How Is Information in Working Memory Selected and Prioritised for Action?



Frida Agnes Bang Printzlau
Department of Experimental Psychology,
St Catherine's College, University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Hilary Term 2021

How Is Information in Working Memory Selected and Prioritised for Action?

Frida Agnes Bang Printzlau, St Catherine's College,
Department of Experimental Psychology, University of Oxford

A thesis submitted for the degree of *Doctor of Philosophy*, Hilary Term 2021

Abstract

Flexible cognition necessary for everyday behaviour depends on working memory (WM)– the ability to keep information in mind for a later use. Not all items in WM are represented equally. Attention can be allocated to select and prioritise task-relevant content for a forthcoming action.

Chapter 2-3 used behaviour, eye-tracking and electroencephalography (EEG) to investigate the consequences of selecting a single feature in WM for the representation of other features belonging to the same item. Orientation judgements were faster and more precise when another feature of the same object was cued for an orthogonal task during the delay. Colour selection in WM engaged spatial attention, indexed by contralateral alpha suppression and gaze bias, and was associated with a boost in the representation of the cued object's memorised location (**Chapter 2**) and orientation (**Chapter 3**) as decoded from EEG voltages. The results provide evidence for obligatory attentional spread between an item's features in WM and support a special status of location in selecting and binding features. Finally, EEG decoding revealed that the cued item was automatically compared with the probe at recall, suggesting selected WM content may be in a prioritised state for interacting with new sensory input and guiding decision-making.

WM is more than a passive store of recent experiences and is fundamentally future-oriented**. Chapter 4** tested whether WM contents are configured as prospective task-sets for guiding behaviour. In one experiment, a prioritised item drove responses in a task-specific manner even when irrelevant in the context of a nested task. In a second experiment, task-foreknowledge improved WM-guided behaviour and enhanced retro-cued prioritisation on reaction times. Together, the results indicate that prioritised WM items are configured as prospective templates optimised for context-dependent action. Neuroimaging studies are needed to elucidate the precise mechanisms of task-specific coding in WM. **Chapter 5** proposes an EEG experiment, in the form of a Registered Report, to test the hypothesis that prioritised content is stored in a functional state that is intrinsically linked to its anticipated use. I predict that patterns of neural activity coding for remembered items will be task-specific. The suitability of the analysis pipeline was demonstrated on simulated neural data.

The results of this thesis highlight a key role of attentional selection in preparing WM representations for future action and motivate further research into the ways information in WM may serve as functional states for decision-making.

Acknowledgements

I have a lot of people to thank for helping me complete this DPhil. First of all I had an incredible supervisory team in Nick Myers, Sanjay Manohar and Mark Stokes. Thank you all for letting me run with my ideas and for helping me shape and sharpen them. Thank you, Nick, for your day-to-day support. Your guidance on scientific and personal questions have helped me find my balance through this DPhil. I have infinite admiration for your seemingly unlimited supply of helpfulness and snacks! Thank you, Sanjay, for asking the tough questions that might have baffled me, but also challenged me to think a step further. I will take your excellent writing tips with me. Thank you, Mark, for your mentorship. Our meetings generally left me with more questions than I started with, but that is what makes science fun, and is a reminder of why I wanted to pursue a DPhil in the first place.

Thank you to the entire Stokes lab–current and former members– for your friendship, support and great chats by the coffee machine or at the pub after Friday meetings. Thanks in particular to DPhil neighbours Dante Wasmuht and Sam Hall-McMaster, bonus mentors Paul Muhle-Karbe and Michael Wolff, as well as Ilenia Salaris and Jake Rich. Thanks also to friends at OHBA–our temporary home after fleeing Tinbergen–in particular Kia Nobre, Freek Van Ede, Sage Boettcher, and Giedrė Čepukaitytė. Thanks to Alex Rollings, Beatriz Moreno Gacía and Lucia Rampazzo for sharing the DPhil journey. We made it! Thanks to Tim Vogels and lab members for giving me a crash course in computational neuroscience. Thanks to the UKRI-BBSRC and the Oxford Interdisciplinary Biosciences DTP for their funding and support through Covid-19 delays. Thanks to all the participants that volunteered their valuable time and for their patience in completing my much-too-long experiments. Thank you to examiners Prof Nick Yeung and Dr Anastasia Kiyonaga for a stimulating conversation during my DPhil Viva and helpful comments on my thesis.

My family and friends provided the laughs, talks and hugs–physical and virtual–that got me through. I am grateful to my family for creating a ring of love around me that I take with me wherever I go: Linda Bang Jessen, Jacob Printzlau, Gertrud Printzlau, Johannes Printzlau, Niklas Printzlau, Imke Printzlau and Renée Van Naerssen. Thank you Mathilde Thykier, Linnea Langvad, Marie Fogh, Katja Andreasen, Ottilia Printzlau, Ada Printzlau and Stine Melsted for your long friendships. Finally, thank you to the man who lives in my house, Cameron Arnold, for being my rock of support and for assuring me I'm always enough while inspiring me to do more. I am looking forward to a new chapter.

Statement of contribution

I carried out the work in this thesis at the Department of Experimental Psychology and Oxford Centre for Human Brain Activity (OHBA) at the University of Oxford. The research was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) and supported by the Oxford Interdisciplinary Biosciences Doctoral Training Partnership. Science is a collaborative process and I would like to recognise the contribution of a number of people.

Chapter 2-3: I developed the ideas for the experiments in collaboration with my supervisory team Nicholas Myers, Sanjay Manohar and Mark Stokes. I designed the experimental task, collected and analysed the data and wrote up the results with the support of my supervisors. Michael Wolff advised on EEG decoding procedures and Freek Van Ede advised on eye-tracking analyses.

Chapter 4: the experiments were developed in collaboration with my supervisory team as well as Paul Muhle-Karbe. I programmed the experimental tasks and collected the data with the support of Jake Rich and Ilenia Salaris. Jake Rich collected the majority of the data for Experiment 1 (Chapter 4) as part of his undergraduate project in Experimental Psychology, which I co-supervised with Mark Stokes. I fully re-analysed the data and wrote up the results for this thesis. The results presented in Experiment 2 (Chapter 4) were published as a preprint (Printzlau, Myers, Muhle-Karbe, Manohar, & Stokes, 2019). As first author, I analysed the data and wrote up the results. My co-authors provided guidance on data analyses and commented on drafts of the paper.

Chapter 5: the chapter is based on a Stage 1 Registered Report, not yet accepted for publication. As first author, I wrote the Registered Report with comments and feedback from co-authors: Michael Wolff, Paul Muhle-Karbe, Kia Nobre, Sanjay Manohar, Nick Myers and Mark Stokes. I developed the experimental design and proposed analyses in collaboration with my co-authors. The data simulations were performed with the support of Nicholas Myers and Mark Stokes.

Table of contents

Abstract ii

Acknowledgements iii

Statement of contribution iv

Table of contents v

Chapter 1. General introduction 1

Chapter 2. Attention spreads between object-features in working memory 66

Chapter 3. Colour-selection in working memory engages spatial attention and strengthens neural representation of associated features 126

Chapter 4. Prospective working memory: Is working memory supported by task-specific representations? 169

Chapter 5. Use-specific working memory: does coding depend on expected future decision-making? 203

Chapter 6. General Discussion 229

References 259

Appendix A. Supplementary material for Chapter 2 278

Appendix B. Supplementary material for Chapter 3 286

Appendix C. Supplementary material for Chapter 4 295

Appendix D. Supplementary material for Chapter 5 296

Approximate word count: 75,000

Chapter 1. General introduction

1. What is working memory?

Working memory (WM) is the ability to hold information in mind for brief periods of time and manipulate memorised content in the service of behavioural goals (Baddeley, 2003). Cognitive flexibility and fluid intelligence, important for solving everyday tasks, crucially depends on WM (Fukuda, Vogel, Mayr, & Awh, 2010; Unsworth, Fukuda, Awh, & Vogel, 2014). For example, when you are in the supermarket and have forgotten to bring a shopping list, you might make a mental list of the items you need. As you walk through the bakery section you are perhaps temporarily distracted by the muffins before you continue to look for the bread that you came for. Once you find it and drop it in your basket you can update your mental list accordingly and start searching for the next item. In that way your WM system allows you to store a mental list, select items from the list to guide your behaviour, and flexibly update the contents of WM according to current goals. In this thesis, I am specifically interested in how information is selected from WM and used to guide decision-making and actions.

In this section, I will briefly introduce three influential perspectives on WM that have shaped the direction of research in the field: the multi-store model of memory (Atkinson & Shiffrin, 1968), the multicomponent model of working memory (Baddeley & Hitch, 1974), and state-based models of working memory (Cowan, 1998; Oberauer, 2002)(Figure 1).

1.1. Multi-store model of memory

In *Principles of Psychology* (1890, p 646-647) William James made a distinction between what would later come to be known as short-term and long-term memory: "*An object which is recollected, in the proper sense of that term, is one which has been absent from consciousness altogether, and now revives anew. […]. But an object of primary memory is*

not thus brought back; it never was lost; its date was never cut off in consciousness from that of the immediately present moment." (James, 1890).

The distinction between short- and long-term memory was formalised by Atkinson and Shriffin in 1968 when they proposed the multi-store model of memory (Atkinson & Shiffrin, 1968). The model consists of three separate but interrelated memory stores: a sensory register for each sensory modality (e.g. iconic memory for visual information) (Sperling, 1960), a short-term memory system (STM) and a long-term memory system (LTM)(Figure 1A). According to the model, incoming sensory information is briefly available in the sensory register (0.5-1 s) after which it decays and disappears. Attended information may then get transferred to the short-term store (up to 30 s), which has a limited capacity of 7 +/- 2 items, or chunks of information (G. A. Miller, 1956). They write "*The short-term store is the subject's working memory; it receives selected input from the sensory register and also from long term store*" (p. 90)(Atkinson & Shiffrin, 1968). Loss of information from STM may be prevented through rehearsal and is thus also transferred to LTM, which has unlimited capacity.

Initial evidence for the existence of multiple memory systems came from neuropsychological studies of patients with bilateral hippocampal lesions (Baddeley, Jarrold, & Vargha-Khadem, 2011; Baddeley & Warrington, 1970; Milner, 1966; Scoville & Milner, 1957; Squire & Zola-Morgan, 1991; Zola-Morgan, Squire, & Amaral, 1986). Patient HM is the most well-known of such patients (Scoville & Milner, 1957). As treatment for epilepsy, he received surgery to remove parts of the medial temporal lobe (MTL), including the hippocampus and anterior parahippocampal gyrus (Corkin, Amaral, Gilberto González, Johnson, & Hyman, 1997; Scoville & Milner, 1957). Following the surgery, HM suffered severe anterograde amnesia as well as retrograde amnesia specific to the recent past (<10 years)(Scoville & Milner, 1957). HM's short-term memory performance on digit-span and dichotic listening tasks was intact (Corkin, 2002;

Milner, 1966), pointing to a dissociation between STM and a declarative LTM system, dependent on the medial temporal lobe (Squire & Zola-Morgan, 1991).

Further evidence for a dissociation came from patient KF, who following damage to left parietal and occipital cortex showed impaired verbal STM, and to a lesser extent visual STM, but had relatively spared LTM function (Shallice & Warrington, 1974; Warrington & Shallice, 1969). Although evidence from any single lesion patient should be interpreted with caution due to the uncertainty around the precise size and location of the lesion, together with functional impairments seen in hippocampal patients, KF's pattern of memory impairment supports a dichotomy between STM and LTM function. However, contrary to a serial encoding model (Atkinson & Shiffrin, 1968), it suggests information may in some cases be encoded into LTM without first entering the STM store. Additionally, the pattern of impairment in patient KF points to a separation between verbal and visual information within STM (Shallice & Warrington, 1974; Warrington & Shallice, 1969), which was later incorporated into the multicomponent model of working memory (Baddeley & Hitch, 1974) (Section 1.2).

Although the division of short- and long-term memory stores is not universally accepted (Nairne, 2002; Ranganath & Blumenfeld, 2005) and the model has been criticised for its simplification of STM and LTM as unitary stores (Baddeley & Hitch, 1974), it has nevertheless been hugely influential in guiding our understanding of memory processes in general and WM specifically–most critically, in the context of this thesis, the formulation of STM as a capacity-limited system for temporary storage that is intimately connected with conscious processing and attention.

1.2. The multicomponent model of working memory

The concept of *working memory* (WM), as it is used today, was first proposed in the multicomponent model of WM (Baddeley, 1992; Baddeley & Hitch, 1974). Rather than a

pure short-term store, WM was proposed to be a system for the concurrent storage and manipulation of information.

The model accepted the dichotomy between STM as a "fluid" system for active storage and processing and LTM as a "crystallized" system for passive storage and learning. It expanded on the theoretical unitary STM store to include three separate components of WM. The initial model consisted of two limited-capacity "buffers" for temporary storage: one for the storage of auditory and verbal information (Phonological Loop) and one for the storage of visual and spatial information (Visuo-Spatial Sketchpad). Finally, the model included a separate Central Executive responsible for coordinating and allocating limited attentional resources between the storage buffers—processes thought to rely on functions of the prefrontal cortex (D'Esposito & Postle, 1999; Wager & Smith, 2004). The model was later updated to include the Episodic Buffer (Baddeley, 2000), a capacity-limited system to hold episodic events in multidimensional code by integrating information from the two other buffers and LTM (Figure 1B).

In addition to neuropsychological evidence described above (Shallice & Warrington, 1974), support for multiple domain-specific WM stores came from the dual-task paradigm (Baddeley & Hitch, 1974). In dual-task paradigms, participants are asked to perform a simultaneous task that demands resources from the limited-capacity WM system, thereby functionally disabling STM temporarily to reveal whether performance on a main task competes for the same WM resources. A visually presented grammatical reasoning task was only somewhat impaired by a concurrent verbal digit span task, suggesting the existence of distinct WM subsystems for processing visuo-spatial and verbal information (Baddeley & Hitch, 1974). Further divisions of the WM subsystems have since been proposed, specifically separate systems for the storage of visual/object-based and spatial information (Klauer & Zhao, 2004; Smith & Jonides, 1997).

In summary, the multi-component model of working memory considers WM to be a limited-capacity system that allows for the temporary storage as well as manipulation of information needed to perform a range of complex cognitive tasks (Baddeley, 2012). The distinction between verbal and visuo-spatial WM has informed a general separation in the field and this thesis will mainly consider visuo-spatial WM. Importantly, with the central executive, the model highlighted the crucial role of attention in directing WM processing in the service of a behavioural goal. The role of attention in WM, including the ability to flexibly select and prioritise information in WM, will be the focus of this thesis.

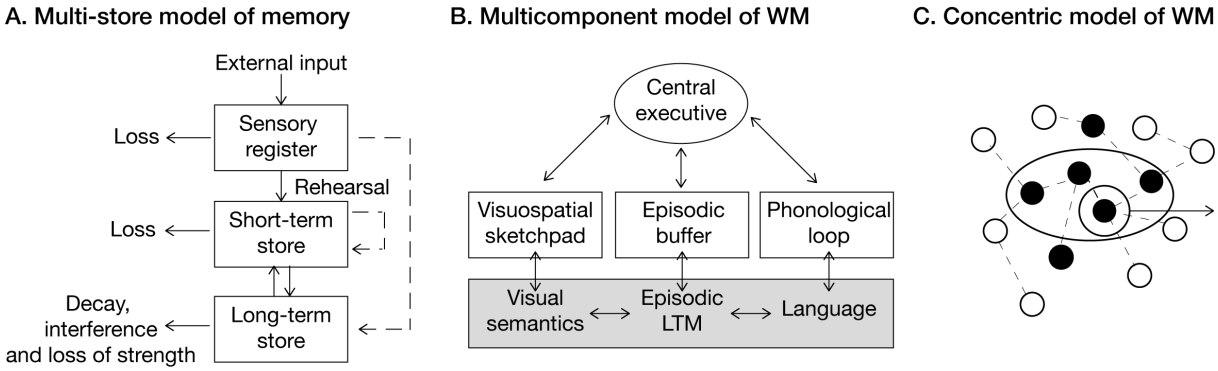

Figure 1. Perspectives on WM. **A.** The multi-store model of memory (adapted from Atkinson & Shriffin, 1968). **B.** The multicomponent model of WM (adapted from Baddeley, 2000). The grey box represents the LTM system and white boxes represent a WM system for temporary storage. The central executive directs resources between the storage buffers. **C.** The concentric model of WM (adapted from Oberauer, 2002) as an example of a state-based model of WM. Lines and nodes represent a network of representations in LTM. Filled nodes are activated LTM representations. Some of these are held in a region of direct access (large oval). The smaller oval represents a single item in the focus of attention, selected for the next cognitive operation (arrow).

1.3. State-based models of working memory

Rather than rely on a distinct WM system to store and process information in separate buffers, a prominent theory suggests WM may operate within the same structures used for perception and long-term storage (Jonides et al., 2008; Postle, 2006). Neuroimaging studies suggest MTL structures, thought to be specialised for LTM, are often recruited during WM tasks (Ranganath & Blumenfeld, 2005). While early neuropsychological

studies showed spared STM performance in patients with MTL lesions, in some cases patients do show STM impairments for new or complex stimuli, perhaps reflective of a general role of the MTL in relational processing and binding (Graham, Barense, & Lee, 2010; Hannula, Tranel, & Cohen, 2006; Olsen, Moses, Riggs, & Ryan, 2012; Pertzov, Miller, et al., 2013).

