas.0907307107

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lee, S.-H., & Baker, C. I. (2016). Multi-Voxel Decoding and the Topography of Maintained Information During Visual Working Memory. *Frontiers in Systems Neuroscience*, *10*. https://doi.org/10.3389/fnsys.2016.00002

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nature Neuroscience*, *16*(8), 997–999. https://doi.org/10.1038/nn.3452

Lennie, P. (2003). The Cost of Cortical Computation. *Current Biology*, *13*(6), 493–497. https://doi.org/10.1016/S0960-9822(03)00135-0

Lenth, R. V., Banfai, B., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Piaskowski, J., Riebl, H., & Singmann, H. (2024, December 12). *Emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.10.6). https://cran.r-project.org/web/packages/emmeans/index.html

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140

Lewis-Peacock, J. A., & Postle, B. R. (2008). Temporary Activation of Long-Term Memory Supports Working Memory. *Journal of Neuroscience*, *28*(35), 8765–8771. https://doi.org/10.1523/JNEUROSCI.1953-08.2008

Lezak, M. D. (2004). *Neuropsychological Assessment*. Oxford University Press.

Li, X., Xiong, Z., Theeuwes, J., & Wang, B. (2020). Visual memory benefits from prolonged encoding time regardless of stimulus type. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000847

Liang, Y., Pertzov, Y., Nicholas, J. M., Henley, S. M. D., Crutch, S., Woodward, F., Leung, K., Fox, N. C., & Husain, M. (2016). Visual short-term memory binding deficit in familial Alzheimer's disease. *Cortex*, *78*, 150–164. https://doi.org/10.1016/j.cortex.2016.01.015

Lin, P.-H., & Luck, S. J. (2009). The influence of similarity on visual working memory representations. *Visual Cognition*, *17*(3), 356–372. https://doi.org/10.1080/13506280701766313

Linden, D. E. J., Bittner, R. A., Muckli, L., Waltz, J. A., Kriegeskorte, N., Goebel, R., Singer, W., & Munk, M. H. J. (2003). Cortical capacity constraints for visual working memory: Dissociation of fMRI load effects in a fronto-parietal network. *NeuroImage*, *20*(3), 1518–1530. https://doi.org/10.1016/j.neuroimage.2003.07.021

Lindsay, G. W., Rigotti, M., Warden, M. R., Miller, E. K., & Fusi, S. (2017). Hebbian Learning in a Random Network Captures Selectivity Properties of the Prefrontal Cortex. *Journal of Neuroscience*, *37*(45), 11021–11036. https://doi.org/10.1523/JNEUROSCI.1222-17.2017

Lindsey, J., & Litwin-Kumar, A. (2022). Theory of systems memory consolidation via recall-gated plasticity. https://doi.org/10.1101/2022.12.08.519638

Liu, B., Li, X., Theeuwes, J., & Wang, B. (2022). Long-term memory retrieval bypasses working memory. *NeuroImage*, *261*, 119513. https://doi.org/10.1016/j.neuroimage.2022.119513

Liu, J., Zhang, H., Yu, T., Ni, D., Ren, L., Yang, Q., Lu, B., Wang, D., Heinen, R., Axmacher, N., & Xue, G. (2020). Stable maintenance of multiple representational formats in human visual short-term memory. *Proceedings of the National Academy of Sciences*, 202006752. https://doi.org/10.1073/pnas.2006752117

Logie, R. H., Brockmole, J. R., & Vandenbroucke, A. R. E. (2009). Bound feature combinations in visual short-term memory are fragile but influence long-term learning. *Visual Cognition*, *17*(1–2), 160–179. https://doi.org/10.1080/13506280802228411

Lohnas, L. J., Duncan, K., Doyle, W. K., Thesen, T., Devinsky, O., & Davachi, L. (2018). Time-resolved neural reinstatement and pattern separation during memory decisions in human hippocampus. *Proceedings of the National Academy of Sciences*, *115*(31), E7418–E7427. https://doi.org/10.1073/pnas.1717088115

Lorenc, E. S., Pratte, M. S., Angeloni, C. F., & Tong, F. (2014). Expertise for upright faces improves the precision but not the capacity of visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 1975–1984.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Lund, U., Agostinelli, C., Arai, H., Gagliardi, A., García-Portugués, E., Giunchi, D., Irisson, J.-O., Pocernich, M., & Rotolo, F. (2024, August 29). *Circular: Circular Statistics* (Version 0.5-1).

Lundqvist, M., Herman, P., & Miller, E. K. (2018). Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. *Journal of Neuroscience*, *38*(32), 7013–7019. https://doi.org/10.1523/JNEUROSCI.2485-17.2018

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., & Miller, E. K. (2016). Gamma and Beta Bursts Underlie Working Memory. *Neuron*, *90*(1), 152–164. https://doi.org/10.1016/j.neuron.2016.02.028

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, *62*, 100–108. https://doi.org/10.1016/j.neubiorev.2016.01.003

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T., & Husain, M. (2019). Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2019.03.017

Marr, D. (2010, July 9). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.

Masse, N. Y., Rosen, M. C., & Freedman, D. J. (2020). Reevaluating the Role of Persistent Neural Activity in Short-Term Memory. *Trends in Cognitive Sciences*, *24*(3), 242–258. https://doi.org/10.1016/j.tics.2019.12.014

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working

memory. *Nature Neuroscience*, *22*(7), 1159–1167. https://doi.org/10.1038/s41593-019-0414-3

Mathy, F., Chekaf, M., & Cowan, N. (2018). Simple and Complex Working Memory Tasks Allow Similar Benefits of Information Compression. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.31

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362. https://doi.org/10.1016/j.cognition.2011.11.003

Matthey, L., Bays, P. M., & Dayan, P. (2015). A Probabilistic Palimpsest Model of Visual Short-term Memory. *PLOS Computational Biology*, *11*(1), e1004003. https://doi.org/10.1371/journal.pcbi.1004003

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. https://doi.org/10.1038/nrn1076

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, *11*(1), 103–107. https://doi.org/10.1038/nn2024

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

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

Miller, J. A., & Constantinidis, C. (2024). Timescales of learning in prefrontal cortex. *Nature Reviews Neuroscience*, 1–14. https://doi.org/10.1038/s41583-024-00836-8

Miller, J. A., Tambini, A., Kiyonaga, A., & D'Esposito, M. (2022). Long-term learning transforms prefrontal cortex representations during working memory. *Neuron*, *110*(22), 3805–3819.e6. https://doi.org/10.1016/j.neuron.2022.09.019

Miner, A. E., Schurgin, M. W., & Brady, T. F. (2020). Is working memory inherently more "precise" than long-term memory? Extremely high fidelity visual long-term memories for frequently encountered objects. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(8), 813. https://doi.org/10.1037/xhp0000748

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Morcom, A. (2014). Re-engaging with the past: Recapitulation of encoding operations during retrieval. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00351

Motter, B. C. (1994). Neural correlates of feature selective memory and pop-out in extrastriate area V4. *Journal of Neuroscience*, *14*(4), 2190–2199. https://doi.org/10.1523/JNEUROSCI.14-04-02190.1994

Mummery, C. J., Frackowiak, R. S. J., & Hodges, J. R. (2000). *A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. Ann Neurol 47: 36–45.*

Musfeld, P., Dutli, J., Oberauer, K., & Bartsch, L. M. (2024). Grouping in working memory guides chunk formation in long-term memory: Evidence from the Hebb effect. *Cognition*, *248*, 105795. https://doi.org/10.1016/j.cognition.2024.105795

Musfeld, P., Souza, A. S., & Oberauer, K. (2024). Testing expectations and retrieval practice modulate repetition learning of visuospatial arrays. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(5), 740–758. https://doi.org/10.1037/xlm0001298

Myers, N. E., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2018). Benefits of flexible prioritization in working memory can arise without costs. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(3), 398. https://doi.org/10.1037/xhp0000449

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: Beyond sustained internal attention. *Trends in cognitive sciences*, *21*(6), 449–461. https://doi.org/10.1016/j.tics.2017.03.010

Nagy, D. G., Török, B., & Orbán, G. (2020). Optimal forgetting: Semantic compression of episodic memories. *PLOS Computational Biology*, *16*(10), e1008367. https://doi.org/10.1371/journal.pcbi.1008367

Ngiam, W. X. Q., Brissenden, J. A., & Awh, E. (2019). "Memory compression" effects in visual working memory are contingent on explicit long-term memory. *Journal of Experimental Psychology: General*, *148*(8), 1373–1385. https://doi.org/10.1037/xge0000649

Ngiam, W. X. Q. (2023). Mapping visual working memory models to a theoretical framework. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02356-5

Ngiam, W. X., Khaw, K. L., Holcombe, A. O., & Goodbourn, P. T. (2019). Visual working memory for letters varies with familiarity but not complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(10), 1761.

Niklaus, M., Nobre, A. C., & Van Ede, F. (2017). Feature-based attentional weighting and spreading in visual working memory. *Scientific reports*, *7*(1), 1–10.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *The Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Norman, K. A., Detre, G., & Polyn, S. M. (2008). Computational models of episodic memory. In *The Cambridge handbook of computational psychology* (pp. 189–225). Cambridge University Press. https://doi.org/10.1017/CBO9780511816772.011

Norris, D., Kalm, K., & Hall, J. (2019). Chunking and redintegration in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000762

Noudoost, B., Clark, K. L., & Moore, T. (2021). Working memory gates visual input to primate prefrontal neurons. *eLife*, *10*, e64814. https://doi.org/10.7554/eLife.64814

Oberauer, K., Awh, E., & Sutterer, D. W. (2017). The role of long-term memory in a test of visual working memory: Proactive facilitation but no proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(1), 1–22. https://doi.org/10.1037/xlm0000302

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

Oberauer, K., & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology: General*, *142*(2), 380–411. https://doi.org/10.1037/a0029588

O'Donnell, R. E., Clement, A., & Brockmole, J. R. (2018). Semantic and functional relationships among objects increase the capacity of visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(7), 1151–1158. https://doi.org/10.1037/xlm0000508

Oh, B.-I., Kim, Y.-J., & Kang, M.-S. (2019). Ensemble representations reveal distinct neural coding of visual working memory. *Nature Communications*, *10*(1), 5665. https://doi.org/10.1038/s41467-019-13592-6

Olkkonen, M., & Allred, S. R. (2014). Short-Term Memory Affects Color Perception in Context. *PLOS ONE*, *9*(1), e86488. https://doi.org/10.1371/journal.pone.0086488

Olkkonen, M., McCarthy, P. F., & Allred, S. R. (2014). The central tendency bias in color perception: Effects of internal and external noise. *Journal of Vision*, *14*(11), 5. https://doi.org/10.1167/14.11.5

Olson, I. R. (2006). Working Memory for Conjunctions Relies on the Medial Temporal Lobe. *Journal of Neuroscience*, *26*(17), 4596–4601. https://doi.org/10.1523/JNEUROSCI.1923-05.2006

Olson, I. R., & Jiang, Y. (2004). Visual short-term memory is not improved by training. *Memory & Cognition*, *32*(8), 1326–1332. https://doi.org/10.3758/BF03206323