State-based models consider WM and LTM to be separate processes within a common memory system and distinguished mainly in terms of their content and attentional state (Larocque, Lewis-Peacock, & Postle, 2014). One example is the Embedded Processes Model proposing that WM consists of two processes that operate within a wider LTM network (Cowan, 1998): (1) a subset of LTM that is currently "activated" or in a heightened state of accessibility (aLTM), equivalent to a short-term store, and (2) a subset of the activated portion of LTM that is currently in the focus of attention (FoA) or in conscious awareness. The FoA has a limited capacity of approximately four integrated items or chunks (Cowan, 2001), whereas aLTM, has no capacity limits as such, but is subject to temporal decay. Another example is the Concentric Model of WM (Oberauer, 2002)(Figure 1C). Like the Embedded Processes Model, it regards WM as an activated portion of LTM. However, instead of a four-item FoA, it posits a region of direct access holding 3-4 items from which a single item is selected for processing at any one time (FoA). As such, it proposes three states in WM: activated portions of LTM that may be retrieved through associative connections, the capacity-limited region of direct access holding items immediately available for selection, and the FoA containing a single item selected for cognitive processing.

State-based models of WM are supported by evidence of different representational states in WM (Oberauer, 2001; Olivers, Peters, Houtkamp, & Roelfsema, 2011; Peters, Goebel, & Roelfsema, 2009). For example, the recency effect–that the last item in a list is retrieved faster and with greater accuracy than earlier items–suggests the last item may

be maintained in a more accessible state than other items in memory (McElree & Dosher, 1989). However, non-attentional processes, such as reduced interference from other memory items, could also contribute to the recency effect (Cowan, 2011). Another line of evidence, making use of retrospective cues ("retro-cues") regarding which items are relevant to an upcoming decision shows that attention can be allocated to privilege relevant items in memory (Griffin & Nobre, 2003; Oberauer, 2001). However, uncued items are not completely lost and may continue to affect processing. This is revealed by an intrusion effect where probes from the uncued set of memory items (lures) are rejected slower than never-encountered probes, indicating residual activation of the recently relevant WM items, perhaps within the activated portion of LTM (Oberauer, 2001, 2002). Neuroimaging studies show greater MTL recruitment for retrieval of earlier items in a list relative to the most recent item, suggesting perhaps that while the most recent item is highly accessible in the FoA, retrieval of earlier items depend on associative mechanisms supported by the MTL, similar to retrieval in LTM (Nee & Jonides, 2011; Oztekin, Davachi, & McElree, 2010).

1.4. Interim summary

In summary, the perspectives on WM presented briefly in this section converge on a view of WM as a capacity-limited system for short-term storage and processing of information that is highly dependent on interactions with attention, perception and long-term memory. Whether WM is best considered an emergent property of attention operating on a shared representational foundation or whether WM should be regarded as a distinct subsystem remains unresolved. Given the highly distributed network of brain regions recruited during WM tasks (Christophel, Klink, Spitzer, Roelfsema, & Haynes, 2017) and an increasing need for separate buffers for information storage in a multi-component model (Baddeley, 2012), it seems likely that WM "piggybacks" on a broader network

involved in perception, LTM and motor planning in order to guide goal-directed behaviour (Jonides et al., 2008; Postle, 2006).

2. Neural basis of working memory

2.1. Prefrontal cortex

Early studies of patients with large lesions to the lateral prefrontal cortex (PFC) point to an important role of this area in executive functions, including planning, attention and WM (D'Esposito & Postle, 1999; Fuster, 2001; Luria, 1966). As discussed in Section 1, WM involves storing relevant sensory information temporarily and using it flexibly to achieve a behavioural goal. Given its position at the apex of the cortical hierarchy, the PFC may be in a unique position to support WM guided-behaviour, by integrating information from the sensory association cortex, motor cortex, and structures involved in processing motivational states, such as the thalamus and limbic system, as well as passing this information on to motor areas to execute behaviour (Miller & Cohen, 2001). Fuster (2001) suggested that the lateral PFC (lPFC), specifically, may be crucial to temporary integration by bridging the gap between incoming sensory information and goal-directed behaviour.

In support of a role of lPFC in WM are a wealth of studies showing sustained activity in lPFC neurons while non-human primates perform delay tasks, thought to depend on WM (Chafee & Goldman-Rakic, 1998; Funahashi, Bruce, & Goldman-Rakic, 1989; Fuster, 1973; Fuster & Alexander, 1971; E. K. Miller, Erickson, & Desimone, 1996). For example, using an oculomotor delayed response task, in which monkeys were required to saccade to a memorised location after a delay, Funahashi et al. (1989) found that a large portion of neurons in dorsolateral PFC showed item-specific persistent activity bridging the delay between sensory input and response. This persistent activity was absent on

error trials, suggesting a functional role in WM performance. Together with earlier indications that lPFC lesions in non-human primates impaired performance on delay tasks (Butters & Pandya, 1969; Fuster, 1980; Jacobsen, 1936), there is strong evidence for an important role of the lPFC in WM function. However, the precise nature of PFC involvement remains a topic of active debate.

2.1.1. Nature of PFC involvement

What is the role of the PFC in WM function? An influential review by Goldman-Rakic (1995) proposed that neurons with memory fields in the PFC displaying item-specific persistent activity may be the neural substrate of WM representations, with different prefrontal areas coding for different features and different levels of complexity (Goldman-Rakic, 1995). In addition, computational models have shown that item-specific persistent delay activity can be achieved in artificial neural networks with recurrent connectivity to sustain mnemonic activity within a stable attractor, mimicking attractor states found in the PFC (Compte, Brunel, Goldman-Rakic, & Wang, 2000; Wang, 2001; Wimmer, Nykamp, Constantinidis, & Compte, 2014).

However, there are challenges to this model (Lebedev, Messinger, Kralik, & Wise, 2004; Shafi et al., 2007; Watanabe & Funahashi, 2014a). What may look like persistent delay activity when averaging a neuron's activity over trials is a lot more variable at the single-trial level and often occurs in discrete "bursts" (Lundqvist et al., 2016; Shafi et al., 2007; Stokes & Spaak, 2016). Furthermore, item-specific neural activity in the PFC seems to depend on task-relevance (Fuster, 2001; Watanabe & Funahashi, 2014a). When monkeys performed a dual WM and attention task, locations maintained in WM and attended perceptual locations seemed to compete for representation in overlapping neural populations (Watanabe & Funahashi, 2014a). Memory-specific delay activity dropped to baseline in the dual task conditions compared to the trials with no dual task

interference. Additionally, the degree of behavioural interference corresponded to the degree of overlap in neural populations, suggesting that processing capacity can be flexibly allocated between different tasks depending on current relevance.

This has led to suggestions that rather than coding for specific WM content, the role of the PFC is to flexibly code for the current internal goal or attentional focus, similar to the role of the central executive in the multicomponent model of WM or the focus of attention in state-based models (Section 1.2-1.3)(Baddeley & Hitch, 1974; Cowan, 1998; Funahashi, 2017; Oberauer, 2002). In a multi-colour WM task, requiring monkeys to detect subtle changes in colour at test, relatively few neurons in the ventrolateral PFC coded for colour information, while the majority coded for the spatial positions of memory items and allocation of spatial attention (Lara & Wallis, 2014). These results suggested that PFC activity might instead reflect control processes that support maintenance in sensory cortices. In other words, the PFC may represent internal goals in order to bias processing along the appropriate pathways in the cortical hierarchy to represent task-relevant sensory information and achieve the desired action (Desimone & Duncan, 1995; Kastner & Ungerleider, 2000; Lara & Wallis, 2015; E. K. Miller & Cohen, 2001; Voytek & Knight, 2010).

A role of the PFC in the control rather than storage of WM is further supported by human fMRI studies (D'Esposito & Postle, 2015). Despite the fact that mean blood oxygen-level dependent (BOLD) activity shows consistent activation in the lPFC during WM tasks (Curtis & D'Esposito, 2003; Duncan & Owen, 2000), several studies requiring participants to store precise representations for subtle discrimination judgements have found no evidence for WM representations in the PFC (Christophel, Hebart, & Haynes, 2012; Postle, 2006; Sreenivasan, Vytlacil, & D'Esposito, 2014). However, interpretations of null-findings should be made with caution, especially since base rates of fMRI decoding accuracy from PFC are low (Bhandari, Gagne, & Badre, 2018). Additionally, at

least one study did find significant decoding of fine-grained WM representations from BOLD activity in the lPFC (Ester, Sprague, & Serences, 2015).

Finally, it is possible that the lPFC is involved in top-down control as well as in the storage of some types of WM content. Specifically, the lPFC may be involved in maintaining abstract, category-level information, whereas maintenance for fine-grained sensory decisions require representation in the sensory cortex (Lara & Wallis, 2014; Lee, Kravitz, & Baker, 2013). Lee et al. (2013) asked participants to perform two types of WM tasks: either to discriminate objects based on subtle, low-level visual features or based on their category. Using multivariate pattern analyses of BOLD activity in fMRI, they found item-specific delay activity in the lPFC for the category task, but not for the sensory-discrimination task, and vice versa for the sensory cortex.

Representations in the PFC tend to be flexible rather than fixed, with many PFC neurons switching their selectivity during a task (Spaak, Watanabe, Funahashi, & Stokes, 2017) or coding for non-linear combinations of task features (i.e. non-linear mixed selectivity)(Rigotti et al., 2013). These properties may put the PFC in a unique position to selectively represent task-relevant features and support task-dependent information processing, as well as filter out irrelevant features to encode content in an abstract, generalizable format depending on current behavioural goals (Bernardi et al., 2020; Fusi, Miller, & Rigotti, 2016; Mack, Preston, & Love, 2020; Rigotti et al., 2013).

There is little doubt that the PFC forms an important part of a larger distributed network of interconnected brain areas, contributing to WM-guided behaviour (Christophel et al., 2017). WM delay representations have also consistently been found in other areas including the posterior parietal cortex (PPC)(Bettencourt & Xu, 2016; Christophel et al., 2012; Jeong & Xu, 2016) and sensory cortex (Emrich, Riggall, La Rocque, & Postle, 2013; Harrison & Tong, 2009; Rademaker, Chunharas, & Serences,

2019; Serences, Ester, Vogel, & Awh, 2009) using multivariate analyses of fMRI data in humans, as well as single unit recordings from the intraparietal sulcus (IPS)(Chafee & Goldman-Rakic, 1998; Gnadt & Andersen, 1988) and sensory cortex (Fuster & Jervey, 1982; Nakamura & Kubota, 1995) of non-human primates. I will briefly discuss the evidence for the involvement of these areas in WM in turn.

2.2. Posterior parietal cortex (PPC)

PPC is part of a frontoparietal network often activated during WM tasks (Chafee & Goldman-Rakic, 1998; Christophel et al., 2017; Curtis, 2006). However, as is the case for the PFC, there is ongoing debate about whether its role in WM is for storage, top-down control, or both.

One line of evidence implicating the PPC in WM storage are studies showing a close link between WM capacity limits and PPC activity (Todd & Marois, 2004, 2005; Vogel & Machizawa, 2004; Vogel, Mccollough, & Machizawa, 2005). During a delayed-match-to-sample task, peak BOLD activity in PPC closely tracked the numbers of items encoded into WM at different set sizes, increasing up to mean capacity of 3-4 items (Luck & Vogel, 1997), before levelling off (Todd & Marois, 2004). Similarly, the amplitude of contralateral delay activity—observed in electroencephalography (EEG) and thought to originate from the parietal cortex (R. Luria, Balaban, Awh, & Vogel, 2016)—tracked the number of maintained items up to mean capacity, while the increase in CDA amplitude correlated with individual differences in WM capacity (Vogel & Machizawa, 2004). Limited capacity is a defining feature of WM function, so an area that tracks the amount of successfully maintained information may play a critical role in its storage.

Together with the lPFC, the PPC is part of an interconnected network involved in the top-down control of attention (Chafee & Goldman-Rakic, 2000; Corbetta & Shulman, 2002; Kastner & Ungerleider, 2000). The PPC is commonly implicated in the guide of

attention (Gottlieb, Kusunoki, & Goldberg, 1998; Jerde, Merriam, Riggall, Hedges, & Curtis, 2012; Shomstein & Gottlieb, 2016). It is possible that the correlations between PPC activity and the number of items in WM instead reflect top-down biasing signals toward items in memory relevant for guiding behaviour. One study suggested PPC activity may reflect attentional demands on WM as PPC activity continued to increase beyond WM capacity limits at longer delays, presumably recruiting greater attentional resources for rehearsal (Magen, Emmanouil, McMains, Kastner, & Treisman, 2009). Another study proposed that recruitment of the PPC is necessary for feature-context binding (Gosseries et al., 2018). While WM load was kept constant, demands on context-binding were modulated by drawing all items from the same category (greater demands on binding) or from different categories (lower demands on binding). Rather than load *per se*, PPC activity seemed to be recruited for distinguishing between objects in memory based on temporal order.

Finally, different portions of the PPC may be recruited for processes related to spatial attention and storage (Xu & Chun, 2006). While the inferior portion of the intraparietal sulcus (IPS) was shown to track the number of items in WM only when these were presented in distinct spatial locations, consistent with a spatial attention account, superior portions of the IPS and lateral occipital complex tracked the number of items in WM independent of the number of locations, consistent with a role in information storage. A storage function of the PPC is further supported by studies showing multivariate decoding of spatial locations, shapes and gratings during the delay of WM tasks in this area (Christophel et al., 2012, 2017; Ester et al., 2015), even when these are outside the focus of attention and can't be decoded from the sensory cortex (Bettencourt & Xu, 2016; Christophel, Iamshchinina, Yan, Allefeld, & Haynes, 2018). This has led to suggestions that the PPC may be important for storing distractor-resistant

representations in WM (Bettencourt & Xu, 2016), perhaps at a lower resolution (Christophel et al., 2018; Rademaker et al., 2019)(discussed further in Section 4.4).

2.3. Sensory cortex

Rather than being stored in a separate region of the brain, such as the PFC or PPC, an influential view proposes that WM content is maintained in the same functionally specialised regions involved in perception, as well as long-term storage of content (introduced in Section 1.3)(Jonides et al., 2008; Postle, 2006). This has become known as the sensory recruitment hypothesis (D'Esposito & Postle, 2015; Postle, 2006; Scimeca, Kiyonaga, & D'Esposito, 2018; Serences, 2016; Sreenivasan, Curtis, & D'Esposito, 2014).

Support for the sensory-recruitment hypothesis gained traction following the development of multivariate analysis methods of neuroimaging data, that consider information stored in the pattern of neural activation, rather than the mean activation profile (Haxby et al., 2001). Sensory cortex activation tends to return quickly to baseline during WM tasks, but WM content can nevertheless be decoded from activity patterns in the sensory cortex (Harrison & Tong, 2009; Serences et al., 2009). For example, using fMRI, Harrison & Tong (2009) showed that oriented gratings could be decoded with high accuracy during the delay of a visual WM task from early visual cortex areas V1-V4. Similar findings have since been reported for many different stimuli in different areas of sensory cortex (Christophel et al., 2017; D'Esposito & Postle, 2015; Sreenivasan, Vytlacil, et al., 2014). The quality of representations in early visual cortex predicts memory precision (Ester, Anderson, Serences, & Awh, 2013; Sligte, Scholte, & Lamme, 2009) and early visual cortex decoding accuracy decreases with increasing WM load, in a way that scales with reductions in memory precision (Emrich et al., 2013). Sensory cortex consequently seems to play an important role in storing high-quality memory representations.

However, one of the hallmarks of WM function is the ability to bridge the gap between sensory input and behaviour, even in the presence of distracting sensory information. Sensory cortex, by nature, needs to process incoming stimuli and as such may not be in a good position to maintain distractor-resistant WM representations. Indeed, Bettencourt and Xu (2016) found that task-irrelevant distractors extinguished stimulus decoding in sensory cortex without costs to behaviour, while decoding remained significant in PPC (Bettencourt & Xu, 2016). However, Rademaker and colleagues found evidence for co-existent representations of WM content and distractors in sensory cortex (Rademaker et al., 2019). It is possible that internally and externally generated input are separated by the laminar organization of visual cortex, as early evidence from laminar fMRI recordings suggest WM representations may be specifically represented in deep and superficial laminar layers of V1, while externally generated input is known to target middle layer 4 (Lawrence et al., 2018; Xu, 2020). An alternative, though not mutually exclusive, possibility is that sensory cortex representations of WM content might depend on the extent to which distractors capture attention (Rademaker et al., 2019). Bettencourt & Xu (2016) used faces and gazebos as distractors, which may be more interesting than the noise patches and oriented gratings used in Rademaker et al. (2019). This would be consistent with work showing that while multiple items can be maintained in WM, only the currently attended item is robustly represented in activity patterns in the sensory cortex (LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013; Larocque et al., 2014; Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012; Sprague, Ester, & Serences, 2016)(see also Section 4.4.3).

2.4. Interim summary

WM is supported by a distributed network of interconnected brain areas, including areas in the lPFC, PPC and sensory cortex. Whether the role of the frontoparietal network is in the control or storage of information is subject to ongoing debate. It is possible that

information in WM is stored in the same regions involved in representing information during perception and long-term memory. If so, a frontoparietal network may play an important function in the control of representations in sensory cortex, as well as in the storage of certain types of content such as abstract, category-level information (PFC) and spatial information (parietal cortex).

3. What is represented in WM: features or objects?

WM is incredibly flexible and allows us to keep anything in mind, including novel combinations of features. One of the key tenets of the sensory recruitment hypothesis is that the site of WM storage is determined by *what* is stored in WM. If an object's features are represented in the same areas of cortex that are responsible for representing these during perception, how are these features combined into coherent object representations in WM? This question has motivated a long-standing debate about whether WM primarily represents integrated objects or an object's component features.

3.1. The feature-binding problem

The ability to organise novel combinations of features (e.g. colours, shapes, orientations) into coherent object representations and maintain these over time is critical for perception, memory and behaviour. In 1980, Feature Integration Theory proposed that spatial attention is needed to bind visual features, represented in functionally specialized channels in visual cortex, into coherent object representations during perception (Treisman and Gelade 1980). In a series of visual search paradigms, search times were fast and independent of the number of distractors in the display when the target could be identified by a single feature (pop-out search). In contrast, search time increased in proportion to the number of distractors when the target could only be identified by a conjunction of features, suggesting identifying conjunctions required serial allocation of spatial attention. Additionally, when serial spatial attention was

prevented, for example, by limiting display times or directing attention toward other objects in a display, participants tended to report "illusory conjunctions" i.e. erroneous combinations of features, suggesting object-features outside the focus of attention may be "free-floating" (Treisman & Schmidt, 1982).

3.2. Unit of representation in WM

Whether focused attention is similarly needed to bind features in WM (Wheeler & Treisman, 2002) or whether, once encoded, features are stored in the form of integrated objects (Luck & Vogel, 1997) has been a topic of active debate. In a seminal study, Luck and Vogel (1997) used a change-detection paradigm to show that the capacity of visual WM was limited to four objects, whether these objects contained only a single feature (e.g. colour) or multiple features from different feature dimensions (e.g. coloured shapes) or from within the same feature dimension (e.g. bi-coloured squares). This led the authors to suggest that the unit of representation in WM was integrated or "bound" objects rather than individual features represented separately. This has since become known as the 'slot-model,' as WM capacity is thought to be limited to four "slots" of information.

Some of the key conclusions reached by Luck and Vogel (1997) have since been disputed. First, a model where individual features are maintained in separate feature stores with individual capacity limits would similarly predict that multiple features on the same object can be stored without costs to performance, as long as these are from different feature dimensions (Wheeler & Treisman, 2002). Although this account could not explain a lack of performance costs when object-features were from the same dimension (e.g. bi-coloured objects)(Luck & Vogel, 1997), others have found that features from the same dimension do compete for WM capacity (Olson & Jiang, 2002; Wheeler & Treisman, 2002). Additionally, even when the number of features to be remembered was the same,

memory performance suffered when participants were required to remember the bindings between features from different dimensions (i.e. only respond 'match' if objects had the same conjunction of features as the memory display)(Wheeler & Treisman, 2002). However, this performance cost was specific to 'whole-display' probe conditions where all memory objects were presented again compared to single-probe conditions, suggesting maintenance of bound objects might be uniquely vulnerable to perceptual distraction from the probe display. This led Wheeler and Treisman (2002) to propose a model of WM whereby features from different dimensions are stored in parallel while maintenance of correct feature-bindings requires sustained attention. However, the requirement for sustained attention is challenged by findings that memory for single features and conjunctions are disrupted to the same extent by a demanding task during a WM delay (J. S. Johnson, Hollingworth, & Luck, 2008).

Another line of critique of the Slot Model questions the suitability of classic change-detection paradigms to measure WM. Successful change-detection can be based on relatively imprecise representations in WM, depending on the degree of change participants are required to detect (Fougnie, Asplund, & Marois, 2010). When the level of detail required to perform successful change-detection is varied, there is a trade-off between the number of items maintained in WM and their resolution, such that capacity is lower when greater precision is required to perform the task (Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007). Similarly, when WM precision is measured using continuous recall paradigms, precision decreases linearly with higher WM load, consistent with a model of WM where capacity is limited by overall information load rather than the number of integrated objects in memory (Bays, Catalao, & Husain, 2009; Wilken & Ma, 2004). This has become known as the resource model of WM.

Continuous recall paradigms have further allowed researchers to model the sources of error in WM using a mixture modelling approach (Zhang & Luck, 2008). WM recall is

modelled as a mixture of items that are stored in memory with a certain precision, and items that are not stored in memory, resulting in random guesses. This approach was initially used to support a modified version of the slot model (Slots+Averaging). According to the Slots+Averaging Model, multiple discrete, fixed-resolution slots can be used to represent a single object in cases where it might be helpful, predicting higher resolution representations at smaller set sizes. However, a subsequent study using a similar approach found that a large proportion of the random guesses reported in the original study instead may have originated from mis-bindings between locations and colours in memory, leading participants to mistakenly report the colour of another memory-item ("swap errors")(Bays et al., 2009). When memory for location is taken into account, the results were better explained by a model where resources are flexibly distributed between items in memory depending on their relevance, with parallel resources for separate feature dimensions. Additionally, memory failures for different features belonging to the same WM object occur independently of each other (Bays, Wu, & Husain, 2011; Fougnie & Alvarez, 2011), which would not be expected if features are maintained purely as integrated objects, with discrete, fixed resolution slots.

Finally, WM capacity might be limited by both information load and the number of objects (Alvarez & Cavanagh, 2004; Oberauer & Eichenberger, 2013). There does seem to be a benefit to storing features of the same object in WM relative to storing the same number of features individually, consistent with a role for object-level information in WM (Fougnie et al., 2010; Olson & Jiang, 2002). For example, storing multiple features of the same object comes at a cost to precision, but increasing the number of features across objects imposed additional costs on the number of items encoded into WM, despite long encoding times (1200 ms)(Fougnie et al., 2010). One theory attempting to integrate the evidence for feature-independence and object-based benefits proposed that information in WM is stored as "hierarchical bundles", containing both feature-level and

object-level information at different levels of the cognitive hierarchy (Brady, Konkle, & Alvarez, 2011). If WM representations are hierarchically structured, this may allow for some degree of independence between features at lower levels of the hierarchy, yet incur additional overhead costs for encoding a new bundle compared to adding a feature to an existing bundle. If so, whether feature-level or object-level representations are emphasised in WM may depend on task-demands and behavioural goals (Vergauwe & Cowan, 2015).

3.3. Binding through location?

Another potential source for the object-based benefit in WM is that features belonging to the same object tend to share a spatial location. WM encoding may involve serial allocation of spatial attention to each object in the display (Bays et al., 2009; Desimone & Duncan, 1995; Treisman, 1998). If features in parallel stores are bound in relation to a specific location, there will be fewer such locations to remember in conjunctive conditions—and consequently a lower 'information load'—than in conditions where the same number of features are presented disjointedly. In support of this hypothesis, the object-based benefit in WM disappeared when item-location was no longer informative (Wang, Cao, Theeuwes, & Olivers, 2016). They presented objects either in distinct spatial locations, or serially in the same central location. In the former condition, the recall probe was also presented in a spatially informative location, whereas in the latter the location of the recall probe was non-informative. If the object-benefit arises because fewer locations need to be stored in memory, the benefit of storing conjunctions rather than disjunctions should only be present when location was informative, which was the case.

A special role for location in encoding and retrieving non-spatial features from visual WM is further supported by studies showing that changes to the spatial configuration

between WM encoding and probe interferes with memory for non-spatial features, and bound object-features in particular, even when memory for location is not needed to perform the task (Jiang, Olson, & Chun, 2000; Treisman & Zhang, 2006). Recent studies have also shown involvement of spatial attention in the retrieval of non-spatial features from WM, using eye-tracking and power modulations in EEG to measure allocation of spatial attention (Poch, Capilla, Hinojosa, & Campo, 2017; van Ede, Chekroud, & Nobre, 2019). For example, retrieving an orientation from WM based on a colour probe was associated with small fixational shifts in eye-position toward the original location of the probed object, even though location was not needed to perform the task, consistent with a role of spatial attention in selecting non-spatial features and their bindings from WM.

A recent model of feature-binding in WM proposes that object-features are bound through their shared location (Schneegans & Bays, 2017a). Colour and orientation are encoded in separate noisy populations of neurons with fixed conjunctive coding for each feature and its location. Non-spatial features such as colour and orientation are bound only through their shared location. The authors fit predictions from the model to data from a continuous recall experiment where participants were required to remember colour-orientation conjunctions at specific locations. They varied the recall probe feature (colour vs orientation probe) and the to-be-reported feature(s) (colour and location vs orientation and location). The model replicated behavioural results showing correlated errors for spatial and non-spatial features, meaning that when participants forget an object's location they also tend to forget other features of the object, while memory errors for two non-spatial features do not tend to co-occur (Bays et al., 2011; Fougnie & Alvarez, 2011). Additionally, when a non-target location is mistakenly reported on a trial, reports for the non-spatial feature centred around the value of the feature that appeared at the reported location.

These findings are not easily explained if non-spatial features are bound directly, independent of their location. However, the model also suffers from limitations. Fixed conjunctive coding between spatial and non-spatial features is biologically plausible for low-level visual features such as colour and orientation, but seems less likely for increasingly complex objects or abstract concepts. Finally, although objects encoded at the same location tend to be more easily confused (Pertzov & Husain, 2014), it is nevertheless possible to store and retrieve objects at overlapping locations, which is not easily explained by the model. Some of these limitations may perhaps be overcome by allowing for flexible conjunctive coding (Manohar, Zokaei, Fallon, Vogels, & Husain, 2019; Swan & Wyble, 2014). One example is the binding pool model where combinations of features are flexibly stored and retrieved through a shared pool of neurons facilitating connections between features and object-tokens (Swan & Wyble, 2014).

3.4. Interim summary

Convergent evidence supports independent stores for non-spatial features in WM (e.g. colour, orientation). A consistent benefit to storing features conjunctively rather than individually nevertheless supports a role for object-level representations in WM, perhaps facilitated through spatial binding mechanisms. WM may store hierarchical representations containing a combination of feature-level and object-level information that are emphasised depending on the requirements of the task. Spatial attention seems to play an important role in encoding and retrieving features from WM, supporting a privileged status of location. I will next consider attentional states in WM.

4. Attentional states in WM

4.1. Overlap between attention and WM

At any one time, humans are faced with an abundance of sensory inputs, but our capacity for processing them is limited (Desimone & Duncan, 1995). Attention refers to the process of selecting and filtering information for cognitive processing and may be either top-down (i.e. voluntarily focusing on goal-relevant information) or bottom-up (i.e. involuntary capture of attention by salient features or events)(Corbetta & Shulman, 2002). WM and attention are intimately connected processes—both are crucial for goal-directed behaviour and top-down control and engage a similar frontoparietal 'control' network including overlapping areas of the PFC, frontal eye fields (FEF), and PPC (Corbetta, Kincade, & Shulman, 2002; A. Gazzaley & Nobre, 2012; Nobre et al., 2004). The substantial overlap between WM and attention has led to proposals that WM is best considered as attention directed toward internal representations rather than a separate cognitive system altogether (Chun, 2011; Cowan, 1998; Kiyonaga & Egner, 2013; Oberauer, 2002; Postle, 2006).

At the very least the two systems are strongly intertwined. For example, successful WM seems to rely on attentional refreshing, as studies have found that a dual spatial attention task directing attention away from memorised locations disrupts spatial WM performance (Awh, Jonides, & Reuter-Lorenz, 1998; Smyth & Scholey, 1994). Vice versa, theories of attention consider WM to play an important role in maintaining the template for top-down visual search, thereby biasing competition toward template-matching targets (Bocincova, Olivers, Stokes, & Manohar, 2020; Bundesen, 1990; Olivers et al., 2011). Simply holding content in WM may guide attention toward similar or matching stimuli in the environment, even when detrimental to performance (Kiyonaga & Egner, 2013). Multiple studies have shown that people are faster to identify a target when it matches contents of WM and slower if the display for visual search instead contains

memory-matching distractors (Awh et al., 1998; Downing, 2000; Kiyonaga, Egner, & Soto, 2012; Olivers, Meijer, & Theeuwes, 2006; Soto, Heinke, Humphreys, & Blanco, 2005; Soto, Hodsoll, Rotshtein, & Humphreys, 2008).

However, subsequent research has identified boundary conditions where automatic capture of attention by WM contents does not seem to apply (Downing & Dodds, 2004; Woodman & Luck, 2007). For example, Woodman and Luck (2007) failed to find evidence for attentional capture in cases where the WM content never matched the target for visual search. Instead, participants were actually better at suppressing memory-matching distractors in some cases, suggesting attention may be strategically shifted toward and away from the item in memory. In a task where a new search target was encoded on every trial, along with an item for subsequent memory recall, Downing and Dodds (2004) found no evidence for attentional capture by a memory-matching distractor. It is possible that the search target maintains priority over the memory stimulus when it changes from trial to trial, reducing the influence of the memory item on attentional capture (Kiyonaga & Egner, 2013). Perhaps only a single memory item, relevant for subsequent behaviour, guides attention. When participants maintained multiple items in WM to be recalled in turn after a delay, only the currently relevant item guided attention on a dual visual search task, while the other memory item did not (Peters et al., 2009). This gave rise to the suggestion that not all information is represented equally in WM. There may be multiple states with different downstream effects on behaviour; an "active" or attended state in memory that automatically drives external attention and a "latent" or unattended state that does not (Olivers et al., 2011). Recall that state-based models of WM predict distinct representational states in WM (Cowan, 1998; McElree & Dosher, 1989; Oberauer, 2002). If the focus of attention selects a single item for the next cognitive operation, this item may be in a state that automatically guides attention (Oberauer, 2002). Next, I will consider how attention may

operate at different stages of WM to facilitate goal-directed behaviour, with particular focus on encoding and maintenance.

4.2. Selective encoding of task-relevant information

According to a prominent theory of visual attention, attention acts as a gatekeeper for encoding into WM (Awh & Vogel, 2006; Bundesen, 1990; Bundesen, Habekost, & Kyllingsbæk, 2005). As posited by the biased competition model of attention, features and objects in the environment compete for neural processing (Desimone & Duncan, 1995). Behaviourally relevant stimuli that win the competition for attentional representation are encoded into WM (Bundesen et al., 2005; Bundesen, Habekost, & Kyllingsbæk, 2011).

Consistent with this model, extensive behavioural and neuroimaging studies show that WM selectively encodes information that is most relevant to the forthcoming task. For example, attentional cues to task-relevant items prior to WM encoding ("pre-cues") are associated with top-down modulation by a frontoparietal network of the sensory cortex at early stages of processing (100-300 ms) that correlates with subsequent WM performance (Adam Gazzaley, 2011; Zanto & Gazzaley, 2009). By selectively coding for relevant content, WM may make efficient use of limited capacity to improve the chances of successfully achieving a behavioural goal (Bundesen, 1990; Bundesen, Vangkilde, & Petersen, 2015). Filtering access into WM seems to play an important role in optimising WM processing for current task demands. For example, selective encoding of items relevant to the task, as measured by contralateral delay activity (CDA) in EEG, correlates with WM capacity (Vogel & Machizawa, 2004) and high capacity individuals were more efficient at selectively representing task-relevant objects and filtering out distractors than were low capacity individuals (Vogel et al., 2005).

While the abovementioned studies indicate that WM selectively encodes task-relevant objects, it is actively debated whether feature-based attention similarly facilitates selective encoding of relevant features (see also Section 3). Evidence against selective encoding of features comes from studies showing performance costs when an irrelevant feature is changed between the encoding and probe display in change-detection paradigms "irrelevant-change distracting effects" (Jiang et al., 2000; Shen, Tang, Wu, Shui, & Gao, 2013; Treisman & Zhang, 2006). These suggest that when an object is attended, all of its features are automatically encoded into WM, whether relevant or not; known as object-based encoding. Similarly, memory-driven attentional capture is found even for irrelevant features of objects stored in memory, such as colour (Soto & Humphreys, 2009) and shape (Gao et al., 2016), in support of object-based encoding, but c.f. Olivers et al. (2006).

Other studies have suggested that it is possible to selectively encode only relevant object-features in WM. For example, one study showed that encoding rates into WM and CDA amplitude during the delay differed for identical displays depending on which feature (colour or orientation) of multi-feature objects was task-relevant (Woodman & Vogel, 2008). This would not be expected if encoding of all object-features were obligatory. Other studies have failed to show evidence of irrelevant-change distracting effects with longer delays (Bocincova, van Lamsweerde, & Johnson, 2017; Jaswal & Logie, 2011) or when WM load is high (Xu, 2010). Attentional capture specific to relevant features in WM was also reported in one case (Olivers et al., 2006). Shape-matching distractors only interfered with visual search performance when shape was the relevant WM-feature and colour-matching distractors only interfered with search when colour was the relevant WM-feature. However, this study used a longer delay between WM encoding and visual search (4000 ms), compared to the two studies that did report attentional capture by irrelevant features (188 ms; 190 ms)(Gao et al., 2016; Soto &

Humphreys, 2009). Perhaps initial encoding is object-based after which irrelevant features are either actively suppressed or not refreshed in memory (Bocincova et al., 2017; Jaswal, 2012; Soto & Humphreys, 2009).

Neuroimaging studies show selective coding for task-relevant features during the WM delay (Bocincova & Johnson, 2019; Serences et al., 2009). In one study, participants were instructed to remember either the colour or the orientation of a memory item (Serences et al., 2009). During the WM delay, stimulus decoding from BOLD responses in early visual cortex were only significant for the cued feature dimension, while classification accuracy was at chance level for the irrelevant feature. To investigate the time-course of WM processing for task-relevant vs -irrelevant features, Bocincova and colleagues collected EEG data while participants performed a similar task (Bocincova & Johnson, 2019). While information about both task-relevant and irrelevant features was present during initial WM encoding, processing of the unattended feature was attenuated and largely disappeared during the delay period, suggesting information about task-irrelevant features are subject to time-based decay or are actively removed over time during the delay.

Preferential neural coding for task-relevant features is not inconsistent with some degree of object-based encoding, resulting in irrelevant-change distracting effects. As noted above, it is possible that all features are initially encoded, after which only relevant features are refreshed and/or irrelevant features are actively suppressed. Internal attention to task-relevant features within WM may result in preferential coding for those features, while neural coding for irrelevant features are at chance, as seen for objects (Larocque et al., 2014)(Section 4.4.3). Even so, task-irrelevant features may not be completely lost and are perhaps represented in residual patterns of activity, resulting in irrelevant-change distraction effects, similar to the finding that a probe containing an uncued memory item is rejected more slowly than never encountered probes (Oberauer,

2001, 2002). Intrusion probe effects were present for up to five seconds, which is somewhat longer than the four second-delay at which attentional capture was no longer observed for irrelevant features (Olivers et al., 2006). However, these two studies cannot be directly compared as the former cued number-lists (Oberauer, 2001) and the latter cued visual features (Olivers et al., 2006).