Olson, I. R., Jiang, Y., & Moore, K. S. (2005). Associative learning improves visual working memory performance. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 889–900. https://doi.org/10.1037/0096-1523.31.5.889

Orban, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750. https://doi.org/10.1073/pnas.0708424105

O'Reilly, R. C., & Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, *18*(2), 283–328. https://doi.org/10.1162/089976606775093909

Orhan, A. E., & Jacobs, R. A. (2014). Toward ecologically realistic theories in visual short-term memory research. *Attention, Perception, & Psychophysics*, *76*(7), 2158–2170. https://doi.org/10.3758/s13414-014-0649-8

Orhan, A. E., Sims, C. R., Jacobs, R. A., & Knill, D. C. (2014). The Adaptive Nature of Visual Working Memory. *Current Directions in Psychological Science*, *23*(3), 164–170. https://doi.org/10.1177/0963721414529144

Öztekin, I., McElree, B., Staresina, B. P., & Davachi, L. (2009). Working Memory Retrieval: Contributions of the Left Prefrontal Cortex, the Left Posterior Parietal Cortex, and the Hippocampus. *Journal of Cognitive Neuroscience*, *21*(3), 581–593. https://doi.org/10.1162/jocn.2008.21016

Palmer, J., Boston, B., & Moore, C. M. (2015). Limited capacity for memory tasks with multiple features within a single object. *Attention, Perception, & Psychophysics*, *77*(5), 1488–1499. https://doi.org/10.3758/s13414-015-0909-2

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601–605. https://doi.org/10.1038/s41586-021-03390-w

Panichello, M. F., DePasquale, B., Pillow, J. W., & Buschman, T. J. (2019). Error-correcting dynamics in visual working memory. *Nature communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-11298-3

Parker, A., Derrington, A., Blakemore, C., Rugg, M. D., Otten, L. J., & Henson, R. N. A. (2002). The neural basis of episodic memory: Evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1424), 1097–1110. https://doi.org/10.1098/rstb.2002.1102

Parra, M. A., Della Sala, S., Logie, R. H., & Morcom, A. M. (2014). Neural correlates of shape–color binding in visual working memory. *Neuropsychologia*, *52*, 27–36. https://doi.org/10.1016/j.neuropsychologia.2013.09.036

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97–107. https://doi.org/10.1038/nrn1603

Pasupathy, A., Popovkina, D. V., & Kim, T. (2020). Visual Functions of Primate Area V4. *Annual Review of Vision Science*, *6*(1), 363–385. https://doi.org/10.1146/annurev-vision-030320-041306

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987. https://doi.org/10.1038/nrn2277

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pessoa, L., Gutierrez, E., Bandettini, P. A., & Ungerleider, L. G. (2002). Neural correlates of visual working memory: fMRI amplitude predicts task performance. *Neuron*, *35*(5), 975–987.

Peters, B., Kaiser, J., Rahm, B., & Bledowski, C. (2015). Activity in Human Visual and Parietal Cortex Reveals Object-Based Attention in Working Memory. *Journal of Neuroscience*, *35*(8), 3360–3369. https://doi.org/10.1523/JNEUROSCI.3795-14.2015

Peterson, D. J., & Berryhill, M. E. (2013). The Gestalt principle of similarity benefits visual working memory. *Psychonomic bulletin & review*, *20*(6), 1282–1289. https://doi.org/10.3758/s13423-013-0460-x

Piekema, C., Rijpkema, M., Fernández, G., & Kessels, R. P. C. (2010). Dissociating the Neural Correlates of Intra-Item and Inter-Item Working-Memory Binding (A. Aleman, Ed.). *PLoS ONE*, *5*(4), e10214. https://doi.org/10.1371/journal.pone.0010214

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2007). Anterior temporal lobes mediate semantic representation: Mimicking semantic dementia by using rTMS in normal participants. *Proceedings of the National Academy of Sciences*, *104*(50), 20137–20141.

Quirk, C., Adam, K. C., & Vogel, E. K. (2020). No evidence for an object working memory capacity benefit with extended viewing time. *Eneuro*, *7*(5). https://doi.org/10.1523/ENEURO.0150-20.2020

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not 'Grandmother-cell' coding in the medial temporal lobe. *Trends in Cognitive Sciences*, *12*(3), 87–91. https://doi.org/10.1016/j.tics.2007.12.003

Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, *13*(8), 587–597. https://doi.org/10.1038/nrn3251

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. https://www.R-project.org/

Rac-Lubashevsky, R., Cremer, A., Collins, A. G., Frank, M. J., & Schwabe, L. (2023). Neural Index of Reinforcement Learning Predicts Improved Stimulus–Response Retention under High Working Memory Load. *The Journal of Neuroscience*, *43*(17), 3131–3143. https://doi.org/10.1523/JNEUROSCI.1274-22.2023

Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature Neuroscience*, *22*(8), 1336–1344. https://doi.org/10.1038/s41593-019-0428-x

Rademaker, R. L., Park, Y. E., Sack, A. T., & Tong, F. (2018). Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 925. https://doi.org/10.1037/xhp0000491

Ranganath, C. (2006). Working memory for visual objects: Complementary roles of inferior temporal, medial temporal, and prefrontal cortex. *Neuroscience*, *139*(1), 277–289. https://doi.org/10.1016/j.neuroscience.2005.06.092

Ranganath, C., & Blumenfeld, R. S. (2005). Doubts about double dissociations between short- and long-term memory. *Trends in Cognitive Sciences*, *9*(8), 374–380. https://doi.org/10.1016/j.tics.2005.06.009

Ranganath, C., Cohen, M. X., & Brozinsky, C. J. (2005). Working memory maintenance contributes to long-term memory formation: Neural and behavioral evidence. *Journal of Cognitive Neuroscience*, *17*(7), 994–1010. https://doi.org/10.1162/0898929054475118

Ranganath, C., Cohen, M. X., Dam, C., & D'Esposito, M. (2004). Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval. *Journal of Neuroscience*, *24*(16), 3917–3925.

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. *Classical conditioning II: Current research and theory*, 64–99.

Rhodes, S., & Cowan, N. (2019). Flexible representations in visual working memory and interactions with long-term learning: Commentary on the special issue. *British Journal of Psychology*, *110*(2), 449–460. https://doi.org/10.1111/bjop.12380

Ricker, T. J., Nieuwenstein, M. R., Bayliss, D. M., & Barrouillet, P. (2018). Working memory consolidation: Insights from studies on attention and working memory. *Annals of the New York Academy of Sciences*, *1424*(1), 8–18. https://doi.org/10.1111/nyas.13633

Riggall, A. C., & Postle, B. R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *32*(38), 12990–12998. https://doi.org/10.1523/JNEUROSCI.1892-12.2012

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Riley, M. R., Qi, X.-L., Zhou, X., & Constantinidis, C. (2018). Anterior-posterior gradient of plasticity in primate prefrontal cortex. *Nature Communications*, *9*(1), 3790. https://doi.org/10.1038/s41467-018-06226-w

Rissman, J., & Wagner, A. D. (2012). Distributed Representations in Memory: Insights from Functional Brain Imaging. *Annual Review of Psychology*, *63*(1), 101–128. https://doi.org/10.1146/annurev-psych-120710-100344

Robinson, M. M., & Brady, T. F. (2023). A quantitative model of ensemble perception as summed activation in feature space. *Nature Human Behaviour*, *7*(10), 1638–1651. https://doi.org/10.1038/s41562-023-01602-z

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation. *Psychological Review*, *111*(1), 205–235. https://doi.org/10.1037/0033-295X.111.1.205

Rolls, E. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience*, *7*. https://doi.org/10.3389/fnsys.2013.00074

Rolls, E. T. (2010). A computational theory of episodic memory formation in the hippocampus. *Behavioural Brain Research*, *215*(2), 180–196. https://doi.org/10.1016/j.bbr.2010.03.027

Roux, F., & Uhlhaas, P. J. (2014). Working memory and neural oscillations: Alpha–gamma versus theta–gamma codes for distinct WM information? *Trends in Cognitive Sciences*, *18*(1), 16–25. https://doi.org/10.1016/j.tics.2013.10.010

Roy, Y., & Faubert, J. (2023). Is the Contralateral Delay Activity (CDA) a robust neural correlate for Visual Working Memory (VWM) tasks? A reproducibility study. *Psychophysiology*, *60*(2), e14180. https://doi.org/10.1111/psyp.14180

Rugg, M. D. (1995). Event-related potential studies of human memory. In *The cognitive neurosciences* (pp. 789–801). The MIT Press.

Saito, J. M., Duncan, K., & Fukuda, K. (2023). Comparing visual memories to similar visual inputs risks lasting memory distortion. *Journal of Experimental Psychology: General*, *152*(8), 2318–2330. https://doi.org/10.1037/xge0001400

Salahub, C., Lockhart, H. A., Dube, B., Al-Aidroos, N., & Emrich, S. M. (2019). Electrophysiological correlates of the flexible allocation of visual working memory resources. *Scientific Reports*, *9*(1), 19428. https://doi.org/10.1038/s41598-019-55948-4

Salazar, R. F., Dotson, N. M., Bressler, S. L., & Gray, C. M. (2012). Content-Specific Fronto-Parietal Synchronization During Visual Working Memory. *Science*, *338*(6110), 1097–1100. https://doi.org/10.1126/science.1224000

Sauseng, P., Klimesch, W., Heise, K. F., Gruber, W. R., Holz, E., Karim, A. A., Glennon, M., Gerloff, C., Birbaumer, N., & Hummel, F. C. (2009). Brain Oscillatory Substrates of Visual Short-Term Memory Capacity. *Current Biology*, *19*(21), 1846–1852. https://doi.org/10.1016/j.cub.2009.08.062

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546. https://doi.org/10.1073/pnas.1820226116

Schacter, D. L., Guerin, S. A., & Jacques, P. L. S. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences*, *15*(10), 467–474. https://doi.org/10.1016/j.tics.2011.08.004

Schall, J., Morel, A., King, D., & Bullier, J. (1995). Topography of visual cortex connections with frontal eye field in macaque: Convergence and segregation of processing streams. *The Journal of Neuroscience*, *15*(6), 4464–4487. https://doi.org/10.1523/JNEUROSCI.15-06-04464.1995

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The Necessity of the Medial Temporal Lobe for Statistical Learning. *Journal of Cognitive Neuroscience*, *26*(8), 1736–1747. https://doi.org/10.1162/jocn_a_00578

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8. https://doi.org/10.1002/hipo.22523

Schiffer, A.-M., Muller, T., Yeung, N., & Waszak, F. (2014). Reward Activates Stimulus-Specific and Task-Dependent Representations in Visual Association Cortices. *The Journal of Neuroscience*, *34*(47), 15610–15620. https://doi.org/10.1523/JNEUROSCI.1640-14.2014

Schneegans, S., & Bays, P. M. (2017). Neural Architecture for Feature Binding in Visual Working Memory. *The Journal of Neuroscience*, *37*(14), 3913–3925. https://doi.org/10.1523/JNEUROSCI.3493-16.2017

Schneegans, S., & Bays, P. M. (2018). Drift in Neural Population Activity Causes Working Memory to Deteriorate Over Time. *The Journal of Neuroscience*, *38*(21), 4859–4869. https://doi.org/10.1523/JNEUROSCI.3440-17.2018

Schneegans, S., & Bays, P. M. (2019). New perspectives on binding in visual working memory. *British Journal of Psychology*, *110*(2), 207–244. https://doi.org/10.1111/bjop.12345

Schon, K., Newmark, R. E., Ross, R. S., & Stern, C. E. (2016). A Working Memory Buffer in Parahippocampal Regions: Evidence from a Load Effect during the Delay Period. *Cerebral Cortex*, *26*(5), 1965–1974. https://doi.org/10.1093/cercor/bhv013

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schurgin, M. W. (2018). Visual memory, the long and the short of it: A review of visual working memory and long-term memory. *Attention, Perception, & Psychophysics*, *80*(5), 1035–1056. https://doi.org/10.3758/s13414-018-1522-y

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Shallice, T., & Warrington, E. K. (1970). Independent Functioning of Verbal Memory Stores: A Neuropsychological Study. *Quarterly Journal of Experimental Psychology*, *22*(2), 261–273. https://doi.org/10.1080/00335557043000203

Shibata, K., Klar, V., Fallon, S. J., Husain, M., & Manohar, S. G. (2024). Working memory as a representational template for reinforcement learning. *Scientific Reports*, *14*(1), 27660. https://doi.org/10.1038/s41598-024-79119-2

Shibata, K., Klar, V., Husain, M., & Manohar, S. (Unpublished manuscript). When reinforcement learning requires working memory.