To summarise, the results are broadly consistent with a model in which WM encoding is initially object-based, after which attention acts on WM to refresh relevant features and/or suppress irrelevant features. Relevant features are associated with a stronger neural signature as decodable from BOLD activity and EEG, but memory traces associated with irrelevant features may linger for at least a few seconds. Next, I will review evidence for internal allocation of attention within WM.

4.3. Selective maintenance: retro-cueing

While attention research is traditionally concerned with how incoming sensory information competes for further processing (Desimone & Duncan, 1995), attention can also be directed toward internal mental states to privilege processing of task-relevant information in WM. Initial evidence for internal attentional orienting came in 2003, when two studies showed that spatial cues presented during the retention interval of a WM task indicating which item was going to be probed at recall ("retro-cues") provided behavioural benefits similar to those observed for pre-cues (Griffin & Nobre, 2003; Landman, Spekreijse, & Lamme, 2003). These benefits occurred even though no additional sensory information was available for sampling at the time of the retro-cue. Since then, the retro-cue benefit has been shown to extend to non-spatial cue features, such as colour (A. Heuer & Schubö, 2016; Li & Saiki, 2014) and the retro-cueing paradigm has become widely used in the study of WM (Souza & Oberauer, 2016).

In a review of the literature, Souza and Oberauer (2016) highlighted six potential mechanisms that may operate independently or in concert to produce the retro-cue benefit: 1) protection from time-based decay; 2) prioritisation for comparison with the probe; 3) removing non-cued WM content; 4) strengthening or "refreshing" relevant representations; 5) retrieval head-start; and, 6) protection from perceptual interference. I will briefly summarise the evidence for each of these in turn.

1) Protection from time-based decay: assuming WM content is subject to time-based decay (Baddeley & Hitch, 1974; Barrouillet, Bernardin, & Camos, 2004) a retro-cue may protect the cued item from such decay. Support for this account comes from a study showing that costs on invalid cue trials increase with longer delays, suggesting that representations of non-cued items decay over time (Pertzov, Bays, Joseph, & Husain, 2013). However, protection from temporal decay cannot explain why retro-cues provide a benefit relative to no-cue conditions where the probe is presented at the same point in time as the retro-cue (Makovski, Sussman, & Jiang, 2008; A. M. Murray, Nobre, Clark, Cravo, & Stokes, 2013; Rerko, Souza, & Oberauer, 2014; Souza, Rerko, & Oberauer, 2014, 2015).
2) Prioritisation for comparison with the probe: Assuming that items in memory are searched through serially at recall for comparison with a probe, a retro-cued item may become prioritised and be the first item that is compared to a probe. While this seems feasible in classic change-detection tasks, it might be a less likely strategy in other types of tasks where the item for comparison is explicitly given at the probe (e.g. continuous recall tasks). However, a new version of this model was proposed recently (Myers, Stokes, & Nobre, 2017)(see Section 5.2).
3) Removing non-cued WM content: A predictive retro-cue may allow removal of uncued content from WM, freeing up WM capacity and resources for processing the cued item. This account is supported by studies showing that the retro-cue

benefit is greater when WM load is high and more items can be dropped from memory (Kuo, Stokes, & Nobre, 2012; Souza et al., 2014; Van Moorselaar, Olivers, Theeuwes, Lamme, & Sligte, 2015) as well as studies showing reduced CDA amplitude following retro-cueing (Kuo et al., 2012), which is thought to reflect the number of items stored in WM (Vogel & Machizawa, 2004). This account predicts behavioural costs on invalid cue trials as non-cued items should be dropped from memory, which some have reported (Astle, Summerfield, Griffin, & Nobre, 2012; Griffin & Nobre, 2003; Williams, Hong, Kang, Carlisle, & Woodman, 2013) but others dispute (Li & Saiki, 2014; Myers, Chekroud, Stokes, & Nobre, 2018). For example, Li & Saiki (2014) showed that a retro-cue followed by a withdrawal cue asking participants to again maintain all items produced a retro-cue benefit without associated costs to other items.

4) Strengthening or 'refreshing' cued representations: retro-cues may "refresh" the representation of cued WM content and strengthen the binding between cued items and their context, thereby improving the item's accessibility for use at the probe (Rerko & Oberauer, 2013; Souza et al., 2015). According to this model, strengthening the cued item does not imply that resources are moved away from other items allowing for retro-cue benefits without costs to non-cued items (Li & Saiki, 2014; Myers et al., 2018). It is also consistent with studies showing benefits when participants are simply asked to "think of" or "refresh" items in turn (Souza et al., 2015; Vergauwe & Langerock, 2017; Yi, Turk-Browne, Chun, & Johnson, 2008), cumulative benefits when items are cued in succession (Rerko & Oberauer, 2013), and "incidental cueing" studies where an item cued for an orthogonal task is recalled more precisely (Zokaei, Manohar, Husain, & Feredoes, 2014; Zokaei, Ning, Manohar, Feredoes, & Husain, 2014).

5) Head-start on retrieval: retrieving a retro-cued item ahead of the probe may allow accumulation of evidence for the target to happen in the absence of noise associated with the probe display and cognitive processes related to decision-making. Allowing more time for evidence accumulation might yield more accurate recall, up to a point. On no-cue trials, evidence accumulation might be cut short due to speed-accuracy trade-offs at the probe. Delaying response selection on no-cue trials does yield a benefit, consistent with the retrieval head-start account, although the benefit is smaller than that afforded by retro-cues (Souza, Rerko, & Oberauer, 2016).
6) Protection from perceptual interference: beyond noise associated with decision-making, the probe display itself may also interfere with recall. For example, in a classic orientation recall WM paradigm, performance is worse when the probe display contains a random orientation than when the probe consists only of a coloured dot (Y. A. Tabi, Husain, & Manohar, 2019). As such, a retro-cue might protect the cued item from interference from the probe display at recall (Makovski, Watson, Koutstaal, & Jiang, 2010), or from perceptual distraction during the delay (Makovski & Jiang, 2007). Protection from interference could also be an ancillary effect of strengthening, or retrieval head-start.

In summary, there is evidence that the retro-cue benefit is supported by multiple, overlapping sources, perhaps to varying degrees in different paradigms. Importantly, attentional shifts following retro-cues only tend to last 300-500 ms (Rerko et al., 2014; Souza et al., 2016), after which the cued content remains in a privileged state, highlighting that sustained attention is not necessary to produce the cueing benefit. Souza et al. (2016) suggested this might be the case if transient attention towards the cued item strengthens the feature-context binding, thereby improving the item's

accessibility in memory and protecting it from interference (Oberauer & Lin, 2017; Rerko & Oberauer, 2013). Another recent theoretical perspective suggests a two-step model for retro-cued prioritisation (Myers et al., 2017): first, attention is directed to select the cued item from memory, after which the cued content is reformatted into an action-oriented state in preparation for the probe. I will discuss evidence for this model of retro-cued prioritisation in more detail in Section 5.2.

4.4. Neural states of representation in WM

Orienting attention in perception and in WM relies on partially overlapping neural mechanisms. Next, I will discuss neural correlates of attentional orienting in WM and the ways they resemble or differ from mechanisms of external attention.

4.4.1. Alpha lateralization

Oscillatory activity in the alpha frequency band (9-14 Hz) is the most prominent signal measured in EEG and MEG (Jensen & Mazaheri, 2010). Synchronization is particularly strong over parietal-occipital areas and has been thought to reflect gating of visual attention. For example, in a multimodal attentional task, alpha power over posterior electrodes was especially strong prior to stimulus onset when participants were instructed to attend to the auditory component of the stimulus and ignore the visual modality, relative to the inverse condition (Foxe, Ca, & Ahlfors, 1998). Increased alpha power during the attend-sound condition may reflect top-down gating or suppression of irrelevant visual input, while the inverse may reflect active engagement in the visual task.

Further support for a role of alpha power in the deployment of visual attention comes from a spatial cueing study (Worden, Foxe, Wang, & Simpson, 2000). An attentional cue to attend either the left or right portion of the visual field was associated with alpha

band desynchronization that was strongest in the hemisphere contralateral to the attended location, indicating that oscillatory activity in the alpha band tracked the locus of spatial attention. While the precise neural mechanisms contributing to alpha synchronization and desynchronization remain uncertain, alpha power may play an active role in guiding neural information flow by inhibiting task-irrelevant areas when synchronized (Bonnefond & Jensen, 2012; Foxe et al., 1998; Jensen & Mazaheri, 2010; Lopes da Silva, 1991; Vanni, Revonsuo, & Hari, 1997) or increasing neural excitability in relevant neural populations when desynchronized (Bollimunta, Mo, Schroeder, & Ding, 2011; Haegens, Nácher, Luna, Romo, & Jensen, 2011).

Alpha power has similarly been shown to track attentional orienting within WM (Myers, Walther, Wallis, Stokes, & Nobre, 2015; Poch, Campo, & Barnes, 2014; Wallis, Stokes, Cousijn, Woolrich, & Nobre, 2015). One study employed an orientation recall WM task to directly compare the dynamics of spatial attention during encoding and maintenance using EEG (Myers, Walther, et al., 2015). Spatial pre-cues before the onset of a WM array and spatial retro-cue during the delay were both associated with contralateral alpha desynchronization relative to the cued position. However, while pre-cues induced sustained lateralization prior to array onset and mostly throughout the delay, lateralization induced by the retro-cue was transient. Similar results were reported by Wallis et al. (2015) using MEG. These findings suggest spatial attention is transient following the retro-cue and are consistent with reports that sustained attention is not required for retro-cueing benefit (Rerko & Oberauer, 2013; Rerko et al., 2014; Souza et al., 2015).

4.4.2. Top-down modulation of sensory regions

Studies of external attention have shown that neural activity in early visual cortex is modulated by attention, such that specialized neural areas respond more strongly to

attended than unattended objects and features (Corbetta, Miezin, Dobmeyer, Shulman, & Petersen, 1991; Desimone & Duncan, 1995; Kastner, De Weerd, Desimone, & Ungerleider, 1998; Moran & Desimone, 1985). Putative sources of such attentional control have been identified in frontoparietal areas including the ventrolateral PFC, inferior frontal junction and frontal eye fields (Bichot, Heard, DeGennaro, & Desimone, 2015; Bichot, Xu, Ghadooshahy, Williams, & Desimone, 2019; Heinen, Feredoes, Weiskopf, Ruff, & Driver, 2014; Moore & Fallah, 2004; Rossi, Bichot, Desimone, & Ungerleider, 2007; Zanto, Rubens, Bollinger, & Gazzaley, 2010). For example, attentional enhancement in the early visual cortex is associated with increased functional connectivity between inferior frontal junction and relevant areas of visual cortex (Zanto et al., 2010).

Increased functional connectivity between a network of frontoparietal regions and sensory cortices has similarly been reported during attentional orienting in WM (de Vries, van Driel, Karacaoglu, & Olivers, 2018; Kuo, Stokes, Murray, & Nobre, 2014; Kuo, Yeh, Chen, & D'Esposito, 2011; Lepsien & Nobre, 2007; Wallis et al., 2015). In one experiment, participants remembered four shapes one of which was retro-cued as relevant (Kuo et al., 2014). At recall, participants were presented with the same four shapes and decided whether the cued shape had switched location. Following retro-cueing, BOLD activity was elevated at the retinotopic location in V1-V4 where the cued object had been presented, compared to unattended locations in memory and the degree of modulation in V4 predicted subsequent reaction times (Kuo et al., 2014). Additionally, functional connectivity between right inferior frontal sulcus and the retinotopic locations in V4 was greater for attended than unattended locations, consistent with a role for top-down modulation of sensory cortex to enhance processing of attended locations in memory.

Switching the priority status of items in WM may be supported by top-down modulation of posterior alpha power by frontal delta-theta oscillations (2-6 Hz)(de Vries et al., 2018). Participants encoded two colour templates presented laterally into WM for two sequential visual searches during EEG. Lateralized alpha suppression shifted according to which template was currently relevant to visual search, consistent with a role for alpha in gating internal attention to select the relevant search template. Crucially, frontal delta-theta power predicted alpha lateralization amplitude following a priority switch and also predicted faster target detection on the second search, supporting a role for top-down delta-theta in switching attention and priority status between items in WM.

Consistent with results from a meta-analysis of fMRI studies, a MEG study showed that WM pre-cues and retro-cues engaged a common frontoparietal control network (Wallis et al., 2015). Theta power was enhanced in the anterior middle frontal gyrus (comprising dorsolateral PFC) and frontal eye fields, and alpha power was suppressed in the intraparietal sulcus, overlapping with lateralized alpha suppression in the occipital cortex. However, the authors also identified differences in the spatiotemporal neural signatures associated with external and internal attentional orienting. For example, cross-temporal correlation analyses suggested neural states were more stable for pre-cues and more dynamic for retro-cues. Additionally, retro-cues were uniquely associated with activity in the pre-supplementary motor area and frontal operculum, consistent with a previous meta-analysis of internal attention (Nee et al., 2013) and perhaps indicative of "output gating" (i.e. selecting a WM item for guiding behaviour)(Chatham, Frank, & Badre, 2014; Tamber-Rosenau, Esterman, Chiu, & Yantis, 2011)(see Section 5.2).

4.4.3. Preferential neural coding for relevant WM content

A wealth of research has shown that stimulus-selective neural areas selectively code for relevant information in WM. As discussed in Section 4.2, pre-cues may allow participants to preferentially encode relevant items or features into WM (Bocincova & Johnson, 2019; Serences et al., 2009; Zanto & Gazzaley, 2009). Retro-cues similarly modulate neural activity coding for relevant vs irrelevant objects in memory (Lepsien & Nobre, 2007; Lepsien, Thornton, & Nobre, 2011). In a dual serial retro-cueing task (DSR), a retro-cue informed participants whether a face or a scene in WM would subsequently be probed, followed by a 'stay' or 'switch' cue instructing them to continue maintaining the cued item or switch to the other memory item (Lepsien & Nobre, 2007). In the second delay, BOLD activity in the parahippocampal gyrus (PHG; selective for scenes) and fusiform gyrus (FG; selective for faces) was differentially modulated depending on the type of cue: PHG activity was significantly higher for scene-stay and face-switch trials than scene-switch and face-stay trials and vice versa for activity in FG. Switch cues, but not stay cues, were associated with engagement of a frontoparietal network, presumably reflecting involvement of top-down control in switching attention between items in memory.

The above study contrasted univariate BOLD activity for attended vs unattended conditions, but was not able to assess the extent to which attended and unattended items in WM were represented in neural activity (Lepsien & Nobre, 2007). This was the aim of two subsequent studies–one using fMRI (Lewis-Peacock et al., 2012) and the other EEG (LaRocque et al., 2013). Both employed a similar DSR task, but used multivariate pattern analyses to test the extent to which each category in memory could be decoded from neural patterns of activity when it was in an attended vs unattended state. Initially following encoding, both categories in memory could be decoded. Following a retro-cue, classification accuracy remained significant for the cued category,

but dropped to chance level for the uncued category. On stay trials, classification accuracy for the uncued item remained at chance until probe onset, but re-emerged on switch trials and instead dropped to baseline for the other (initially cued) item. Even though both items were maintained in WM during the first delay, neural activity exclusively coded for the currently relevant item. These results suggest that WM content inside and outside the focus of attention may be supported by different representational states, where only the former is decodable from neural activity (Larocque et al., 2014).

However, these conclusions relied on category-level representations (LaRocque et al., 2013; Lepsien & Nobre, 2007; Lepsien et al., 2011; Lewis-Peacock et al., 2012). This leaves open the alternative possibilities that participants broadly attended the relevant category and not the specific memory item, or that neural decoding for the attended category reflected anticipatory attention for the probe category rather than the memory representation itself. These limitations were ruled out in a study using an inverted encoding model to reconstruct spatial locations maintained in WM from fMRI voxel patterns (Sprague et al., 2016). Participants encoded two locations on every trial and on separate trials were either cued to one item immediately following encoding (Remember 1), later during the delay (Retro-cue) or at the probe (Remember 2). In the first delay, reconstruction quality was good for the cued item's location in the Remember 1 condition and for both locations in the other two conditions, but representations degraded over the course of the delay. Following the retro-cue, reconstruction quality was restored for the cued location on retro-cue trials, but continued to degrade on Remember 1 and Remember 2 trials. The results suggest that representations in WM – as reconstructed from neural activity patterns – track item relevance, but degrade over time. However, representational quality can be restored by focusing attention toward a single item in memory, in a way that mimics the behavioural benefits of retro-cueing. These findings

have since been reproduced for directions of motion and orientations in WM (Larocque, Riggall, Emrich, & Postle, 2017; Wolff, Jochim, Akyürek, & Stokes, 2017).

4.4.4. Fate of unattended items in WM

While the neural representational quality for unattended items in memory is poor (LaRocque et al., 2013; Lewis-Peacock et al., 2012; Sprague et al., 2016; Wolff et al., 2017), representational quality can be re-established by a retro-cue labelling a specific item as task-relevant (Sprague et al., 2016). This led to preliminary suggestions that only attended content in WM is stored in neural activity space, while unattended memory content may be represented in a latent or "activity-silent" neural state, for example through temporary changes in synaptic plasticity (Fiebig & Lansner, 2017; Kamiński & Rutishauser, 2020; Manohar et al., 2019; Mongillo, Barak, & Tsodyks, 2008; Stokes, 2015). An influential computational model demonstrated how activity-silent WM could be supported by short-term synaptic facilitation, through a transient increase in presynaptic calcium concentration following neural firing (Mongillo et al., 2008). Subsequent activation of the network will reflect the code established for the memory stimulus during perception as potentiated synaptic connections reactivate the same population of neurons through pattern completion. While traditional models of WM posited that information was represented in persistent activity in selective populations of neurons, this model suggests information may be maintained through synaptic mechanisms in the absence of sustained activity.

Recent studies have provided indirect support for the presence of latent states in WM for unattended items (Rose et al., 2016; Wolff, Ding, Myers, & Stokes, 2015; Wolff et al., 2017). If memory for unattended content is supported by activity-silent synaptic traces, these might be revealed by stimulating the system in a way that would drive latent connections to become active, such as a strong visual impulse (for visual WM) (Wolff et

al., 2017) or transcranial magnetic stimulation (TMS) (Rose et al., 2016). In a combined EEG-TMS study, participants performed a version of the DSR task in which they encoded two memory items from three possible categories (faces, words and motion direction) on each trial (Rose et al., 2016). During the delay, participants received a retro-cue and were probed about one item, after which a second retro-cue either cued the same item again ("stay") or cued the other memory item ("switch"). Participants received a TMS pulse following each cue. Replicating previous studies, the cued item could be decoded from EEG activity patterns in the first delay, while classification accuracy dropped to baseline for the uncued item. Crucially, the TMS impulse brought back decoding for the unattended item temporarily, specifically when TMS targeted the portion of cortex selective for the unattended item's category. Interestingly, this was not the case after the second retro-cue when the unattended item could be dropped from memory.

Similar reinstatement of neural decoding from posterior EEG channels was revealed following a strong visual impulse (Wolff et al., 2017). Participants encoded two oriented gratings on each trial to be probed in turn and were informed of the probe order before the start of each block of trials. During the delay prior to the first probe, decoding accuracy remained significant for the currently relevant memory item to be tested early, but dropped to baseline for the item to be tested later, consistent with previous results (LaRocque et al., 2013; Lewis-Peacock et al., 2012). A strong visual impulse during this delay triggered an increase in decoding accuracy for both memory items, though more strongly for the currently relevant item. During the second delay following the first probe, the impulse triggered re-emergence of decoding only for the tested-late item that was now relevant, while the tested-early item that was no longer relevant could not be decoded. The authors proposed that neural activity codes for the currently relevant items in memory, while unattended items are stored in a flexible, hidden state. This hidden

state may be revealed by a strong visual impulse that drives the visual system to "echo" latent patterns coding for unattended WM content.

An inherent limitation of studies proposing activity-silent mechanisms for maintenance of unattended content is that they are based on null-results of non-significant decoding, specifically in visual cortex. It is possible that unattended content is coded in neural activity patterns, but that these patterns are attenuated compared to patterns for attended content, resulting in non-significant decoding from a noisy neural population (Schneegans & Bays, 2017b) or that these patterns exist in other areas of cortex. If so, it should be possible to detect activity patterns coding for unattended WM content with greater statistical power.

To boost power, one study collected fMRI data from a large number of participants (n=87) while they performed a version of the DSR task with orientation stimuli (Christophel et al., 2018). Replicating previous results, the authors found that only currently relevant memory items were decodable from visual cortex (V1-V4). However, the unattended memory item could be decoded from other areas of cortex, specifically the intraparietal sulcus (IPS) and frontal eye fields (FEF). The authors proposed that visual cortex may store high-resolution representations of currently relevant content, while posterior parietal cortex (PPC) may store lower-resolution representations of both attended and unattended content.

This reflects results discussed in Section 2.2 showing that visual distractors eliminated WM-related activity patterns in visual cortex, while PPC continued to code for WM-content under distraction (Bettencourt & Xu, 2016), but c.f. Rademaker et al. (2019) who found that representations of mnemonic content and distractors co-existed in sensory cortex. However, the degree of neural interference from distractors in sensory cortex correlated with behavioural precision, consistent with a hypothesis that sensory cortex

may store high-resolution representations, supplemented by distractor-resistant, low-resolution mnemonic representations in PPC and PFC (Rademaker et al., 2019). Storing WM content in a different area of cortex in anticipation of distraction may be an efficient mechanism for protecting WM representations from perceptual interference, which may be more likely if both perceptual and mnemonic representations are represented in the same areas of cortex. In WM, it may similarly be advantageous to separate representations for attended and unattended content, since only attended content is immediately relevant and should drive behaviour (Olivers et al., 2011; Olivers & Roelfsema, 2020).

An alternative mechanism by which currently relevant and prospectively relevant WM representations may be separated is if these are coded in opposite and/or anti-correlated patterns of activity within the same area of cortex (van Loon, Olmos-Solis, Fahrenfort, & Olivers, 2018). During fMRI, participants were presented with an object from one of three categories to remember for a visual search task on each trial. This object was either relevant to a first visual search task (current) or to a second search task (prospective). The other visual search involved a stable target across all trials. In the first delay, the object category in WM could be classified from posterior fusiform cortex (selective for objects) regardless of priority status, though decoding was better for current compared to prospectively relevant object categories. Interestingly, during the first search, the prospectively relevant object category could also be decoded. However, using a cross-generalisation approach, training on currently relevant object-categories and testing on prospectively relevant categories, classification accuracy was negative for the prospectively relevant item. This suggests the representational format of the prospectively relevant object category was inverted compared to the currently relevant category.

Similar evidence for priority-dependent "re-mapping" of the representational format was reported for parametric stimuli in WM (Q. Yu, Teng, & Postle, 2020). In a DSR task, participants were presented with two orientations sequentially, each in one of nine possible locations. Using an inverted encoding model trained on the attended item orientations, they could decode the orientation of both the attended and unattended memory item from visual cortex. However, the representational format of the unattended item's orientation was inverted relative to the attended memory item. The representational format for the unattended item's location was similarly inverted in IPS, but not in the visual cortex. The priority-dependent representational reformatting reported in these two fMRI studies could serve as a flexible mechanism for separating representations of content that is relevant for current vs future goals (van Loon et al., 2018; Q. Yu et al., 2020).

Beyond separating currently relevant from prospectively relevant representations in WM, representational re-formatting may serve as a mechanism for transforming *selected* representations into a neural subspace for guiding behaviour (Panichello & Buschman, 2020). Non-human primates performed a continuous recall WM task involving a retro-cue to one of two possible colours in WM, encoded at different locations, while single-unit activity was recorded from neurons in the lPFC, FEF, lateral intraparietal area (LIP) and visual cortex. While the colour of memory items presented at different locations were represented in orthogonal neural codes before retro-cue onset—perhaps a useful mechanism for preventing interference between items in memory—after cue-onset the colour of selected representations was aligned into a common neural subspace regardless of their original location. Thus, selection in WM may involve re-configuration of the representation into a subspace used for guiding behaviour.

In contrast to the abovementioned human neuroimaging studies, the relative coding scheme for the two colours in memory was *orthogonal* before the cue, rather than

inverted. However, the experiments were designed to test different, though related, questions: the neuroimaging studies assessed the relative coding scheme of prioritised vs unprioritised items after one was selected as currently relevant. Panichello and Buschman (2020) assessed the relative coding scheme of two stimuli presented at different locations *before* cueing (*orthogonal* codes) and after each was selected as relevant to decision-making (*parallel* codes), but there was no 'prospectively relevant' item. Consequently, the two types of rotational dynamics are not necessarily inconsistent. It is possible that two memory items are represented in orthogonal subspaces prior to selection. Selection may configure a relevant item into a common subspace used for guiding behaviour, regardless of the item's original position (Panichello & Buschman, 2020). If another item continues to be relevant to a future decision, the currently unprioritised item may be represented in an inverted coding format relative to the prioritised item (van Loon et al., 2018; Q. Yu et al., 2020). Together, these studies show that the same memory content may be represented in different neural formats depending on its priority status in WM. However, the precise role of these coding schemes in preventing interference between items in memory and/or translating the memory representation into an appropriate action remains a question for future research (Stokes, Muhle-Karbe, & Myers, 2020) (see Section 5.3).

4.5. Neural models of attention in WM

Traditionally, computational models of WM have not incorporated a focus of attention and consequently do not explain how WM contents may be maintained in different states depending on their current relevance. For example, networks with recurrent connectivity that sustain sensory activity within a stable attractor state (Compte et al., 2000; X. J. Wang, 2001; Wimmer et al., 2014) are successful at modelling short-term maintenance of individual features, but do not include mechanisms for selective attention. Recent computational models have built on traditional attractor models to incorporate

mechanisms for attentional orienting within WM. Next, I will introduce two examples of such models with different proposed model architectures and different neural and behavioural predictions (Figure 2).

4.5.1. Model with persistent activity and synaptic plasticity

This computational model relies on a combination of neural activity and synaptic plasticity to support flexible attention between items in WM (Manohar et al., 2019)(Figure 2A). The model architecture consists of two layers: a sensory layer and a conjunctive layer. The sensory layer is made up of separate feature stores coding for different feature dimensions (e.g. colour, location and orientation). Neurons in the sensory layer have fixed selectivity such that they respond only to features within their receptive fields. The feature stores within the sensory layer are not directly connected to each other, but all connect to the conjunctive layer. Neurons in the conjunctive layer have flexible selectivity and because connection weights between the feature- and conjunctive layers are initially random, neurons in the conjunctive layer compete to represent feature-conjunctions through lateral inhibition. The winning conjunctive neuron codes for the combination of active features and after stimulus offset persistent activity is maintained through rapid Hebbian plasticity that establishes reciprocal excitatory connections between simultaneously active neurons. In this way, fast plasticity enables flexible, conjunctive coding and maintenance in WM. When a new object is presented, the activity pattern is disrupted and a new conjunctive representation is formed, but the modified synaptic weights continue to represent the previous object. The number of neurons in the conjunctive layer determines WM capacity as they correspond to the number of objects that can be maintained in WM at any one time. If the WM set size exceeds the number of conjunctive units, a “used” conjunctive unit will be employed to represent new input and thus overwrite previously established plasticity. Consequently, forgetting in the model occurs by interference (Oberauer,

Farrell, Jarrold, & Lewandowsky, 2016; Oberauer & Lewandowsky, 2008). An item stored in WM may be retrieved by a partial recall cue. The cue will turn on the appropriate sensory neuron activating the relevant conjunctive unit through the strengthened connections, which in turn reactivates the remaining object-features by synaptic pattern completion.

Importantly, the model makes distinctions between items that are represented inside and outside the focus of attention. Items represented in persistent activity are in the "focus of attention" while items outside the focus of attention are maintained in latent, "activity-silent" traces of strengthened synaptic connectivity. Actively represented items are more accessible and less susceptible to interference, as they do not depend on "re-activation" through pattern completion, but can be read out directly from active neurons. Since persistent activity is not disrupted for the last item in a sequence, this item continues to be in the focus of attention, explaining behavioural recency effects (McElree & Dosher, 1989). Similarly, a retro-cue may bring an item into the focus of attention by re-establishing the active traces for an item in memory, as during cued recall. Consequently, the retro-cue benefit in the model is the combined result of three mechanisms introduced previously: strengthening/refreshing, retrieval head-start and protection from perceptual interference (Souza & Oberauer, 2016).

Neurally, the model predicts neural decoding of attended items from the sensory layer, but relatively poor neural decoding from the conjunctive layer since the same units may represent different feature conjunctions on different trials, consistent with sensory recruitment as well as distributed coding (Christophel et al., 2017; Harrison & Tong, 2009; Serences et al., 2009). Unattended items are not generally represented in neural activity patterns, but may be brought back into an active state by retro-cueing or a strong neural impulse echoing latent neural traces (Rose et al., 2016; Sprague et al., 2016; Wolff et al., 2017). Interestingly, although features of an unattended item are not

represented in neural activity, the model does predict some level of neural decoding of unattended items in the sensory layer, but in a reversed neural code (van Loon et al., 2018; Q. Yu et al., 2020). In the model, this occurs because lateral inhibition between conjunctive units spreads to feature neurons that form part of a previously established memory engram with strengthened synaptic traces.

While the model accounts for several behavioural and neural correlates of flexible attention within WM, it also has certain limitations. First, as discussed in previous sections, attention is not necessarily sustained following retro-cueing (Myers, Walther, et al., 2015; Wallis et al., 2015) and the retro-cue benefit survives even when attention is subsequently directed away from a cued item (Rerko & Oberauer, 2013; Rerko et al., 2014). The model predicts that a retro-cued item remains in an active focus of attention, but perhaps cueing strengthens the neural traces, constituting binding between features, resulting in a benefit even if attention is subsequently directed away (Souza & Oberauer, 2016). Second, in the model, a retro-cue works by re-activating neural activity patterns through traces established during perception. However, the neural pattern coding for a stimulus during perception does not necessarily cross-generalise to neural activity patterns coding for the same stimulus during a WM delay or following retro-cueing (i.e. "dynamic coding")(King & Dehaene, 2014; Panichello & Buschman, 2020; Spaak et al., 2017; Stokes et al., 2013; Wolff et al., 2015). These limitations are not unique to this model and similarly apply to other models where WM content is represented in persistent activity in the same neural populations involved in perception, including the model described below (Compte et al., 2000; Schneegans & Bays, 2017b; Wimmer et al., 2014). Finally, the model predicts no privileged status of location in WM (Jiang et al., 2000; Pertzov & Husain, 2014; Schneegans & Bays, 2017a; Treisman & Zhang, 2006; B. Wang et al., 2016). However, this assumption of the model could be relaxed, for example by introducing retinotopic coding of features in the sensory layer (Bocincova et al., 2020).

4.5.2. Conjunctive population code model with persistent activity

This model suggests a potential mechanism for attentional orienting in WM in the absence of activity-silent mechanisms (Schneegans & Bays, 2017b)(Figure 2B). Instead the model represents colour-location conjunctions in a population code with self-excitation and lateral inhibition that is able to sustain activity in an attractor state following stimulus offset (X. J. Wang, 2001). This architecture is similar to previous attractor models of WM with recurrent connectivity (Compte et al., 2000; X. J. Wang, 2001), but instead of representing a single stimulus feature it has fixed colour-location conjunctive coding such that each colour (two in this case) is represented in a separate neural field with retinotopic selectivity. In addition to local excitation and lateral inhibition within each colour field, the model further includes global inhibition over both colour fields, introducing competition between memory items. At recall, the model decodes features from the population activity and the item most similar to the cue in the cue-dimension is selected, producing its conjunctive feature value as the response. Memory errors in the model arise due to sub-optimal encoding, drift in the attractor state or decay in the neural representation due to noise. A similar network architecture with additional fields for conjunctive coding between orientation and location was proposed to account for feature binding in WM through shared spatial location (Schneegans & Bays, 2017a)(see Section 3.3). Consequently, to retrieve any non-spatial feature, the model will always retrieve its corresponding location. However, it does not predict auto-associative pattern completion between object-features like the model described above (Manohar et al., 2019).

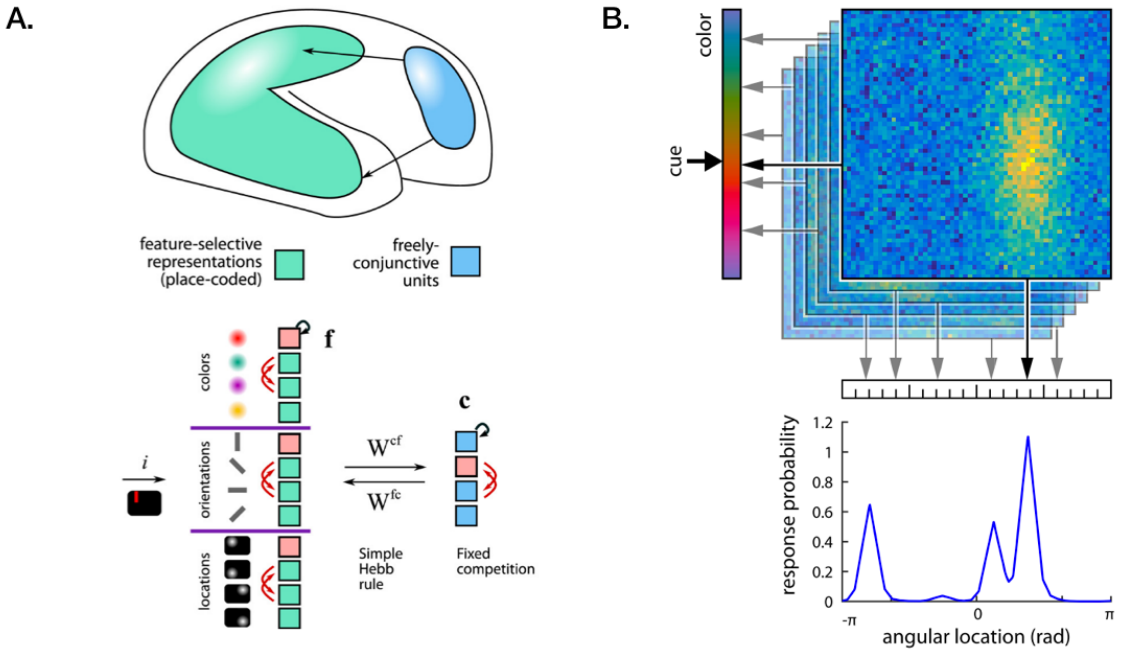

Figure 2. Models of attention in WM. **A.** Single architecture with persistent activity and synaptic plasticity (diagram reproduced from Manohar et al., 2019). Feature-selective units (green) encode stimulus features and freely conjunctive units (blue) compete to represent feature conjunctions. Attended features are represented in active units (red). If a new item is encoded, lateral inhibition will extinguish activity from the previously encoded item, but short-term plasticity (W) between feature and conjunctive units continues to represent unattended items in patterns of connectivity. Items may be retrieved through pattern completion following cued recall. **B.** Conjunctive population code with persistent activity (diagram reproduced from Schneegans & Bays, 2017a). A feature is coded in a noisy population of neurons with fixed conjunctive coding between non-spatial features (e.g. colour) and location. Items encoded in distinct population fields compete for neural resources and competition may be biased by top-down attention to increase signal-to-noise for attended relative to unattended items. At recall, the model decodes features from the population activity and the item most similar to the cue is selected, producing its conjunctive feature value as the response.