Shiffrin, R. M., & Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review*, *76*(2), 179–193. https://doi.org/10.1037/h0027277

Shimi, A., & Logie, R. H. (2019). Feature binding in short-term memory and long-term learning. *Quarterly Journal of Experimental Psychology*, *72*(6), 1387–1400. https://doi.org/10.1177/1747021818807718

Silva, C. F. da, Yao, Y.-W., & Hare, T. A. (2018, July 26). *Can model-free reinforcement learning operate over information stored in working-memory?* https://doi.org/10.1101/107698

Simons, J. S., & Spiers, H. J. (2003). Prefrontal and medial temporal lobe interactions in long-term memory. *Nature Reviews Neuroscience*, *4*(8), 637–648. https://doi.org/10.1038/nrn1178

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198. https://doi.org/10.1016/j.cognition.2016.03.020

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656. https://doi.org/10.1126/science.aaq1118

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–830. https://doi.org/10.1037/a0029856

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *Journal of Neuroscience*, *37*(27), 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707. https://doi.org/10.1016/j.neuron.2016.07.006

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. https://doi.org/10.1037/0033-295X.99.2.195

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177. https://doi.org/10.1016/j.nlm.2004.06.005

Squire, L. R., Stark, C. E., & Clark, R. E. (2004). The Medial Temporal Lobe. *Annual Review of Neuroscience*, *27*(1), 279–306. https://doi.org/10.1146/annurev.neuro.27.070203.144130

Squire, L. R., & Wixted, J. T. (2011). The Cognitive Neuroscience of Human Memory Since H.M. *Annual Review of Neuroscience*, *34*(1), 259–288. https://doi.org/10.1146/annurev-neuro-061010-113720

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222. https://doi.org/10.1080/14640747308400340

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron*, *78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Storkey, A. (1997). Increasing the capacity of a hopfield network without sacrificing functionality. In W. Gerstner, A. Germond, M. Hasler, & J.-D. Nicoud (Eds.), *Artificial Neural Networks — ICANN'97* (pp. 451–456). Springer. https://doi.org/10.1007/BFb0020196

Stroud, J. P., Duncan, J., & Lengyel, M. (2024). The computational foundations of dynamic coding in working memory. *Trends in Cognitive Sciences*, *0*(0). https://doi.org/10.1016/j.tics.2024.02.011

Strzelczyk, D., Clayson, P. E., Sigurdardottir, H. M., Mushtaq, F., Pavlov, Y. G., Devillez, H., Lukashevich, A., Rocha, H. A., Chung, Y. H., Ortego, K. M., Störmer, V. S., García Alanis, J. C., Löffler, C., Schubert, A.-L., Biel, A. L., Tretow, A., Xu, W., Hamalainen, J., Lu, Z., … Langer, N. (2023, August 31). *Contralateral delay activity as a marker of visual working memory capacity: A multi-site registered replication* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/shdea

Sučević, J., & Schapiro, A. C. (2023). A neural network model of hippocampal contributions to category learning (D. Zeithamova & M. J. Frank, Eds.). *eLife*, *12*, e77185. https://doi.org/10.7554/eLife.77185

Summerfield, C., Greene, M., Wager, T., Egner, T., Hirsch, J., & Mangels, J. (2006). Neocortical Connectivity during Episodic Memory Formation. *PLOS Biology*, *4*(5), e128. https://doi.org/10.1371/journal.pbio.0040128

Sun, W., Advani, M., Spruston, N., Saxe, A., & Fitzgerald, J. E. (2023). Organizing memories for generalization in complementary learning systems. *Nature Neuroscience*, *26*(8), 1438–1448. https://doi.org/10.1038/s41593-023-01382-9

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Swan, G., & Wyble, B. (2014). The binding pool: A model of shared neural resources for distinct items in visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 2136–2157. https://doi.org/10.3758/s13414-014-0633-3

Takahashi, E., Ohki, K., & Kim, D.-S. (2013). Dissociation and convergence of the dorsal and ventral visual working memory streams in the human prefrontal cortex. *NeuroImage*, *65*, 488–498. https://doi.org/10.1016/j.neuroimage.2012.10.002

Tark, K.-J., Kang, M.-S., Chong, S. C., & Shim, W. M. (2021). Neural representations of ensemble coding in the occipital and parietal cortices. *NeuroImage*, *245*, 118680. https://doi.org/10.1016/j.neuroimage.2021.118680

Taylor, R., & Bays, P. M. (2018). Efficient Coding in Visual Working Memory Accounts for Stimulus-Specific Variations in Recall. *Journal of Neuroscience*, *38*(32), 7132–7142. https://doi.org/10.1523/JNEUROSCI.1018-18.2018

Tehovnik, E. J., Sommer, M. A., Chou, I.-H., Slocum, W. M., & Schiller, P. H. (2000). Eye fields in the frontal lobes of primates. *Brain Research Reviews*, *32*(2), 413–448. https://doi.org/10.1016/S0165-0173(99)00092-2

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 37–55. https://doi.org/10.1037/xlm0000578

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*(6984), 751–754. https://doi.org/10.1038/nature02466

Toniolo, S., Zhao, S., Scholcz, A., Amein, B., Ganse-Dumrath, A., Heslegrave, A. J., Thompson, S., Manohar, S., Zetterberg, H., & Husain, M. (2024). Relationship of plasma biomarkers to digital cognitive tests in Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *16*(2), e12590. https://doi.org/10.1002/dad2.12590

Torres, R. E., Duprey, M., Molokwu, N., Campbell, K. L., & Emrich, S. M. (2023, March 3). *Not all objects are created equal: Greater visual working memory for real-world objects is related to item memorability* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/v2ta5

Trepka, E., Spitmaan, M., Qi, X.-L., Constantinidis, C., & Soltani, A. (2023). Training-dependent gradients of timescales of neural dynamics in the primate prefrontal cortex and their contributions to working memory. *The Journal of Neuroscience*, *44*(2), e2442212023. https://doi.org/10.1523/JNEUROSCI.2442-21.2023

Tulving, E. (1986). Episodic and semantic memory: Where should we go from here? *Behavioral and Brain Sciences*, *9*(3), 573–577. https://doi.org/10.1017/S0140525X00047257

Tulving, E. (1993). What Is Episodic Memory? *Current Directions in Psychological Science*, *2*(3), 67–70. https://doi.org/10.1111/1467-8721.ep10770899

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, *53*, 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Uchimura, M., Panichello, M. F., & Buschman, T. J. (Unpublished manuscript). Context-dependent changes in neural representation of working memory.

Udale, R., Tran, M. T., Manohar, S., & Husain, M. (2022). Dynamic in-flight shifts of working memory resources across saccades. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(1), 21. https://doi.org/10.1037/xhp0000960

Urgolites, Z. J., Wixted, J. T., Goldinger, S. D., Papesh, M. H., Treiman, D. M., Squire, L. R., & Steinmetz, P. N. (2022). Two kinds of memory signals in neurons of the human hippocampus. *Proceedings of the National Academy of Sciences*, *119*(19), e2115128119. https://doi.org/10.1073/pnas.2115128119

Utochkin, I. S., & Brady, T. F. (2020). Individual representations in visual working memory inherit ensemble properties. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(5), 458. https://doi.org/10.1037/xhp0000727

Van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, *7*, e34963. https://doi.org/10.7554/eLife.34963

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124–149. https://doi.org/10.1037/a0035234

Vergauwe, E., & Cowan, N. (2015). Working memory units are all in your head: Factors that influence whether features or objects are the favored units. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1404–1416. https://doi.org/10.1037/xlm0000108

Verschooren, S., Kessler, Y., & Egner, T. (2021). Evidence for a single mechanism gating perceptual and long-term memory information into working memory. *Cognition*, *212*, 104668. https://doi.org/10.1016/j.cognition.2021.104668

Visser, M., Jefferies, E., & Lambon Ralph, M. A. (2009). Semantic Processing in the Anterior Temporal Lobes: A Meta-analysis of the Functional Neuroimaging Literature. *Journal of Cognitive Neuroscience*, *22*(6), 1083–1094. https://doi.org/10.1162/jocn.2009.21309

Vo, V. A., Sutterer, D. W., Foster, J. J., Sprague, T. C., Awh, E., & Serences, J. T. (2022). Shared Representational Formats for Information Maintained in Working Memory and Information Retrieved from Long-Term Memory. *Cerebral Cortex*, *32*(5), 1077–1092. https://doi.org/10.1093/cercor/bhab267

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. https://doi.org/10.1038/nature02447

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of experimental psychology: human perception and performance*, *27*(1), 92. https://doi.org/10.1037//0096-1523.27.1.92

Voytek, B., & Knight, R. T. (2010). Prefrontal cortex and basal ganglia contributions to visual working memory. *Proceedings of the National Academy of Sciences*, *107*(42), 18167–18172.

Wagner, A. D., Shannon, B. J., Kahn, I., & Buckner, R. L. (2005). Parietal lobe contributions to episodic memory retrieval. *Trends in Cognitive Sciences*, *9*(9), 445–453. https://doi.org/10.1016/j.tics.2005.07.001

Wallis, G., Stokes, M. G., Arnold, C., & Nobre, A. C. (2015). Reward boosts working memory encoding over a brief temporal window. *Visual Cognition*, *23*(1–2), 291–312. https://doi.org/10.1080/13506285.2015.1013168

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, *27*(4), 635–657. https://doi.org/10.1080/14640747508400525

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829–853. https://doi.org/10.1093/brain/107.3.829

Wechsler, D. (1939). *The measurement of adult intelligence*. Williams & Wilkins Co. https://doi.org/10.1037/10020-000

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

Whittington, J. C. R., Warren, J., & Behrens, T. E. J. (2021). Relating transformers to models and neural representations of the hippocampal formation.