The model was specifically used to simulate the experimental finding that a retro-cue may re-instate latent representations in WM (Sprague et al., 2016). Two colours in two spatial locations were encoded on each trial and a retrospective colour cue informed participants which item to remember for subsequent spatial recall. In the model, attentional orienting within WM occurs through biased competition, similar to the prominent theory of attention in perception (Bundesen, 1990; Desimone & Duncan, 1995). When the two items are encoded into WM they compete for representation and decay over time due to noise in the neural population. An early cue following encoding (Remember 1 condition) may bias neural competition toward the cued item, resulting in better spatial reconstruction from the model, while the representation for the uncued

item degrades. In the retro-cue condition, a colour cue may strengthen the relevant spatial representation by means of the conjunctive colour-location code and thereby bias neural resources toward the attended memory item, enhancing the representation of the cued item at the cost of the uncued item. Consequently, location reconstruction after the retro-cue was comparable to when a single item was maintained (Remember 1). When both items are relevant (Remember 2), resources are evenly distributed between the two items throughout the delay, resulting in worse reconstruction of spatial locations in the model overall compared to the other two conditions. Thus, the model is able to reproduce effects of retro-cueing on neural decoding from patterns of BOLD activity.

Beyond limitations of fixed conjunctive coding covered in Section 3.3., an important limitation of the model is that the retro-cue benefit is achieved purely by re-allocating neural resources away from uncued item representations and toward the representations of the cued item, reducing noise and protecting it from time-based decay. This can be thought of as the neural equivalent of the "removal" hypothesis of retro-cueing, introduced in Section 4.3 (Souza & Oberauer, 2016). While removal of uncued items may be one source of the retro-cue benefit, it cannot explain why retro-cue benefits do not always come at the expense of other items in memory (Li & Saiki, 2014; Myers et al., 2018). Because retro-cueing in the model relies on resolving competition from uncued items it also cannot easily explain why non-predictive cues support behaviour when uncued items cannot be dropped from memory (Souza et al., 2015; Zokaei, Ning, et al., 2014) or how retro-cue benefits remain even after attention is subsequently directed away from a cued item (Rerko & Oberauer, 2013; Rerko et al., 2014).

4.5.3. Interim summary

In summary, both models account for supposed latent states in WM as decoded from EEG and BOLD activity. In one model, latent WM states are supported by short-term

synaptic plasticity and may be re-instated by a feature-cue that activates the entire population coding for the cued item through pattern completion mechanisms (Manohar et al., 2019). In the other, seeming latent states are coded in neural activity patterns but are below the detection-threshold of multivariate decoding due to reduced signal-to-noise (Schneegans & Bays, 2017b). In their current form, both models struggle to account for retro-cue benefits in the absence of sustained attention, though synaptic plasticity between features and conjunctions in the former may support consolidated memory traces for a cued item even after attention is removed (Manohar et al., 2019; Rerko & Oberauer, 2013). A third model similarly suggests attention strengthens the bindings between an item's features, but remains ambivalent as to whether the neural mechanisms supporting binding are synaptic (e.g. short-term Hebbian plasticity) or active (e.g. conjunctive coding)(Oberauer & Lin, 2017).

5. Prospective WM

In the previous section, I discussed how attention may modulate what is represented in WM depending on current goals. This demonstrates that, beyond representing the past, a core purpose of WM is to guide future behaviour (Nobre & Stokes, 2019). Even so, prospective aspects of WM functions have traditionally received less consideration in the literature relative to the retrospective mechanisms by which the brain maintains high-quality representations of sensory history (Compte et al., 2000; Goldman-Rakic, 1995; Harrison & Tong, 2009; Serences, 2016). While this literature has been pivotal in shaping our understanding of the neural mechanisms supporting short-term information storage, a singular focus on sensory-like representations risks missing important mechanisms by which WM guides flexible behaviour. Next, I will discuss theoretical and empirical work in support of prospective coding in WM.

5.1. Task-dependent modulation of WM content

It is well-established that WM preferentially codes for objects and features that are relevant to the forthcoming task while filtering out task-irrelevant information (Bundesen et al., 2005; Serences et al., 2009) (see Sections 4.2 and 4.4.3). This may be an efficient mechanism for ensuring that limited cognitive resources are devoted towards processing information that is most likely to help us achieve a behavioural goal. As a consequence, WM representations for the same external objects may look different depending on which object-features are most important to perform the desired task and therefore attended, as demonstrated in humans using fMRI (Lee et al., 2013; Mack et al., 2020; Serences et al., 2009).

Similar attentional modulation of the neural code for WM content has been found in studies of non-human primates. In a delayed paired-associate task, monkeys learned to associate each sample image with a different target image (Rainer, Rao, & Miller, 1999). The monkeys were presented with a new sample on each trial and, after a delay, then presented with a test image that either matched the associated target or not. Even though the target image was not presented at encoding, neural activity coded for the anticipated target stimulus during the delay, suggesting WM flexibly represents the content relevant for decision-making, rather than a record of sensory history.

Similarly, task context (recall vs recognition) affected neural selectivity for two images in memory, even though the trial sequence was identical in the two tasks until probe onset (Warden & Miller, 2010). In the recall task, monkeys sequentially saccaded to two memorised images at the probe, whereas in the recognition task the monkeys had to respond when a matching sequence was presented. During the delay of the recall task, neural representations in the monkey PFC preferentially coded for the first image in the sequence, while the second image was preferentially represented during the delay of the

recognition task, suggesting WM flexibly prioritised the WM item most relevant to the expected task.

The literature reviewed above provides strong evidence that attention flexibly selects objects, features and/or level-of-abstraction for representation in WM depending on their prospective relevance to a behavioural goal. But beyond selecting *which* information is represented, evidence is starting to emerge that WM may also modulate *how* the same relevant task-feature is represented depending on the expected task-demands at retrieval.

A large number of neurons in the PFC display non-linear mixed selectivity, rather than pure selectivity to a single task variable such as task type or specific images encoded into WM (Rigotti et al., 2013). A re-analysis of the data from Warden and Miller (2010) revealed that neurons responded nonlinearly to combinations of variables. For example, one neuron fired specifically when a certain picture was presented in the second position during the recognition task, but not during the recall task. In this way, complex task information may be distributed across a population of neurons, resulting in high dimensional representations. High dimensional representations have the advantage that they allow simple linear read-out of responses that depend on combinations of task-variables in a way that is not possible in a population with low-dimensional representations consisting of neurons with classical or linear mixed selectivity (Fusi et al., 2016). In line with this prediction, the dimensionality of population activity in the monkey dorsolateral PFC predicted behavioural errors on the task, with error trials displaying lower dimensionality (Rigotti et al., 2013). In this way, neurons with non-linear mixed selectivity may allow the same feature to be represented in a different pattern of neural activity depending on the task context.

Another line of evidence indicating that WM modulates how the same features are stored depending on task-goals comes from studies showing that templates for visual search are systematically biased depending on the anticipated visual search context (Geng, Di Quattro, & Helm, 2017; Navalpakkam & Itti, 2007; X. Yu & Geng, 2019). For example, a memory representation of a target colour for visual search was shifted in the opposite direction of anticipated distractors in the visual search display when these were linearly separable from the target (X. Yu & Geng, 2019). When target-distractor similarity was high, the memory representation was asymmetrically sharpened. Thus, the memory template stored in WM for visual search was not a true representation of the sensory input, but was shifted and sharpened to increase the distinctiveness of the template, thereby optimising the template for the current context (Navalpakkam & Itti, 2007). An fMRI study demonstrated that biased search templates might be established through adaptive allocation of attentional gain to neurons in early visual cortex that are selective to the part of feature-space most instructive to discrimination judgements in the current task (Scolari, Byers, & Serences, 2012). Together, the results indicate that the format of WM representations depends not only on the sensory history but also on expectations for future use.

5.2. Working memory for action

Another way in which memory representations could be optimised for the current task is by coupling the representation with the anticipated motor response, which are arguably confounded in some research purporting to study memory (Funahashi, 2015; Funahashi et al., 1989; Ikkai & Curtis, 2011; Takeda & Funahashi, 2002). In a modified delayed saccade task, in which monkeys saccaded to a target 90 degrees rotated relative to a memorised location, delay activity in the PFC coded for both the retrospective visual information as well as prospective motor planning (Takeda & Funahashi, 2002). In humans, a recent study similarly showed that when possible, the brain will link a

sensory representation to the appropriate motor plan in WM (van Ede, Chekroud, Stokes, & Nobre, 2019). During EEG, participants performed a continuous recall WM task that instructed them to remember two tilted bars on each trial. Instead of uncoupling the memorised orientations from the relevant motor plan, each memory item was associated with a specific action in this paradigm. This was achieved by placing restrictions on the responses at probe: the left hand should always be used to dial up the orientation of a leftward titled item (relative to the vertical axis) and the right hand should always be used to respond to a rightward titled item. Selection of the relevant content in memory (left or right item) was thus dissociated from the selection of the relevant action (left vs right hand). Decoding from EEG activity showed that selection of the task-relevant sensory representation and the appropriate motor plan occurred in parallel and not serially (visual, then motor) as would be expected if the WM representation was predominantly sensory in nature. Further, this tight link between sensory representations and action planning was shown to occur early during the WM delay (Boettcher, Gresch, Nobre, & van Ede, 2020).

Establishing strong links between representations in memory and motor systems might be adaptive for a memory item that has been selected as relevant for the current task, but perhaps not for items that might become relevant later. One theory suggests attentional enhancement of representations in WM may be a by-product of establishing links between currently relevant memory content and the appropriate motor plan (Olivers & Roelfsema, 2020). When an item in WM is selected for action, recurrent feedback connections between action-related systems and the relevant sensory representation may produce selective neural enhancement of the selected representation in sensory cortex. The model further predicts that attentional selection of an item in WM is associated with anticipated reward. This may itself engage reward-based plasticity mechanisms to strengthen synaptic connections in the neural population coding for the

selected item. According to the model, attentional selection supports memory-guided behaviour through a combination of consolidated memory traces, as predicted by previous literature (Manohar et al., 2019; Rerko & Oberauer, 2013; Souza et al., 2015), and strengthened sensory-motor links (A. Heuer, Ohl, & Rolfs, 2020; Myers et al., 2017; Olivers & Roelfsema, 2020).

Preliminary support that attentional selection may strengthen sensory-motor connections comes from studies showing attentional focusing within WM engages the oculomotor system (Van Ede, Board, & Nobre, 2020; van Ede, Chekroud, & Nobre, 2019). Retro-cued selection of non-spatial contents from WM was associated with small fixational shifts in gaze toward the original location of the cued object. Microsaccades toward the original position of a selected item in WM may reflect a type of motor plan to saccade to the remembered location. Even if spatial location was not directly relevant in this study, it may generally be useful for relevant information in WM to guide where to look next in order to achieve a behavioural goal.

Further support comes from a recent study showing that the content of WM drives motor behaviour on a secondary task (J. A. Miller, Kiyonaga, Ivry, & D'Esposito, 2020). Participants remembered verbal directions (up, down, left, right) for a delayed match/mismatch task. During the delay, participants performed a dual motor task in which they were cued to move their mouse and click on one of four boxes. The position of the cued box was either compatible with the direction in memory (e.g. left and left) or incompatible (e.g. left and up). The speed and trajectory of mouse movements on the motor task were influenced by the directions maintained in WM, even when WM contents were anti-predictive of the motor location. Additionally, when two items were maintained in WM, the prioritised memory item, indicated by a retro-cue, had greater influence on movements than the unprioritised item. The results are consistent with a model of WM where prioritised WM content is represented in an action-oriented state

that drives ongoing behaviour (A. Heuer et al., 2020; Myers et al., 2017; Olivers & Roelfsema, 2020).

Myers et al. (2017) proposed a two-step theory of retro-cued prioritisation in WM. First, a frontoparietal attentional network is engaged to select the relevant representation in memory (A. Gazzaley & Nobre, 2012)(see Section 4.4.2). Second, a prefrontal network comprising the ventromedial PFC and potentially the anterior cingulate cortex and pre-supplementary motor area supports a re-configuration of stimulus-response mappings thereby transforming the selected representation into an action-oriented task-set optimised for performing the current task. Once a WM representation has been re-formatted into an action-oriented state, the need for sustained attention may no longer be required, which could help account for the puzzling finding that retro-cues may improve behavioural performance even when attention is subsequently directed away from the cued item (Rerko & Oberauer, 2013; Souza & Oberauer, 2016). It also predicts behavioural benefits for the retro-cued item without associated costs for uncued items in memory (Li & Saiki, 2014; Myers et al., 2018).

Support for this theory of prioritisation comes from neuroimaging studies reporting that areas traditionally involved in motor planning are also recruited during retro-cueing in WM (Chatham et al., 2014; Nee et al., 2013; Tamber-Rosenau et al., 2011; Wallis et al., 2015). While there are substantial overlaps in the neural areas engaged during external attention and attention to contents in WM, one study found activation in basal ganglia that was specific to internal attentional shifts (Tamber-Rosenau et al., 2011). Additionally, areas within a cingulo-opercular network, including the frontal operculum/anterior insula and presupplementary motor cortex were exclusively recruited for retro-cueing, not pre-cueing, in WM (Wallis et al., 2015). Engagement of these areas is also seen during task switching and response selection (Higo, Mars, Boorman, Buch, & Rushworth, 2011; Richter & Yeung, 2015; Rushworth, Hadland,

Paus, & Sipila, 2002) and might reflect an additional process of 'output gating' during retro-cueing that is not involved in pre-cueing, namely selecting which content is allowed to drive behaviour (Allport, 1987; Chatham et al., 2014).

Output gating in WM has been linked to increased recruitment of lateral frontal cortex as well as enhanced connectivity between the striatum and lateral frontal cortex compared to "input gating" (i.e. gating which information is allowed to enter WM)(Chatham et al., 2014). Participants performed a hierarchical WM task where a context cue informed them which of two items should guide behaviour during fMRI scanning. The context cue was either presented before stimulus presentation (analogous to a pre-cue) or was presented after the WM items (analogous to a retro-cue). In the context-first condition, participants knew which item was relevant and so could use a strategy to selectively encode a single item into WM ("input gating"). In the context-last condition, participants had to encode both items into memory and only after knowing the context could they select which item should be allowed to influence behaviour ("output gating"). Input and output gating recruited overlapping cortico-striatal areas, but output gating was associated with greater BOLD activity in the lPFC, premotor cortex and striatum and greater connectivity between the lateral frontal cortex and striatum. The authors proposed that the lateral frontal cortex is involved in maintaining task rules that influence the striatum to gate which lower order information in WM is allowed to guide behaviour.

A recent EEG study investigated the contribution of motor planning to the retro-cueing benefit (Schneider, Barth, & Wascher, 2017). Participants were presented with three WM items and a retro-cue validly cued one, two, or all three of the items, or was a neutral cue. Alpha power suppression over occipitoparietal areas was present for all conditions but scaled with cue-type, consistent with a role in freeing up attentional resources for processing. Prior to probe onset, the single item retro-cue was uniquely

associated with mu/beta power suppression contralateral to the response side over parietal electrodes, with a likely source in sensorimotor cortex. This suppression was similarly present for the other conditions *after* probe-onset, but was specific to the single-cue condition *before* probe onset, indicating a potential role in motor preparation. Thus, the results support a two-stage model of retro-cueing with an initial role of attentional selection reflected in posterior alpha power, and a second stage of response planning that is limited to a single item reflected in parietal mu/beta power.

While the literature suggests that selection in WM engages brain regions involved in motor planning and that this may play a role in prioritising selected content for action, it is not yet clear what would constitute an "action-oriented state" in WM (Myers et al., 2017; Olivers & Roelfsema, 2020). One option is that selection in WM strengthens sensory-motor links without changing the underlying coding scheme of the memory representation (Olivers & Roelfsema, 2020). Another option is that attentional selection re-configures the representation in WM, perhaps by establishing a functional template for decision-making in the context of the current task (Myers et al., 2017; Navalpakkam & Itti, 2007; Stokes, 2015). If so, the memory representation may be used as a decision-boundary to prepare a prospective task set that links potential sensory inputs to the appropriate response (Myers et al., 2017).

5.3. Prospective task sets in WM

Oberauer (2010) proposed a distinction between declarative WM and procedural WM with declarative WM holding representations about the world and procedural WM maintaining task sets for guiding action (Oberauer, 2010). A task set refers to the prospective configuration of cognitive processes, including perception, attention, memory and motor planning, in order to form stimulus-response (S-R) mappings in preparation for a certain task or behaviour (Monsell, 2003; Sakai, 2008).

Task-sets are commonly invoked in the literature on cognitive control as a way of routing stimulus and response selection along the appropriate pathway for guiding behaviour. This may be especially important in situations involving task-switching or multi-tasking where the same stimuli are associated with different responses depending on the context. For example, an amber traffic light may signal you to stop if you're arriving at the intersection, but signal you to press the accelerator if you are already stopped and waiting for the green light.

In classic task-switching studies, participants are instructed to switch between tasks involving the same stimulus set and response modes, for example to classify numbers based on their magnitude or parity (Monsell, Sumner, & Waters, 2003). Task switches relative to task repetitions are associated with performance costs, but these may be reduced by cueing the relevant task ahead of time, perhaps by allowing top-down re-configuration of the neural state for the task to occur before target onset (Meiran, 1996; Rogers & Monsell, 1995), but c.f. (Allport, Styles, & Hsieh, 1994). Configuration of a task set can be thought of as retrieval of the appropriate context-dependent S-R mappings (e.g. 'if amber traffic light, press accelerator'). Thereby a task cue may allow for a faster and more accurate translation from stimulus to motor output once the target appears, whereas the full reconfiguration would have to occur after target onset in the absence of a task cue. An alternative account of task switch costs suggest these may be the result of task-set inertia (Allport et al., 1994). When both tasks involve overlapping stimuli, the target stimuli may prime the recently completed task-set resulting in proactive interference on task-switch trials compared to task-repeat trials. A combination of preparatory task-set configuration and task-set interference may contribute to the task-switch costs (Kiesel et al., 2010), but importantly both theories invoke task-sets to explain performance costs.

Another index of task-set configuration is task-rule congruency effects (Braem, Deltomme, & Liefooghe, 2019; Liefooghe, De Houwer, & Wenke, 2013; Liefooghe, Wenke, & De Houwer, 2012; Monsell, 2003). Task-rule congruency effects are found in dual-task paradigms when participants perform two tasks with overlapping stimuli and response modes. For example, participants encoded a pair of stimulus-response mappings ('if N, press left, if Q press right') to be executed after a delay (Liefooghe et al., 2012). During the delay, a nested task required participants to categorise letters based on their tilt (left vs right), while ignoring their identity. Performance on the nested task was faster and more accurate when the identity and the tilt of the letter pointed to the same response (congruent) relative to when they pointed to conflicting responses (incongruent). In this way, instructed stimulus-response mappings encoded into WM to be executed after a delay drove responses on a nested delay task even when irrelevant, indicating that novel task-sets maintained in WM may interact with incoming sensory input to drive behaviour in a reflex-like manner (Braem et al., 2019; Liefooghe et al., 2013, 2012). Importantly, task-rule congruency effects are only seen when task-sets are actively prepared for implementation and not when they are merely memorised for later recognition (Liefooghe et al., 2013).

Recent literature suggests WM may similarly be supported by a translation of the relevant sensory representation into a prospective task set in preparation for the probe (González-García, Formica, Liefooghe, & Brass, 2020; Muhle-Karbe, Duncan, De Baene, Mitchell, & Brass, 2017; Muhle-Karbe, Myers, & Stokes, 2020; Myers et al., 2017). For example, in a classic delayed change-detection task, a red memory stimulus may be reconfigured into the prospective S-R mapping 'press YES if target is red, press NO for any other colour'. Encoding WM content in such a response-oriented format might facilitate context-dependent processing of the probe stimulus and enable faster and more accurate decision-making.

A frontoparietal network may be recruited to reformat novel S-R mappings into decision-circuits in preparation for speeded decision-making (Muhle-Karbe et al., 2017). In one study, participants were presented with two images on every trial, each associated with a different response finger during fMRI scanning. On half of the trial blocks, participants remembered the instructed stimulus-response mapping for subsequent execution following a target image (implementation block) and on the other half, participants remembered the stimulus-response mappings for subsequent recognition (memorization block). Decoding of the instructed S-R mappings was enhanced during the delay for the implementation blocks compared to the memorization blocks in a network of frontoparietal regions, comprising the inferior frontal junction, inferior frontal sulcus, intraparietal sulcus, as well as object-selective areas of visual cortex. Interestingly, decoding of instructions from visual cortex during the delay correlated with response times only in implementation task blocks. Together, the results suggest a network of frontoparietal areas may be involved in reformatting the novel, instructed S-R mappings into decision-circuits in preparation for a speeded response specifically when the instructions need to be executed.

It is possible that this format prepared for execution in the implementation condition resembles a prioritised state in WM when an item has been selected as relevant for guiding behaviour (Myers et al., 2017). This is supported by two recent studies (González-García et al., 2020; Muhle-Karbe et al., 2020). In a behavioural study, participants encoded two pairs of S-R mappings on each trial (four unique stimuli each associated with one of two response keys)(González-García et al., 2020). After an initial delay, participants were retro-cued to one pair of S-R mappings, followed by a nested delay task in which participants were presented with one of the memorized stimuli and had to respond with its colour. The response key on the nested task could overlap with the S-R mappings in memory (congruent) or not (incongruent). If retro-cued

prioritisation re-configures the memorized S-R mapping into an action-oriented state, then task-rule congruency effects should be present for the cued S-R mappings, but not for the uncued S-R mappings, which is what was found.

In an EEG study, participants encoded two orientations into WM that served as decision boundaries for a series of targets for which participants had to make clockwise/counter-clockwise decisions (Muhle-Karbe et al., 2020). Before each target, participants were cued to which of the two memorized orientations should be used for decision-making. Decoding from EEG data revealed that the relevant, but not irrelevant, orientation could be decoded after the cue, consistent with previous studies showing preferential coding for relevant WM content (LaRocque et al., 2013; Wolff et al., 2017). Following target presentation, both relevant and irrelevant orientations could be decoded, suggesting the target may serve as an impulse to drive latent states coding for the irrelevant item in memory (Rose et al., 2016; Wolff et al., 2017). However, the neural pattern for active and latent memory items were qualitatively different and had distinct roles on behaviour: trial-wise decoding strength of the active memory item was a strong predictor of performance for the current target, while decoding of the latent item was unrelated to performance for the current target, but predicted performance over a longer time scale for the series of targets. Moreover, the latent item interfered with current trial performance to the extent that its neural format resembled the pattern of the active item. Finally, drift diffusion modelling suggested that stronger decoding of the active memory item was associated with faster decision-time (i.e. drift rate). This is consistent with a hypothesis of prioritisation in which the active memory item is transformed into a decision-circuit by acting as a matched filter, directly interacting with new sensory input by filtering these through the neural pattern coding for the mnemonic template, to compute the relevant decision (Hayden & Gallant, 2013; Myers, Rohenkohl, et al., 2015; Sugase-Miyamoto, Liu, Wiener, Optican, & Richmond, 2008).

5.4. Interim summary

In summary, *what* is represented in WM as well as *how* it is represented in WM may depend on specific task demands. Additionally, evidence is starting to emerge that attentional selection may play an important role in establishing sensory-motor links for guiding behaviour. It is possible that selected representations in WM become linked to their context-dependent action in preparation for behaviour, such that a probe triggers the appropriate response in a reflex-like manner. Some have suggested a 1:1 relationship between the cognitive state in WM (currently relevant vs prospectively relevant) and the neural activity state (active vs silent)(Larocque et al., 2014; Manohar et al., 2019). However, activity-silent brain states may actually be in a good position to interact with incoming sensory information as the pattern of activation should reflect synaptic traces coding for the items in memory (Mongillo et al., 2008; Stokes, 2015; Stokes et al., 2020). If attentional selection, in part, serves to consolidate synaptic memory traces in the sensory cortex (Manohar et al., 2019; Olivers & Roelfsema, 2020; Rerko & Oberauer, 2013), these could serve as functional templates for comparison between a memory item and new sensory input (Hayden & Gallant, 2013; Rademaker et al., 2019; Scolari et al., 2012). However, empirical evidence for prospective states in WM is still lacking and further research is needed to elucidate the ways in which anticipated task-demands modulate states in WM to guide behaviour.

6. Thesis scope

The aim of the thesis was to investigate how WM content is selected and prioritised for action. To shed light on this general question, I developed two primary research questions that informed the experiments presented in this thesis: (1) How is information selected in working memory? And (2) How is WM content prioritised for future action?

6.1. How is information selected in working memory?

A wealth of evidence supports the existence of distinct representational states in WM: selected items in WM are recalled faster and with greater precision and can be decoded to a higher degree from patterns of neural activity, particularly in the sensory cortex. A frontoparietal network of attentional control may provide top-down signals to bias processing in sensory regions toward selected WM content. But what is selected in WM: features or objects?

Previous work has shown that WM preferentially processes task-relevant features, while evidence of object-based encoding in WM suggests there might be boundaries to WM selectivity for individual features. A recent computational model of attention in WM predicts that attentional selection of a single feature in WM reactivates associated item-features through obligatory pattern completion, thereby bringing the entire item into an attended state (Manohar et al., 2019)(see Section 4.5.1).

In a first set of experiments (Chapter 2-3), I investigated the consequences of selecting a single feature in WM for the representation of other features belonging to the same object. In **Chapter 2**, I present behavioural evidence that attention spreads between object-features in WM. EEG and eye-tracking data further suggest that spatial attention plays an important role in selecting non-spatial features in WM. The results are consistent with a privileged role of location in WM. In **Chapter 3**, I replicate these results and show spatial attention is involved in selecting non-spatial features from WM, even when location is redundant. Selecting a colour from WM further enhances the neural representation of the item's orientation, consistent with auto-associative spread of attention between an item's features in WM, and the selected item remains in a privileged state for comparison with the probe at recall.

6.2. How is working memory content prioritised for future action?

Recent theoretical and experimental work has emphasised the core purpose of WM in guiding future behaviour (Nobre & Stokes, 2019). Specifically, because attention selects content that is relevant for a forthcoming task, selected WM content may be held in a state that is optimised for guiding action. However, it is not yet clear what might constitute an action-oriented state in WM. One theory suggests attentional selection supports a re-configuration of a retrospective sensory representation into an action-oriented code that supports decision-making (Myers et al., 2017).

The results of two behavioural experiments (**Chapter 4**) support a model where prioritised content in WM is configured as a prospective task-set in a way that facilitates speeded WM-guided decision-making. If so, the neural code for prioritised information in WM should depend on its anticipated use. In **Chapter 5**, I present a planned EEG experiment, in the format of a Stage 1 Registered Report, designed to test if the neural format of relevant content in WM is task-specific. I include simulated neural data validating the analysis approach.

Chapter 2. Attention spreads between object-features in working memory

Working memory (WM) allows humans to preferentially process information depending on its relevance to current task demands (Bundesen, 1990). However, whether encoding into WM is object-based or whether WM selectively encodes task-relevant features has been the subject of ongoing debate (Bocincova & Johnson, 2019; Gao et al., 2016; Olivers et al., 2006; Serences et al., 2009; Shen et al., 2013; Woodman & Vogel, 2008; Xu, 2010). Moreover, once information is encoded into WM it is unclear whether internal attention selects individual features or spreads between an object's features in WM. In two experiments, I used behavioural methods, eye-tracking and electroencephalography (EEG) to show that selecting one feature of a multi-feature object in WM engages spatial attention and privileges associated item-features. The results are consistent with recent computational models proposing auto-associative retrieval of an item's features (Manohar et al., 2019; Oberauer & Lin, 2017) and further suggest a special role of location in selecting and binding features in WM (Schneegans & Bays, 2017a).

Selective attention is thought to play a key role in the top-down control of WM–including the selection of relevant information for storage in the service of a behavioural goal (Baddeley, 2012; Cowan, 2005; A. Gazzaley & Nobre, 2012; Oberauer, 2002). Specifically, attention is important for filtering information during WM encoding such that task-relevant objects are selectively encoded into WM while filtering out distractors (Bundesen, 1990; Vogel & Machizawa, 2004; Vogel et al., 2005), thereby optimizing the use of limited WM capacity (Luck & Vogel, 1997). Further, attention may enable selective encoding of task-relevant features into WM such that, for example, only the colour and not the shape of an object is stored in memory (Bocincova & Johnson, 2019; Serences et al., 2009; Woodman & Vogel, 2008). However, selective encoding of features

is disputed by evidence from irrelevant-change distraction paradigms where WM performance is impaired by a change in an irrelevant feature-dimension at probe (Gao et al., 2016; Jiang et al., 2000; Shen et al., 2013; Soto & Humphreys, 2009; Treisman & Zhang, 2006). These suggest encoding into WM is object-based, imposing boundaries on WM selectivity. It is likely that WM maintains a combination of feature- and object-level representations, with emphasis on either or both depending on factors such as instructions, task-relevance, and probe display (Brady et al., 2011; Fougnie et al., 2010; Vergauwe & Cowan, 2015).

During the maintenance period, retrospective cues ('retro-cues') that direct internal attention to task-relevant WM objects improve subsequent behavioural performance for cued content (Griffin & Nobre, 2003; A. Heuer & Schubö, 2016; Landman et al., 2003; Pertzov, Bays, et al., 2013; Souza & Oberauer, 2016). Additionally, retro-cues that inform participants which *feature* of multi-feature objects will be relevant for decision-making provide behavioural benefits similar to object-based retro-cues in WM, suggesting attention can also be allocated to prioritise features (Hajonides, van Ede, Stokes, & Nobre, 2019; A. Heuer & Schubö, 2016; Niklaus, Nobre, & Van Ede, 2017).

A wealth of evidence has now shown that attention can be flexibly shifted between items in WM depending on their momentary relevance (LaRocque et al., 2013; Lewis-Peacock et al., 2012; van Ede, Niklaus, & Nobre, 2017; Zokaei, Ning, et al., 2014), sometimes without costs to other items (Hollingworth & Maxcey-Richard, 2013; Myers et al., 2018; Rerko et al., 2014). Even when all encoded information is equally relevant to a future decision, an 'incidental' task directing attention to one memory item for an orthogonal dual-task can push this item into a privileged state and improve behavioural performance (Zokaei, Manohar, et al., 2014; Zokaei, Ning, et al., 2014). Similarly, refreshing cues to "think of" items in memory may improve memory performance in proportion to the number of times each item was refreshed (M. R. Johnson, Mitchell,

Raye, D'Esposito, & Johnson, 2007; Souza et al., 2015; Vergauwe & Langerock, 2017; Yi et al., 2008). The process of attending to an item during the memory delay facilitates memory performance, even if attention is subsequently removed or directed elsewhere (Rerko & Oberauer, 2013; Rerko et al., 2014). This led to the proposal that attentional selection during WM enhances the memory traces for the selected item, perhaps through strengthened connections between the object-features and their context (e.g. location) (Oberauer & Lin, 2017; Rerko & Oberauer, 2013).

Traditionally, computational models of WM have not incorporated markers of selective attention within WM. For example, networks with recurrent connectivity that sustain sensory activity within a stable attractor state (Compte et al., 2000; X. J. Wang, 2001) are successful at modelling short-term maintenance of individual features, but do not include mechanisms for internal attentional orienting. Recent work has built on traditional attractor models to incorporate different attentional states (Manohar et al., 2019; Oberauer & Lin, 2017; Schneegans & Bays, 2017b). For example, a recent model has incorporated a focus of attention in WM by proposing an equivalence between the neural activity state (active vs silent mechanisms) and the cognitive state (attended vs unattended)(Manohar et al., 2019). Interestingly, this model predicts that attended WM content is represented in spiking neural activity in populations or neurons representing each object-feature, while items outside the focus of attention are maintained through associative short-term synaptic traces between feature units and flexible conjunctive units, established at encoding (Manohar et al., 2019). Oberauer and Lin (2017) similarly proposed that encoding into WM establishes associative connections between features and their encoding context (Oberauer & Lin, 2017). In this model, the item in the focus of attention is represented in an active neural state, while recently selected objects may be maintained in a *region of direct access* (Oberauer, 2002) through strengthened "bindings" between object-features. In both models, a retrieval cue acts by activating the

relevant feature unit, which in turn activates the associated object-features through pattern completion mechanisms bringing the entire object into the focus of attention.

If features in WM are bound through auto-associative connections and retrieved through pattern completion, as predicted by these models, a retrieval cue that directs attention to one item-feature in memory should strengthen all other features encoded as belonging to that object, even when these are not currently relevant. While previous research has suggested that task-relevant features can be selectively encoded into WM (Serences et al., 2009; Woodman & Vogel, 2008), it remains unclear whether it is similarly possible to select only task-relevant object-features during *retrieval*. In one behavioural experiment and one EEG and eye-tracking experiment, I test the primary hypothesis that attention automatically spreads between object-features in WM. Alternatively, it may be possible to select an individual feature in WM without strengthening other features belonging to the same object.

The primary hypothesis assumes that attention spreads symmetrically between features. However, WM may not represent all object-features equally. For example, a wealth of evidence suggests space may be special, perhaps because of its role in defining objects (Kahneman, Treisman, & Gibbs, 1992; Pertzov & Husain, 2014; Schneegans & Bays, 2017a; Treisman & Zhang, 2006; B. Wang et al., 2016; Wood, 2011). In general, features presented in the same location (at any one time) form part of the same object. Consequently, location may play a special role in binding non-spatial features in WM, as proposed by a recent computational model (Schneegans & Bays, 2017a). Therefore, a second question concerns whether selection of spatial and non-spatial features induce similar attentional spread to associated object-features in WM.

Experiment 1

To test the prediction that attention spreads between object-features in WM, I conducted a multi-item, multi-feature visual WM experiment. Participants encoded three oriented, coloured bars with different spatial locations on each trial. Critically, during the maintenance delay participants were required to retrieve only one feature of a single item stored in memory (colour or location). After the delay, participants were probed to recall the orientation of one of the items. The orientation probed for recall at the end of the trial either belonged to the same object as the feature cued during the maintenance period (congruent trials, 1/3 of trials) or belonged to a different object (incongruent, 2/3 of trials). Consequently, the cue during the maintenance period was not predictive of which of the three items would be probed for orientation recall ('incidental' cue; Zokaei, Ning, et al., 2014).

If selectively attending the cued feature automatically privileges associated but currently redundant item-features, orientation judgements should be faster and more precise on congruent than incongruent trials. Alternatively, if it is possible to selectively access the object's colour or location without strengthening associated object-features, there should be no congruency benefit. To test whether selection of spatial and non-spatial features similarly induce attentional spread to associated object-features, I independently manipulated the feature dimension of the cue and the probe (colour vs location). This manipulation also served to test whether a potential congruency effect was supported by increased accessibility of the cue-feature in memory, in which case it should be specific to conditions where the cue and probe shared the same feature dimension ('colour cue + colour probe '/'location cue + location probe'). Alternatively, if the full object was brought into a privileged state as predicted, the congruency effect should extend to conditions were the cue and probe did not share the same feature dimension ('colour cue + location probe'/'location cue + colour probe').

By modelling the sources of error using mixture modelling, I also tested the extent of feature-binding of attended vs unattended items (Bays et al., 2009; Zhang & Luck, 2008). If object features are stored as conjunctions in WM regardless of attentional state (Luck & Vogel, 1997), the probability of swap errors should not change as a function of congruency. In contrast, if object-features are bound together associatively, and their binding is strengthened by attention (Manohar et al., 2019; Oberauer & Lin, 2017; Rerko & Oberauer, 2013), the probability of swap errors should be lower on congruent relative to incongruent trials.

To preview the results from Experiment 1, orientation responses were faster and more precise for congruent than incongruent trials. This was the case for both colour and location incidental cues and retrieval cues, though location may play a special role in selecting non-spatial features from WM.