Wimmer, G. E., & Poldrack, R. A. (2021, August 19). *Reward learning and working memory: Effects of massed versus spaced training and post-learning delay period*. https://doi.org/10.1101/2020.03.19.997098

Wixted, J. T., Goldinger, S. D., Squire, L. R., Kuhn, J. R., Papesh, M. H., Smith, K. A., Treiman, D. M., & Steinmetz, P. N. (2018). Coding of episodic memory in the human hippocampus. *Proceedings of the National Academy of Sciences*, *115*(5), 1093–1098. https://doi.org/10.1073/pnas.1716443115

Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2020). Drifting codes within a stable coding scheme for working memory. *PLOS Biology*, *18*(3), e3000625. https://doi.org/10.1371/journal.pbio.3000625

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871. https://doi.org/10.1038/nn.4546

Xie, W., & Zhang, W. (2018). Familiarity Speeds Up Visual Short-term Memory Consolidation: Electrophysiological Evidence from Contralateral Delay Activities. *Journal of Cognitive Neuroscience*, *30*(1), 1–13. https://doi.org/10.1162/jocn_a_01188

Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.-J., Yang, T., Dehaene, S., Tang, S., Min, B., & Wang, L. (2022). Geometry of sequence working memory in macaque prefrontal cortex. *Science*, *375*(6581), 632–639. https://doi.org/10.1126/science.abm0204

Xu, Y. (2002). Encoding color and shape from different parts of an object in visual short-term memory. *Perception & Psychophysics*, *64*(8), 1260–1280. https://doi.org/10.3758/BF03194770

Xu, Y. (2017). Reevaluating the Sensory Account of Visual Working Memory Storage. *Trends in Cognitive Sciences*, *21*(10), 794–815. https://doi.org/10.1016/j.tics.2017.06.013

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91–95. https://doi.org/10.1038/nature04262

Xue, G. (2018). The Neural Representations Underlying Human Episodic Memory. *Trends in Cognitive Sciences*, *22*(6), 544–561. https://doi.org/10.1016/j.tics.2018.03.004

Yonelinas, A. (2001). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1363–1374. https://doi.org/10.1098/rstb.2001.0939

Yonelinas, A. P. (2013). The hippocampus supports high-resolution binding in the service of perception, working memory and long-term memory. *Behavioural Brain Research*, *254*, 34–44. https://doi.org/10.1016/j.bbr.2013.05.030

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the Brain Regions Involved in Recollection and Familiarity in Recognition Memory. *Journal of Neuroscience*, *25*(11), 3002–3008. https://doi.org/10.1523/JNEUROSCI.5295-04.2005

Yoo, A. H., & Collins, A. G. E. (2022). How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, *34*(4), 551–568. https://doi.org/10.1162/jocn_a_01808

Yoo, A. H., Klyszejko, Z., Curtis, C. E., & Ma, W. J. (2018). Strategic allocation of working memory resource. *Scientific Reports*, *8*(1), 16162. https://doi.org/10.1038/s41598-018-34282-1

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Zimmer, H. D., Popp, C., Reith, W., & Krick, C. (2012). Gains of item-specific training in visual working memory and their neural correlates. *Brain Research*, *1466*, 44–55. https://doi.org/10.1016/j.brainres.2012.05.019

Zipser, D., Kehoe, B., Littlewort, G., & Fuster, J. (1993). A spiking network model of short-term active memory. *Journal of Neuroscience*, *13*(8), 3406–3420. https://doi.org/10.1523/JNEUROSCI.13-08-03406.1993

Zipser, D. (1991). Recurrent Network Model of the Neural Mechanism of Short-Term Active Memory. *Neural Computation*, *3*(2), 179–193. https://doi.org/10.1162/neco.1991.3.2.179

Zokaei, N., Manohar, S., Husain, M., & Feredoes, E. (2014). Causal Evidence for a Privileged Working Memory State in Early Visual Cortex. *Journal of Neuroscience*, *34*(1), 158–162. https://doi.org/10.1523/JNEUROSCI.2899-13.2014

Zokaei, N., Ning, S., Manohar, S., Feredoes, E., & Husain, M. (2014). Flexibility of representational states in working memory. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00853

Appendix

7.4 Chapter 3

7.4.1 Tables

Table S1
Exp. 1 Effect of retention interval on recognition accuracy

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	3.00	2367.86	2386.39	–1180.93	2361.86			
Model 2	4.00	2354.62	2379.32	–1173.31	2346.62	15.24	1.00	0.00
Model 3	5.00	2353.72	2384.60	–1171.86	2343.72	2.90	1.00	0.09

Winning Model

	Est.	Std. Error	z-value	p-value
(Intercept)	3.196	0.221	14.469	0.000
Retention interval	–0.022	0.007	–3.402	0.001
t2	–0.005	0.001	–3.934	0.000

Table S2
Exp. 1 Effect of retention interval on colour memory error

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	4.00	39058.47	39083.33	–19525.24	39050.47			
Model 2	5.00	39060.19	39091.26	–19525.10	39050.19	0.28	1.00	0.60
Model 3	6.00	39060.40	39097.68	–19524.20	39048.40	1.79	1.00	0.18

Winning Model

	Est.	Std. Error	df	t-value	p-value
(Intercept)	43.720	2.427	138.411	18.017	0.000
Retention interval	0.478	0.091	3649.439	5.248	0.000

Table S3
Exp. 1 Effect of interference on recognition accuracy

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	5.00	4421.91	4456.23	–2205.96	4411.91			
Model 2	6.00	4423.36	4464.54	–2205.68	4411.36	0.55	1.00	0.46
Model 3	9.00	4428.27	4490.03	–2205.13	4410.27	1.10	3.00	0.78

Winning Model

	Est.	Std. Error	z-value	p-value
(Intercept)	2.544	0.172	14.812	0.000
Interference — Filler	–0.049	0.127	–0.387	0.698
Interference — Lure	–0.176	0.119	–1.475	0.140
Interference — Target	–0.107	0.129	–0.833	0.405

Table S4
Exp. 1 Effect of interference on colour memory error

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	6.00	77637.99	77679.41	–38812.99	77625.99			
Model 2	7.00	77621.08	77669.40	–38803.54	77607.08	18.90	1.00	0.00
Model 3	10.00	77624.28	77693.31	–38802.14	77604.28	2.80	3.00	0.42

Winning Model

	Est.	Std. Error	df	t-value	p-value
(Intercept)	44.112	2.630	175.420	16.771	0.000
Interference — Filler	1.073	1.689	7305.429	0.636	0.525
Interference — Lure	2.902	1.621	7304.367	1.790	0.073
Interference — Target	3.625	1.732	7313.931	2.093	0.036
No. of trials	2.186	0.503	7320.058	4.349	0.000

Table S5

Exp. 1 Effect of number of items in memory on recognition accuracy

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	3.00	1334.85	1351.74	–664.43	1328.85			
Model 2	4.00	1327.52	1350.04	–659.76	1319.52	9.33	1.00	0.00
Model 3	5.00	1325.57	1353.71	–657.78	1315.57	3.95	1.00	0.05

Winning Model

	Est.	Std. Error	z-value	p-value
(Intercept)	4.891	1.004	4.872	0.000
No. of items in memory	–0.228	0.122	–1.862	0.063
Retention interval	–1.606	0.595	–2.699	0.007
No. of items in memory:Retention interval	0.149	0.074	2.007	0.045

Table S6

Exp. 1 Effect of number of items in memory on colour memory error

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	4.00	23167.15	23189.92	–11579.58	23159.15			
Model 2	5.00	23154.04	23182.50	–11572.02	23144.04	15.11	1.00	0.00
Model 3	6.00	23155.88	23190.04	–11571.94	23143.88	0.15	1.00	0.70

Winning Model

	Est.	Std. Error	df	t-value	p-value
(Intercept)	36.621	5.932	693.499	6.174	0.000
No. of items in memory	0.183	0.575	1910.779	0.318	0.750
Retention interval	7.796	2.003	2148.733	3.892	0.000

Table S7

Exp. 2 Effect of retention interval on recognition accuracy

Model Comparison

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
Model 1	3.00	889.26	905.21	–441.63	883.26			
Model 2	4.00	887.65	908.92	–439.83	879.65	3.61	1.00	0.06
Model 3	5.00	889.25	915.84	–439.63	879.25	0.40	1.00	0.53

Winning Model

	Est.	Std. Error	z-value	p-value
(Intercept)	3.590	0.319	11.243	0.000
Retention interval	–0.061	0.011	–5.371	0.000

Table S8

Exp. 2 Effect of retention interval on colour memory error

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	4.00	16407.93	16429.21	–8199.96	16399.93			
Model 2	5.00	16406.73	16433.32	–8198.36	16396.73	3.20	1.00	0.07
Model 3	6.00	16407.81	16439.73	–8197.91	16395.81	0.91	1.00	0.34

Winning Model

	Est.	Std. Error	df	t-value	p-value
(Intercept)	44.770	3.975	77.589	11.263	0.000
Retention interval	0.682	0.167	1493.081	4.083	0.000

Table S9

Exp. 2 Effect of interference on recognition accuracy

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	5.00	1675.00	1705.17	–832.50	1665.00			
Model 2	6.00	1655.39	1691.60	–821.70	1643.39	21.60	1.00	0.00
Model 3	9.00	1653.16	1707.48	–817.58	1635.16	8.23	3.00	0.04

Winning Model

	Est.	Std. Error	z-value	p-value
(Intercept)	4.221	0.583	7.239	0.000
Interference — Filler	–0.334	0.662	–0.504	0.614
Interference — Lure	–0.870	0.582	–1.495	0.135
Interference — Target	–1.437	0.620	–2.319	0.020
No. of trials	–0.415	0.166	–2.506	0.012
Interference — Filler:n	0.023	0.212	0.108	0.914
Interference — Lure:n	0.035	0.193	0.182	0.856
Interference — Target:n	0.433	0.205	2.110	0.035

Table S10

Exp. 2 Effect of interference on colour memory error

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	6.00	33641.27	33677.49	–16814.63	33629.27			
Model 2	7.00	33630.95	33673.21	–16808.47	33616.95	12.32	1.00	0.00
Model 3	10.00	33633.32	33693.70	–16806.66	33613.32	3.62	3.00	0.31

Winning Model

	Est.	Std. Error	df	t-value	p-value
(Intercept)	42.910	4.385	93.745	9.784	0.000
Interference — Filler	5.321	3.106	3075.476	1.713	0.087
Interference — Lure	9.599	2.923	3072.292	3.284	0.001
Interference — Target	3.260	3.143	3081.529	1.037	0.300
No. of trials	3.187	0.908	3075.392	3.510	0.000

Table S11

Exp. 2 Effect of number of items in memory on recognition accuracy

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	3.00	444.95	459.00	–219.48	438.95			
Model 2	4.00	429.32	448.05	–210.66	421.32	17.63	1.00	0.00
Model 3	5.00	431.31	454.72	–210.65	421.31	0.01	1.00	0.91

Winning Model

	Est.	Std. Error	z-value	p-value
(Intercept)	4.644	0.811	5.725	0.000
No. of items in memory	–0.008	0.073	–0.110	0.912
Retention interval	–1.158	0.285	–4.067	0.000

Table S12

Exp. 2 Effect of number of items in memory on colour memory error

Model Comparison

	N param.	AIC	BIC	logLik	Deviance	χ^2	Df	p-value
Model 1	4.00	8734.23	8752.98	–4363.11	8726.23			
Model 2	5.00	8729.53	8752.97	–4359.76	8719.53	6.70	1.00	0.01
Model 3	6.00	8730.75	8758.88	–4359.37	8718.75	0.78	1.00	0.38

Winning Model

	Est.	Std. Error	df	t-value	p-value
(Intercept)	39.383	10.650	263.760	3.698	0.000
No. of items in memory	0.566	1.051	653.704	0.538	0.591
Retention interval	10.031	3.872	783.679	2.591	0.010

Table S13
Exp. 1 Clustering analysis

No. of clusters	Min. abs. error	Max. abs. error
1	75.27	99.98
2	45.54	74.21
3	40.28	69.64
4	31.67	65.89
5	34.67	64.45
6	32.72	64.38
7	30.09	64.74
8	29.63	64.62

Table S14
Exp. 2 Clustering analysis

No. of clusters	Min. abs. error	Max. abs. error
1	72.12	112.15
2	43.63	80.56
3	39.71	82.75
4	39.47	85.44
5	35.10	81.00
6	33.50	83.40
7	33.77	83.19
8	31.71	82.06

Table S15
List of used images and their semantic categories

Category	N	Example
Arts And Crafts	9	Violin, Wool, Harp
Beverages	6	Beer, Cocktail, Wine
Body Parts	7	Eye, Hand, Nose
Clothes	16	Trousers, T-shirt, Flip-flops
Electronics	14	Computer mouse, Clock, Microphone
Everyday Objects	7	Old key, Toothbrush, Lock
Food	20	Peperoni,Cheese, Peanut
Games + Leisure	10	Dice, Book, Opera masks
Home And Furniture	12	Lamp, House, Shower
Insects + Sea Animals	15	Octopus, Shell, Whale
Mammals + Birds	15	Bird of prey, Cocktail, Bunny
Misc	13	Scales, Anchor, Present
Plants	11	Acorn Leaf, Clove, Flower
Sports	13	Canoe, Table tennis ball, Badminton ball
Tools	10	Paint brush, Wrench, Saw
Transport	13	Sail boat, Car, Helicopter
Universe	9	Rain cloud, Full moon, Sun

7.5 Chapter 4

7.5.1 Tables

Model Comparison

Table S16
Exp. 1a Accuracy

N param.	AIC	BIC	logLik	deviance	χ^2	Df	p-value
4.000	12829.819	12858.931	–6410.909	12821.819			
9.000	12766.201	12831.703	–6374.101	12748.201	73.617	5.000	0.000

Minimal random effects structure: accuracy ∼ condition + (1 | id)

	Est.	Std. Error	z-value	p-value
(Intercept)	0.971	0.093	10.489	0.000
Relevant BL	–0.235	0.053	–4.447	0.000
Irrelevant BL	–0.116	0.054	–2.169	0.030

Full random effects structure: accuracy ∼ condition + (1 + condition | id)

	Est.	Std. Error	z-value	p-value
(Intercept)	0.963	0.084	11.472	0.000
Relevant BL	–0.198	0.105	–1.890	0.059
Irrelevant BL	–0.093	0.074	–1.256	0.209

Table S17
Exp. 1b Accuracy

	Est.	Std. Error	z-value	p-value
(Intercept)	0.931	0.084	11.138	0.000
BL	–0.114	0.055	–2.059	0.040

Table S18
Exp. 2a Accuracy

	Est.	Std. Error	z-value	p-value
(Intercept)	0.951	0.100	9.504	0.000
WM^+ BL	0.010	0.118	0.084	0.933
WM^-	–0.014	0.123	–0.114	0.910
WM^- BL	–0.015	0.130	–0.114	0.910

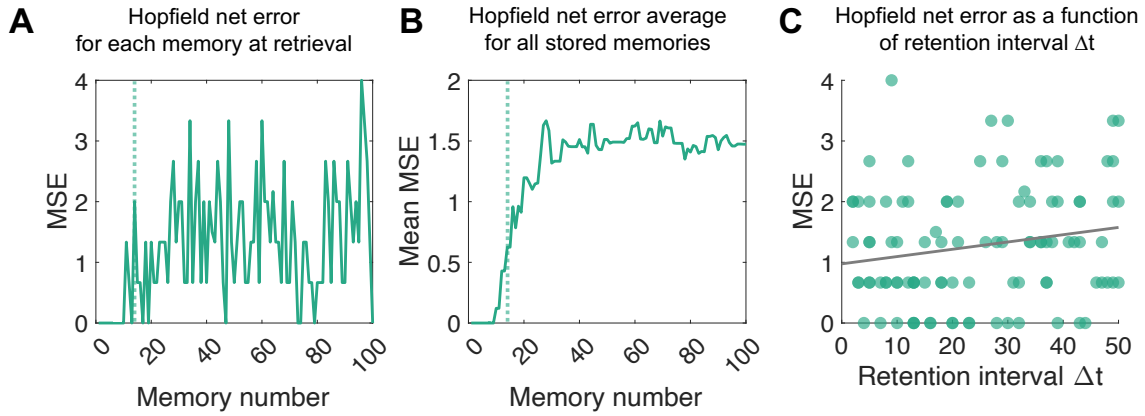
Table S19
Exp. 2b Accuracy

	Est.	Std. Error	z-value	p-value
(Intercept)	1.575	0.088	17.950	0.000
WM^+ BL	0.068	0.076	0.904	0.366
WM^-	–0.078	0.075	–1.036	0.300
WM^- BL	0.024	0.075	0.320	0.749

7.6 Chapter 5

7.6.1 Figures

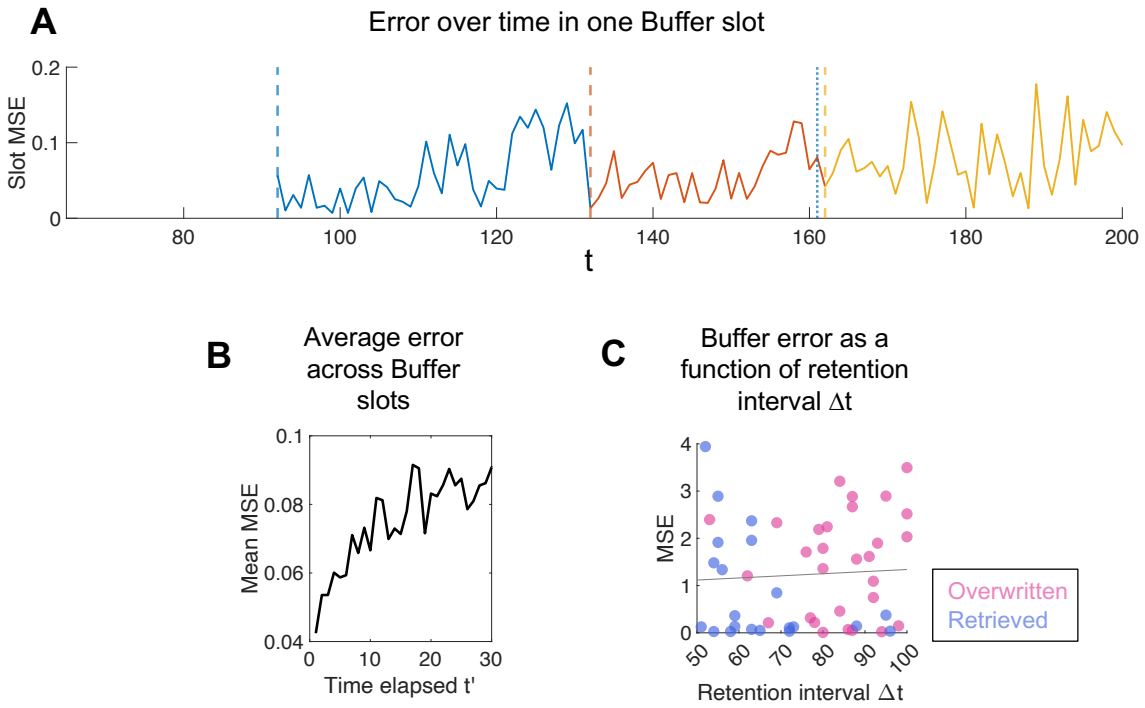
Figure S1
Hopfield model performance decreases rapidly when overloaded with memories



Simulation details: Single trial shown. $N_h = 100$, C ≈ 7.53, $\Delta t \sim$ Uniform [0, 50], $\lambda = 10$, set size = 100. **(A)** Mean-squared error in the Hopfield model as a function of number of patterns stored. MSE was evaluated for each memory at its retrieval time t_2. **(B)** Average mean-squared error in the Hopfield model across previous memories as a function of number of memories stored. **(C)** Hopfield MSE as a function of retention time Δt.

Figure S2

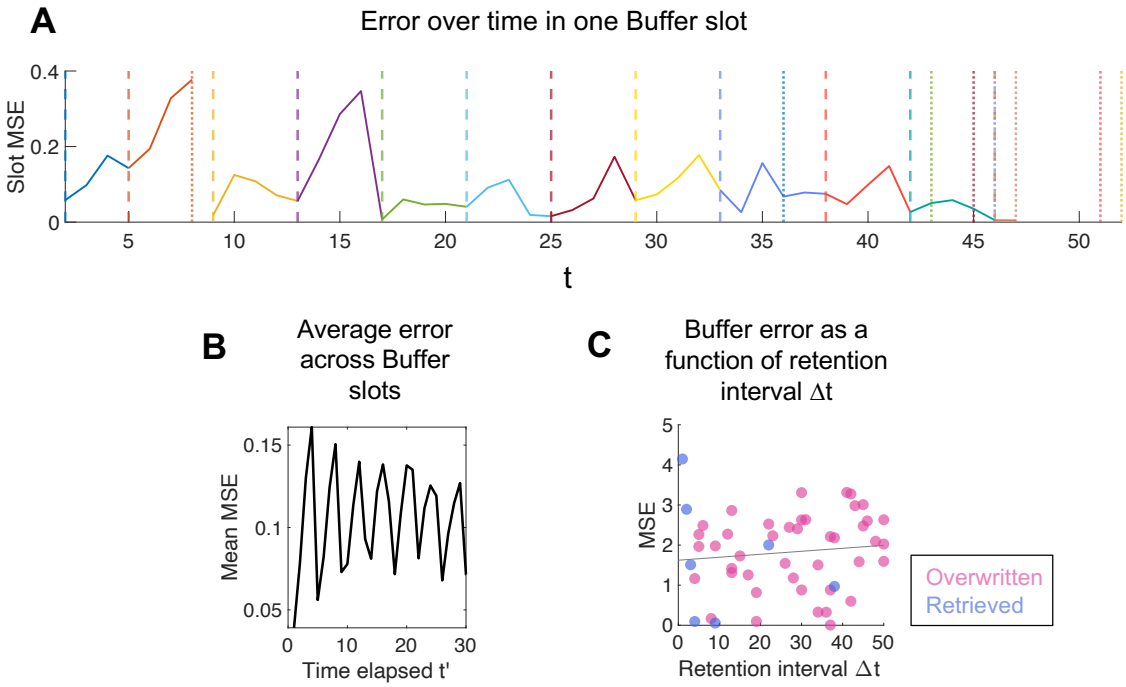
Buffer performance decreases when the retention interval was long



Simulation details: Single trial shown. $N_b = 100$, $\Delta t \sim$ Uniform [50, 100], $\lambda = 16$, set size $= 50$. **(A)** MSE in one buffer slot evaluated at every time step after encoding. Dashed lines represent a new memory being encoded, dotted lines represent a memory being retrieved. Colours indicate different memories. Most memories are forgotten before they are retrieved (e.g. the blue memory). Although new memories arrived very slowly, old ones could not be retrieved before a new memory arrived since they were needed at a much later time point. **(B)** MSE in the buffer for all slots as a function of time elapsed after encoding (t'). With memories arriving slowly, individual memories accumulated noise over time, leading to higher error. **(C)** Buffer MSE as a function of retention time Δt. Since the arrival rate was high, most memories were overwritten even at short retention intervals, driving up the overall error. Error was already high for retrieved memories since time elapsed $t' \geq 50$, and even longer memories were overwritten.

Figure S3

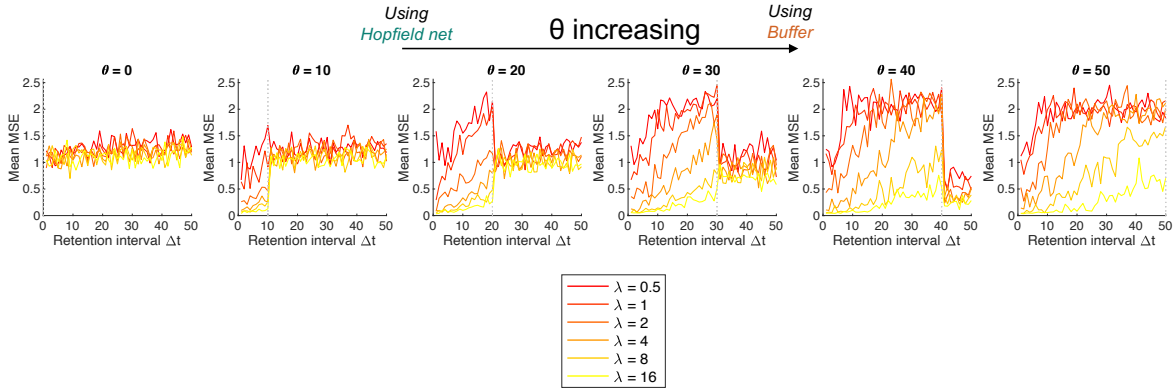
Buffer performance decreases rapidly when memories arrive in short succession



Simulation details: Single trial shown. $N_b = 100$, $\Delta t \sim$ Uniform $[0, 50]$, $\lambda = 0.25$, set size $= 50$. **(A)** MSE in one buffer slot evaluated at every time step after encoding. Dashed lines represent a new memory being encoded, dotted lines represent a memory being retrieved. Colours indicate different memories. Most memories were forgotten before they are retrieved (e.g. the blue memory). Memories arrived at such a high rate, few old ones could be retrieved without being interfered with. **(B)** MSE in the buffer for all slots as a function of time elapsed after encoding (t'). Error was independent of the time elapsed t' since memories could be overwritten at any time point after encoding. **(C)** Buffer MSE as a function of retention time Δt. Since the arrival rate was extremely high, most memories were overwritten even at short retention intervals, driving up the overall error.

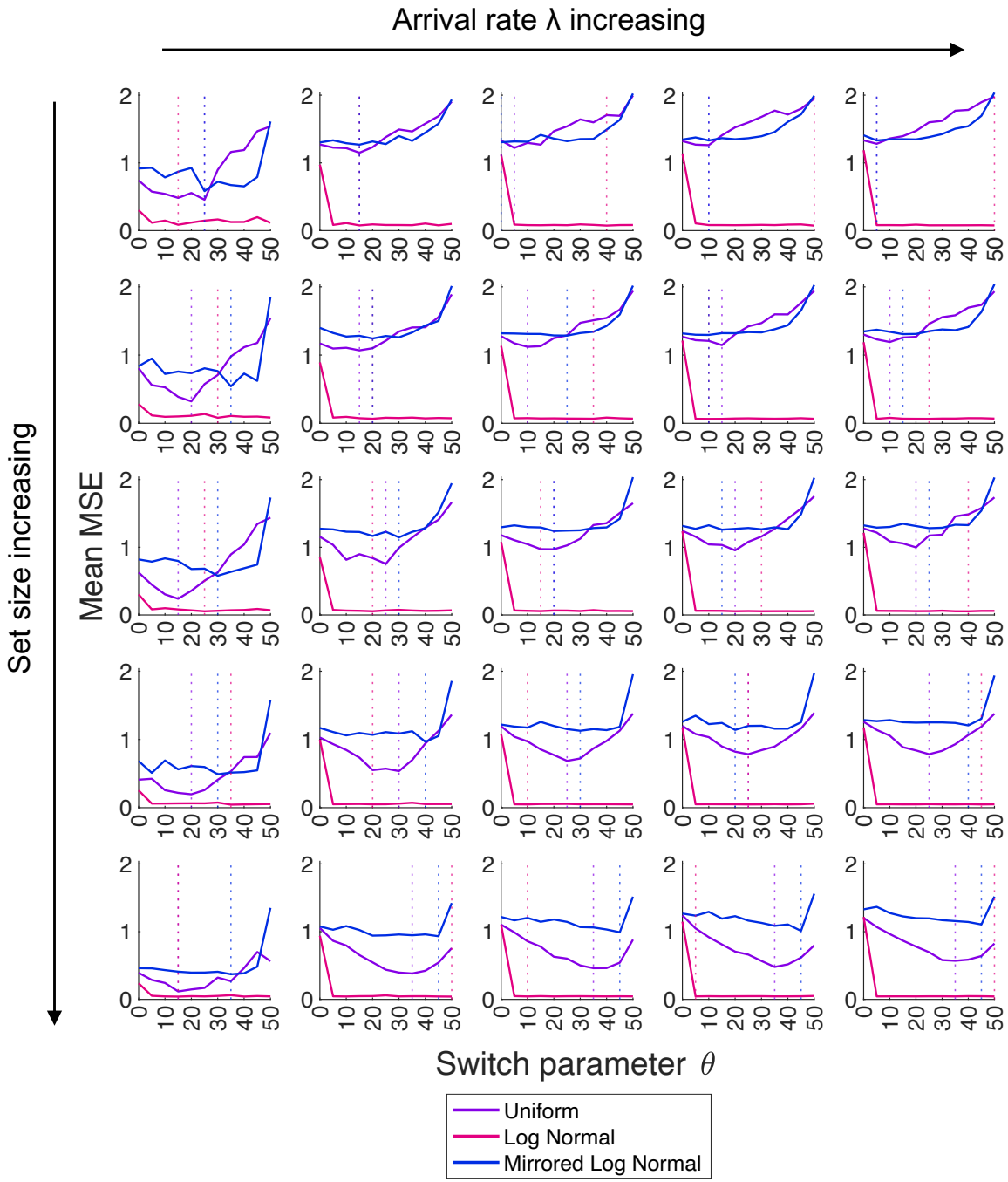
Figure S4

Testing different environments



Simulation details: 100 trials shown. $N_h = N_b = 100$, $\Delta t \sim$ Uniform [0, 50], $\lambda = 0.5, 1, 2, 4, 8, 16$, $\theta = 0, 10, 20, 30, 40, 50$, set size = 100. MSE as a function of arrival rate λ. Overall error decreases as λ increases. When distributing memories in a balanced way between both systems (switch parameter $\theta = 30$), it was evident that error increases more rapidly in the buffer the faster memories arrive, whereas the Hopfield model was less affected. As θ increased (and the Buffer encodes more memories), the overall error was more dependent on λ.

Figure S5
Testing different environments

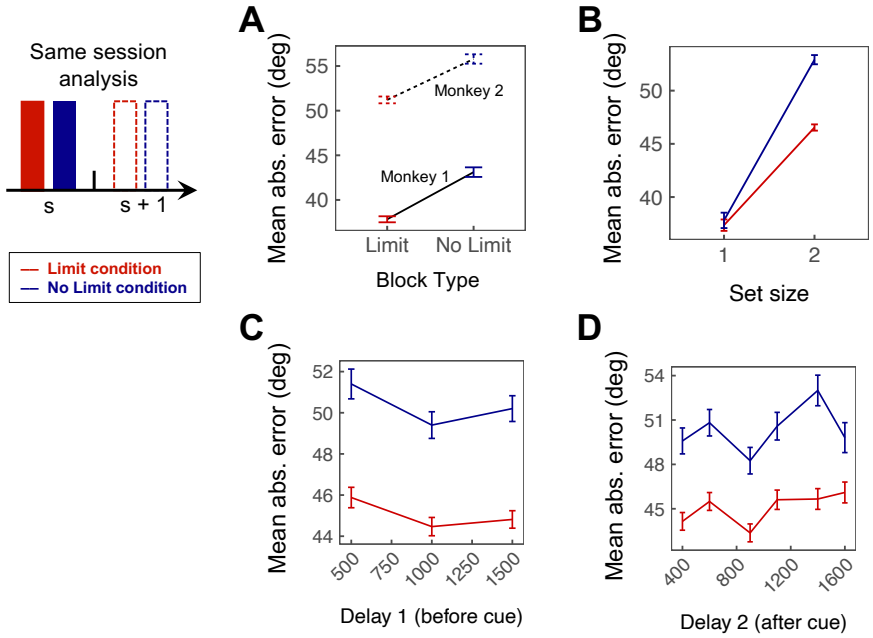


Simulation details: 100 trials shown. $N_h = N_b = 100$, $\Delta t \sim$ Uniform [0, 50], Lognormal $\sim \mu = 1, \sigma = 2, \lambda = 0.5, 1, 2, 4, 8, 16$, $\theta = 0, 10, 20, 30, 40, 50$, set size = 100. Average MSE as a function of switch parameter θ, tested environments, arrival rate λ, and set size (No. of memories in one trial). As λ increased, errors for the slow and uniform environment increased, whereas errors for the bursty environment stayed low, except for the case where only the buffer was used. As set size increases, error increases specifically for the buffer, and for faster environments (low λ).

7.7 Chapter 6

7.7.1 Figures

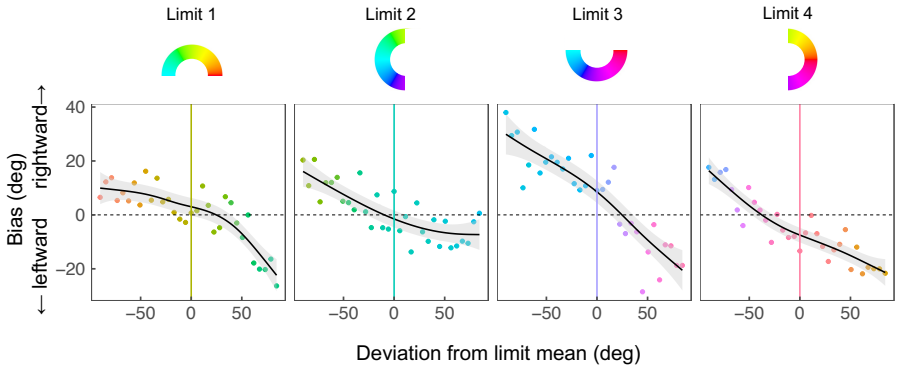
Figure S6
Same Session: Effects of Set Size and Delay



(A) Monkey 1 performed better than Monkey 2, but Block Type effects were comparable. **(B)** A Block Type effect was only observed at Set Size 2. **(C)** Longer delays between stimuli and retro-cue improved performance slightly, but did not change between Block Types. **(D)** Longer delays between retro-cue and report slightly worsened performance, but this did not change between Block Types.

Figure S7

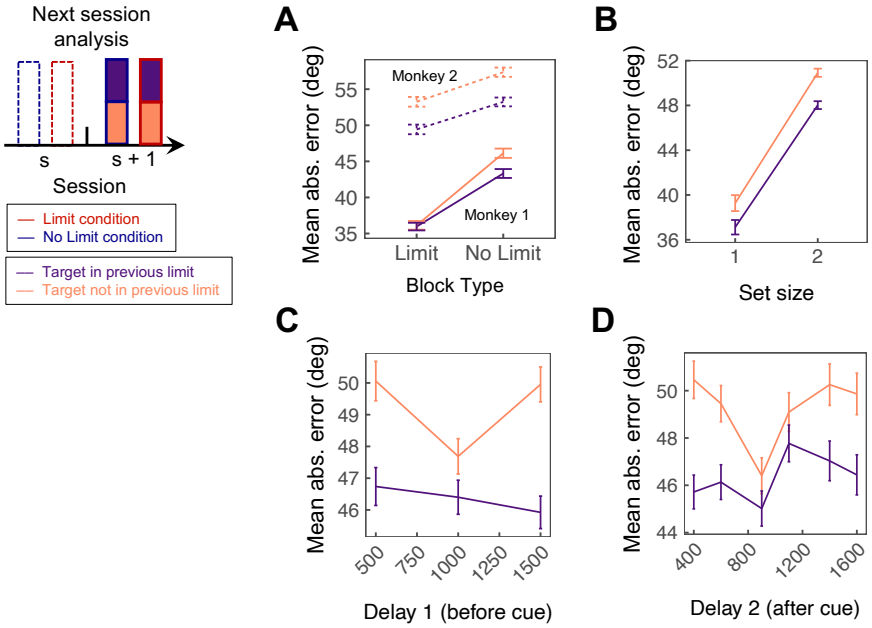
Same Session: Biases for all Limits



Biases towards the Limit mean were present in all Limit conditions. They seemed weaker for colours the monkey preferred (red: lower rightward biases in Limit 1; blue: lower biases in Limit 2).

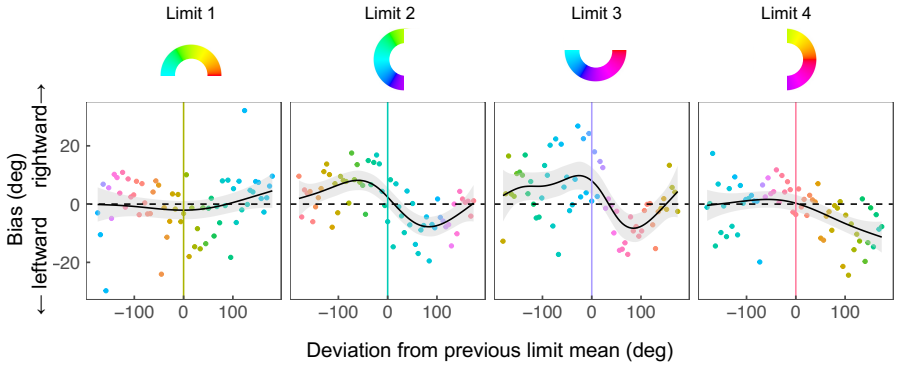
Figure S8

Next Session: Effects of Set Size and Delay



(A) The benefit for colours within the previous Limit were present in both blocks for Monkey 2, but only in the No Limit blocks for Monkey 1. **(B)** The performance benefit for colours in the previous Limit statistically disappeared when comparing between set sizes. The performance benefit for colours in the previous Limit did not depend on the delay between stimuli and retro-cue **(C)**, nor between retro-cue and report **(D)**.

Figure S9
Next Session: Biases for Previous Limits



Biases towards the Limit mean were not present in all Limit conditions. They seemed to only be present for means in Limit 2 (blueish) and Limit 4 (reddish).

Figure S10
PCA same session results for area 7ab

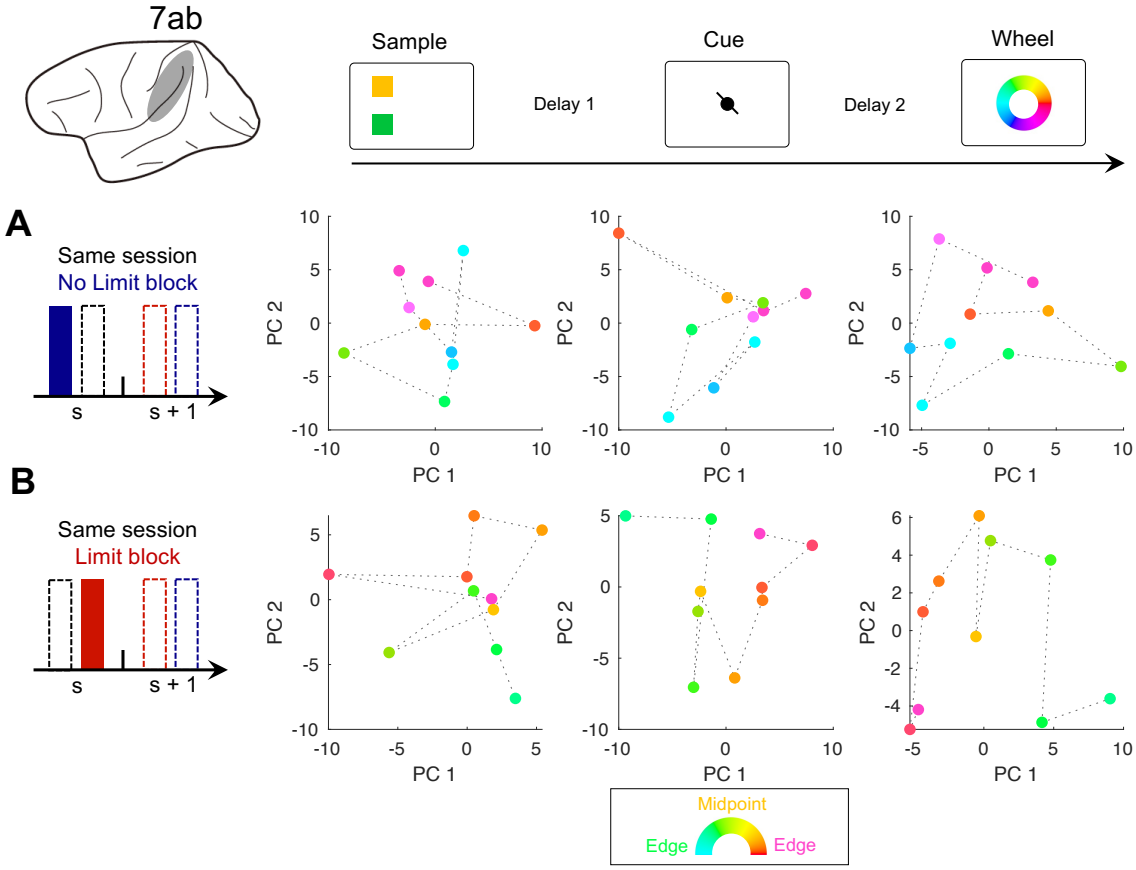


Figure S11
PCA same session results for area V4pIT

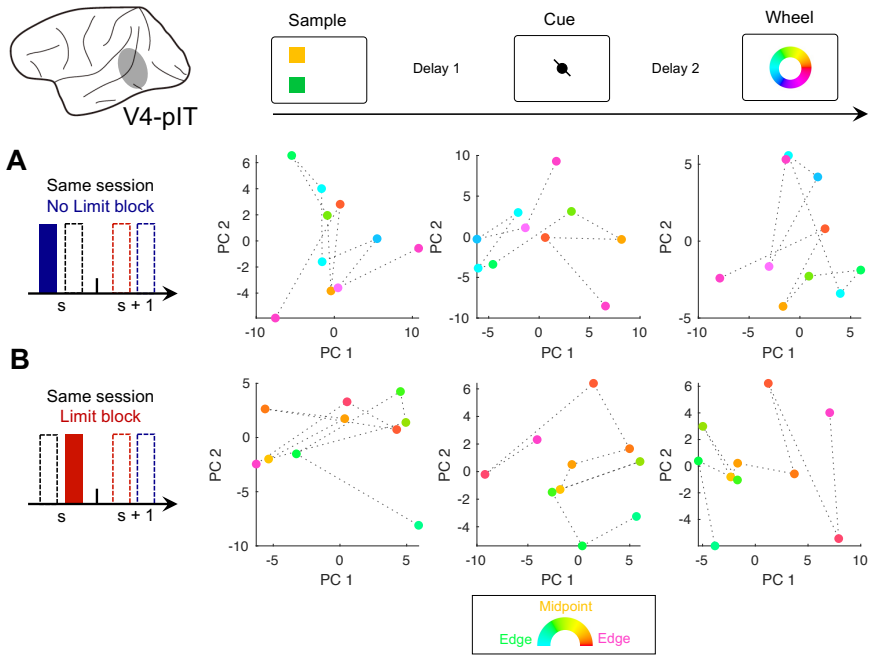


Figure S12
PCA same session results for Motor area

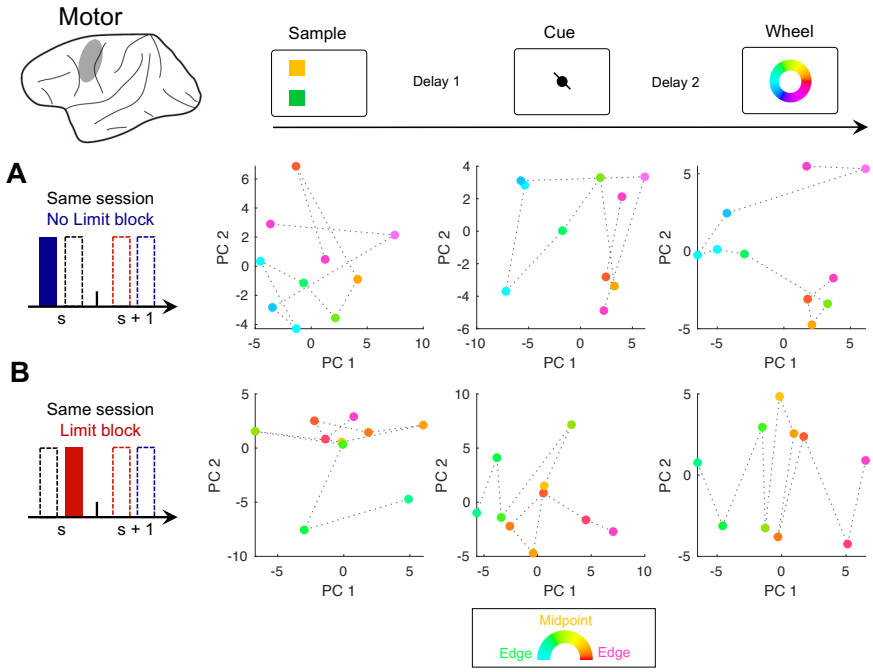


Figure S13
PCA next session results for area 7ab

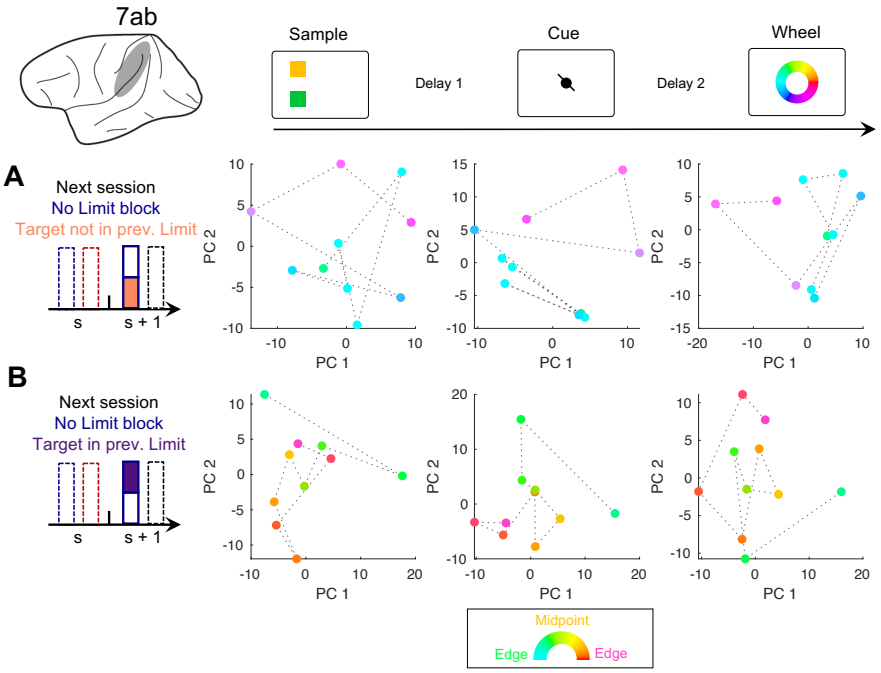


Figure S14
PCA next session results for area V4pIT

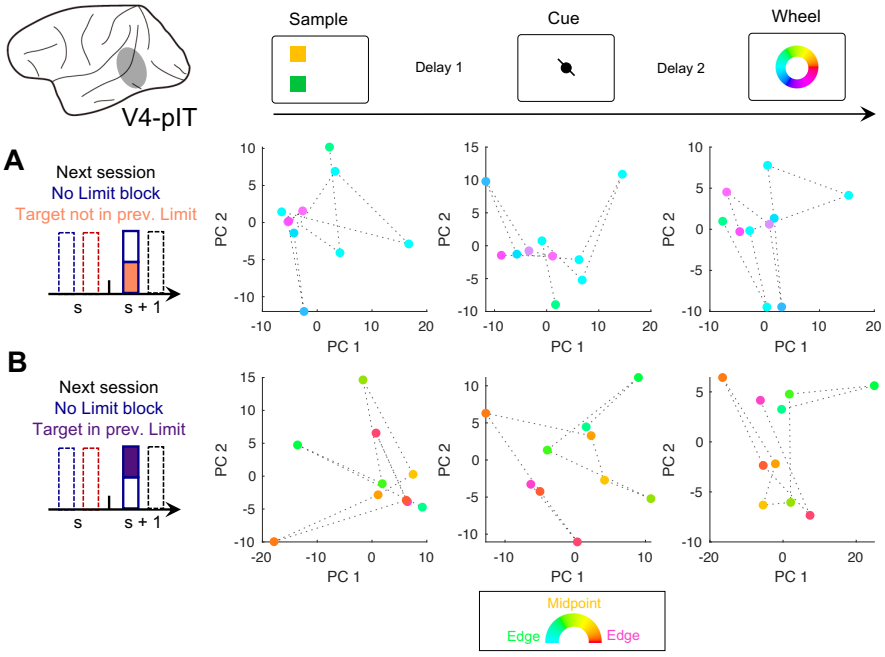
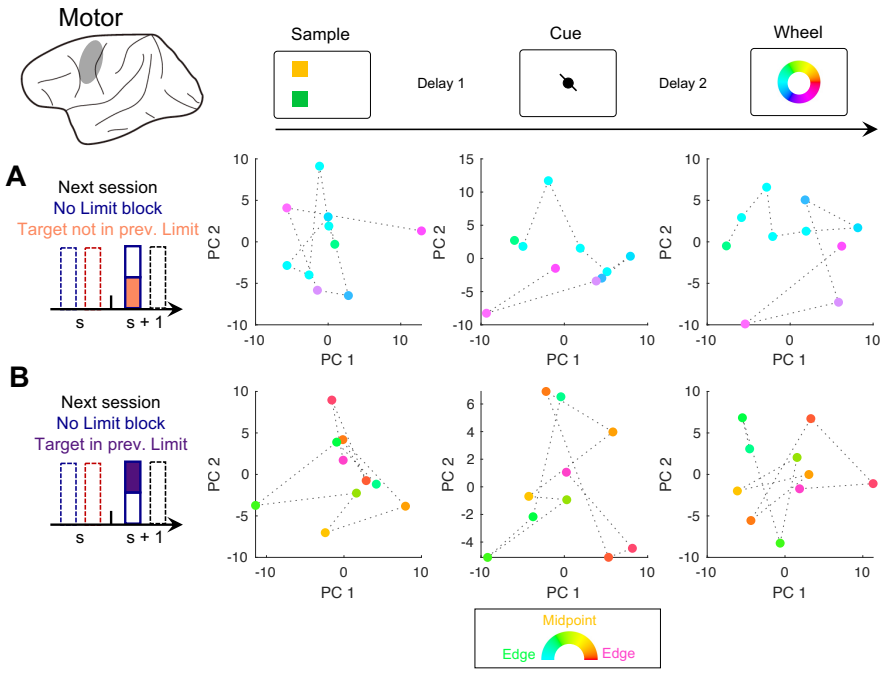


Figure S15
PCA next session results for Motor area



7.7.2 Tables

Table S20
Biases in different Limits

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	–0.146	1.946	1.530	–0.075	0.949
Dev. from circle mean	–0.166	0.018	21817.056	–9.223	0.000
Limit 2	0.624	1.218	21817.115	0.512	0.608
Limit 3	6.045	1.210	21797.090	4.997	0.000
Limit 4	–5.807	1.174	21817.175	–4.946	0.000
Dev. from circle mean:Limit 2	0.021	0.023	21817.011	0.904	0.366
Dev. from circle mean:Limit 3	–0.128	0.023	21817.102	–5.549	0.000
Dev. from circle mean:Limit 4	–0.038	0.023	21817.168	–1.682	0.093

Table S21
Performance in Limit vs. No Limit and In/Out prev. Limit

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	45.209	6.341	1.019	7.130	0.086
Limit Block	6.629	0.685	7354.595	9.675	0.000
Target in prev. Limit	–2.176	0.678	31369.565	–3.211	0.001
Limit Block:Target in prev. Limit	–1.341	0.905	31373.271	–1.481	0.139

Table S22
Performance in No Limit and In/Out prev. Limit

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	45.372	6.361	1.024	7.133	0.085
Limit Block	6.568	0.779	2969.663	8.434	0.000
Target in prev. Limit	–2.171	0.673	27237.028	–3.226	0.001
Limit Block:Target in prev. Limit	–1.614	0.958	27238.995	–1.685	0.092

Table S23
Performance in Limit and In/Out prev. Limit

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	53.326	6.497	1.033	8.208	0.072
Trial No.	–0.011	0.001	31382.051	–7.761	0.000
Target in prev. Limit	–4.701	0.883	31370.224	–5.324	0.000
Trial Number:Target in prev. Limit	0.004	0.002	31371.604	2.340	0.019

Table S24
Biases by previous Limits

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	–0.685	3.028	1.275	–0.226	0.852
Dev. from circle mean	–0.009	0.009	17571.250	–0.955	0.340
Prev. Limit 2	1.601	1.560	39.840	1.026	0.311
Prev. Limit 3	2.598	1.716	67.544	1.514	0.135
Prev. Limit 4	–1.679	1.532	77.706	–1.096	0.276
T:Prev. Limit 2	–0.030	0.012	17571.576	–2.410	0.016
T:Prev. Limit 3	–0.028	0.014	17574.538	–1.949	0.051
T:Prev. Limit 4	–0.026	0.013	17578.899	–2.005	0.045

Table S25
Behavioural set size effects

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	28.010	6.600	1.084	4.244	0.133
Limit Block	–4.995	2.196	33899.004	–2.274	0.023
Set Size	9.015	0.710	33899.004	12.704	0.000
Limit Block:Set Size	5.454	1.176	33899.013	4.636	0.000

Table S26
Behavioural effects of delay 1

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	46.600	6.640	1.023	7.018	0.087
Limit Block	4.477	1.229	33899.041	3.643	0.000
Delay 1	–0.002	0.001	33899.168	–3.019	0.003
Limit Block:Delay 1	0.000	0.001	33899.015	0.351	0.725

Table S27
Behavioural effects of delay 2

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	42.348	6.636	1.018	6.382	0.096
Limit Block	5.588	1.119	33899.001	4.994	0.000
Delay 2	0.002	0.001	33899.161	3.566	0.000
Limit Block:Delay 2	–0.001	0.001	33899.018	–0.692	0.489

Table S28
Behavioural effects of set size and previous Limit

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	31.329	6.896	1.311	4.543	0.093
Limit	–9.863	3.376	31372.058	–2.922	0.003
Target in prev. Limit	–4.073	3.392	31369.577	–1.201	0.230
setSize	7.417	1.293	31383.897	5.738	0.000
Limit:Target in prev. Limit	5.636	4.703	31370.744	1.198	0.231
Limit:setSize	8.896	1.790	31382.091	4.970	0.000
Target in prev. Limit:setSize	1.076	1.812	31369.667	0.594	0.553
Limit:Target in prev. Limit:setSize	–3.737	2.500	31369.848	–1.495	0.135

Table S29
Behavioural effects of delay 1 and previous Limit

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	49.782	6.464	1.053	7.701	0.075
Target in prev. Limit	–2.140	1.245	31371.195	–1.719	0.086
Delay 1	–0.001	0.001	31370.639	–0.793	0.428
Target in prev. Limit:Delay 1	–0.001	0.001	31370.517	–0.636	0.525

Table S30
Behavioural effects of delay 2 and previous Limit

	Est.	Std. Error	df	t value	Pr(>|t|)
(Intercept)	48.461	6.456	1.046	7.507	0.078
Target in prev. Limit	–3.664	1.129	31368.982	–3.245	0.001
Delay 2	0.001	0.001	31369.974	0.914	0.361
Target in prev. Limit:Delay 2	0.001	0.001	31370.090	0.757	0.449