Methods

Participants

28 participants took part in the experiment. A threshold of maximum 40 % errors of omission was set for performance on the incidental task to ensure participants were paying attention to the incidental cue. Five participants were excluded due to low performance on the incidental task (>40 % errors of omission). Another two participants were excluded due to having misunderstood task instructions. The remaining 21 participants (16 female) aged 18-55 (M = 26.38, SD = 8.73) were included in the analyses. One participant was left handed, the rest were right handed. The study was approved by the Oxford Central University Research Ethics Committee (R55073/RE001). All participants gave informed consent.

Procedure

The task consisted of a continuous report WM task and a dual match/non-match incidental task (Figure 1). Following an initial fixation period of 500 ms (fixation point size=0.3° visual angle), participants were presented with three oriented, coloured bars on each trial (height=3°, width=0.5°). The colours and locations of the stimuli were picked at random without replacement out of six possible on each trial. The six possible colours were red, yellow, green, cyan, blue and magenta (RGB values: [210 0 10], [230 210 50], [0 150 0], [0 210 210], [0 50 255], [200 50 170]). The six possible locations were equally spaced on a circle, 60 deg apart, with radius of 7° visual angle from the fixation point. The location set was counterbalanced across participants (from zero: 15°, 75°, 135°, 195°, 255°, 300° or -15°, 45°, 105°, 165°, 225°, 285°). The orientation angles of the bars were chosen at random with replacement between 1-180 degrees on each trial. The sample display was on the screen for 250 ms followed by a delay period of 1500 ms. For the incidental task, participants were presented with a ring-shaped colour- or location cue (blocked condition; diameter=0.7°, width=0.2°) for 200 ms. Participants were asked to make a match/non-match judgement whether the cue matched the feature of one of the items in memory. On non-match trials (20%), participants were asked to press the space bar within 1000 ms of incidental cue onset, after which the rest of the trial was skipped. Participants received audio-feedback on error trials (high-pitched tone). On match trials (80%), participants made no response and the trial continued. After another delay of 1000 ms, the probe display appeared on the screen. It consisted of a dial for orientation recall (3°), the orientation of which was chosen at random between 1-180 degrees on each trial. The object for orientation recall was cued either by colour or location (trial wise condition) and was either the same object as was cued by the incidental task (congruent condition, 1/3 of trials) or a different object (incongruent condition, 2/3 of trials). As a result, the incidental cue was non-predictive of the item cued for orientation recall.

Participants used the arrow keys to turn the on-screen dial to match the orientation of the cued object and pressed the space bar to respond. Participants had as much time as they needed to start responding, but once they initiated a response there was a time-out after 2500 ms. They received visual feedback of the correct orientation for 200 ms.

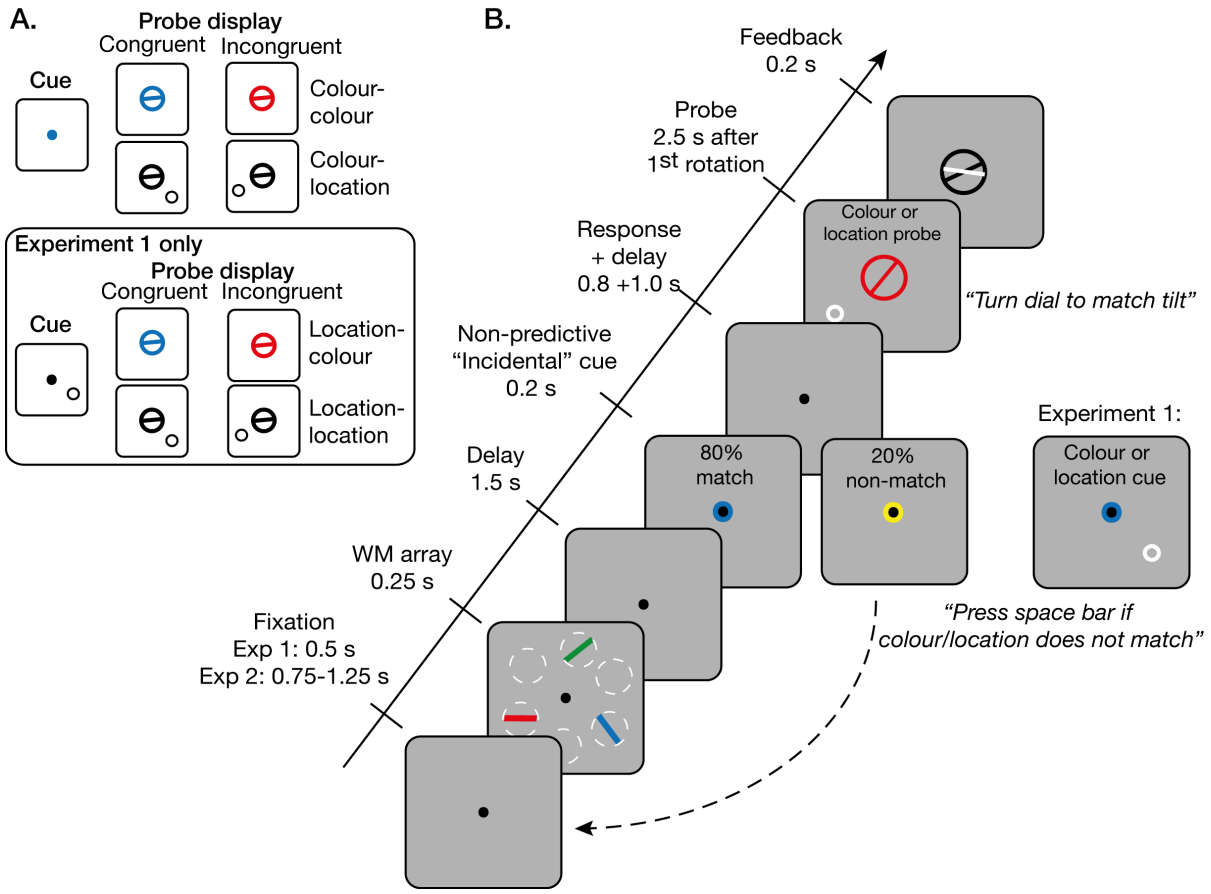

Figure 1. Task design for Experiment 1 and 2. **A.** Task conditions. The feature dimension of the incidental cue was either colour or location in Experiment 1 and always colour in Experiment 2. The feature dimension of the probe was either colour or location in both experiments. If the cued feature belonged to the same object as the one probed for orientation recall, the trial was congruent (1/3). If a different object was probed, the trial was incongruent (2/3). **B.** Trial sequence. Participants were presented with three oriented, coloured bars on each trial. After an initial delay, an incidental cue (Exp 1: location or colour, Exp 2: colour only) was presented and participants pressed the space bar if the cued feature did not match one of the items in WM (20% of trials). After a second delay, a probe was presented and participants turned the bar to match the orientation of the probed object (probed by colour or location). Participants received feedback after responding. Dashed white circles represent the possible locations of the stimuli on the screen (for illustration purposes only).

Participants completed 600 trials in total, split into 20 blocks of 30 trials, lasting about 1-1.5 hours including breaks. The 20% non-match incidental cue trials were skipped leaving 480 orientation recall trials for the main analyses.

The task was presented on a 19 inch, 1280×1024 pixel Dell monitor. Participants were seated 57 cm from the screen. The task was programmed and stimuli presented in Matlab version R2015b (The MathWorks Inc., 2015) and Psychtoolbox (Brainard, 1997a).

Analyses

Statistical analyses were performed in Matlab R2015b (The MathWorks Inc., 2015) and SPSS v24 (IBM Corp., 2016). A 2×2×2 repeated measures ANOVA was performed with factors congruency (congruent vs incongruent), cue feature dimension (colour vs location) and probe feature dimension (colour vs location) on mean absolute error (MAE; in radians) and median reaction times (RT; time from probe onset until first key press).

When the cue and probe shared a feature dimension, it is possible the congruency benefit was supported by increased accessibility of the cued feature in memory. To test whether potential congruency benefits were supported by overlapping mechanisms when the incidental cue and the probe shared the same cue characteristics (i.e. 'colour cue-colour probe' and 'location cue-location probe') vs when they had distinct cue characteristics (i.e. 'colour cue-location probe' and 'location cue-colour probe'), I computed Pearson correlations between the size of the congruency benefit (congruent–incongruent) for the same vs distinct cue characteristics conditions, separately for MAE and RT. To ensure potential correlations were not determined by overall performance, I computed the abovementioned correlations on the residuals from linear regression between the size of the congruency benefit for each condition and overall performance (i.e. overall MAE and overall RT).

To delineate between potential sources of memory errors and how they were affected by incidental-cueing, I fit a mixture model to trial-wise raw error using Bays Lab toolbox (Bays et al., 2009). The model is an extension to the Zhang and Luck (2008) model and can be described in the following way:

$$P(\theta') = (1 - \gamma - \beta)\phi_\sigma(\theta' - \theta) + \gamma\frac{1}{2\pi} + \beta\frac{1}{m}\sum_{i}^{m}\phi\,(\theta' - \theta_i^*)$$

Where θ denotes the target orientation in radians. θ' denotes the reported orientation. γ is the proportion of trials where the subject responds at random. ϕ_σ denotes the circular equivalent of the normal distribution (Von Mises distribution) with mean 0 and precision σ. The swap model extension also includes the β parameter, which is the probability of responding with one of the two non-target values: $\{\theta_1^*, \theta_2^*, \theta_m^*\}$, for m non-targets.

There were insufficient trial numbers to fit the model separately for colour and location cues or probes, so conditions were combined where the incidental cue and the probe shared the same cue characteristics (i.e. 'colour-colour' or 'location-location') and conditions where the cue and probe had different cue characteristics (i.e. 'colour-location' or location-colour'). The parameters obtained from the model (Von Mises distribution precision parameter, proportion target responses, proportion non-target responses, proportion guesses) were analysed in a 2×2 repeated measures ANOVA with factors congruency (congruent vs incongruent) and cue-probe feature dimension (same vs different).

For the incidental task, I compared errors of omission, errors of commission and reaction times for colour and location incidental cues in paired samples t-tests.

I complement the frequentist statistics with Bayes Factor Analysis, performed using JASP v0.9.0.1 (JASPTeam, 2018) with default priors. These analyses indicate the

likelihood of the data given H_1 relative to the null hypothesis H_0 (reported here as BF_{10}; values > 1 suggest evidence in favour of the H_1, while values <1 suggest evidence in favour of the null). For the Bayesian repeated measures ANOVA analyses interaction terms I report the inclusion Bayes Factor across matched models (BF_{incl}). This compares the models that contain that effect to the equivalent models stripped of this effect by dividing the sum of the P(M | data) of the former by the P(M | data) of the latter. However, the jury is still out on the best way of reporting Bayes Factors for interaction terms in repeated measures ANOVAs involving multiple factors. As of December 2020, the authors of the JASP package (JASPTeam, 2018) are writing an article concerning discrepancies between frequentist and Bayesian ANOVAs for higher-order interactions[1].

Results

Only trials with correct performance on the incidental task (i.e. non-responses on match trials) were included in the main analyses of orientation recall performance. Table 1 shows descriptive statistics for mean absolute error (MAE) and reaction times (RT).

Table 1. Descriptive statistics for mean absolute error (rad) and reaction time (s). N = 21.

	Mean absolute error				Reaction time			
	Congruent		Incongruent		Congruent		Incongruent	
Cue feature – Probe feature	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Colour – Colour	.707	.212	.925	.231	.736	.226	.908	.241
Colour – Location	.745	.257	.969	.209	.832	.192	.992	.287
Location – Colour	.817	.190	.931	.248	.777	.242	.896	.300
Location – Location	.704	.176	1.031	.234	.794	190	.964	.322

[1] https://forum.cogsci.nl/discussion/6501/when-the-bf-and-the-classical-results-diverge Accessed 11/01/2021

Mean absolute error

Figure 2A shows the distribution of response errors for all participants, separately for congruent and incongruent trials. I computed MAE for each participant and analysed these in a 2×2×2 repeated measures ANOVA (Figure 2B & C). Consistent with the primary hypothesis, a main effect of congruency showed that error for congruent trials (M=.745, SD=.184) was significantly lower than for incongruent trials (M=.965, SD=.197): $F_{1,20}$=60.78, p<.001, η^2=.752, BF_{10}=7.25×10^{14}.

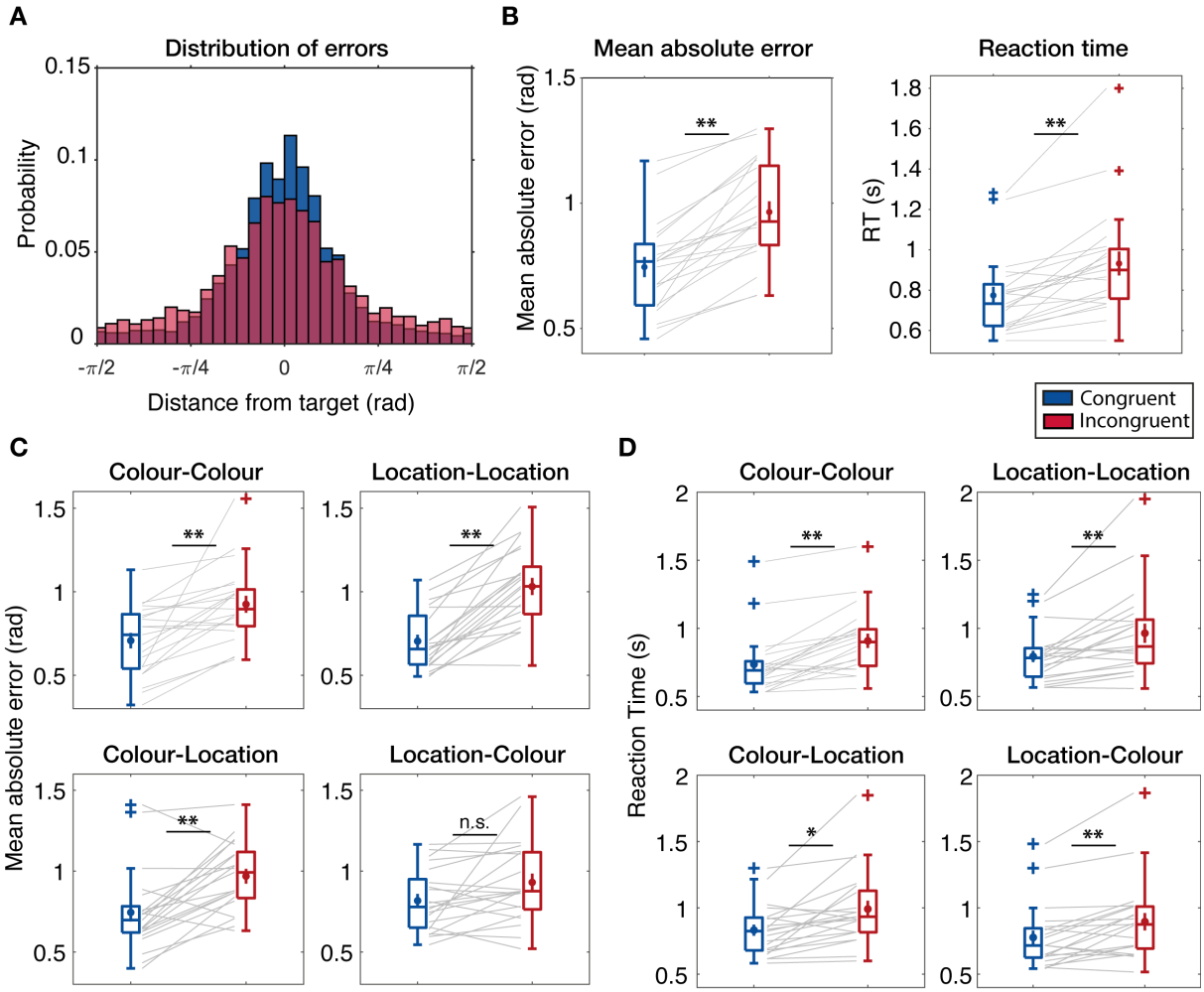

Figure 2. Mean absolute error and reaction times. **A.** Distribution of errors for all participants relative to target angle for congruent (blue) and incongruent (red) trials. **B.** Left: Boxplot of mean absolute error (rad) for congruent and incongruent trials. Mean and standard error across participants overlaid. Grey lines show individual participants. Right: same as left panel but for reaction times (s). **C.** Mean absolute error split by cue and probe feature dimensions. Top left: colour cue, colour probe. Top right: location cue, location probe. Bottom left: colour cue, location probe. Bottom right: location cue, colour probe. **D**. Same as C, but for reaction times (s). *p<.05, **p<.001.

Additionally, there was a significant three-way interaction between congruency, incidental cue feature and probe feature: $F_{1,20}$=9.66, p=.006, η^2=.326, BF_{incl}=5.59. This interaction was followed up with 16 pairwise comparisons (Bonferroni corrected alpha level=0.0031). There were significant congruency effects when the incidental cue and the probe shared cue characteristics, whether this was colour ('colour-colour': t_{20}=-4.38, p<.001, d=.955, BF_{10} =107) or location ('location-location': t_{20}=-7.60, p<.001, d=1.66, BF_{10}=63087). However, there was an asymmetry when the cue and probe had distinct features. When the incidental cue feature was colour and the probe feature was location, there was a significant congruency effect ('colour-location': t_{20}=-4.60, p<.001, d=1.00, BF_{10}=169), but this was not the case when the incidental cue feature was location and the probe feature was colour and Bayes Factor suggested only limited evidence for this effect ('location-colour'; t_{20}=-2.49, p=.022, d=.544, BF_{10}=2.68).

The lack of a significant congruency effect for the 'location-colour' cueing condition was driven by higher error on congruent trials in this condition compared to other congruent conditions. Error was higher for the 'location-colour' congruent condition compared to the 'location-location' congruent condition (t_{20}=3.76 p=.001, d=.819, BF_{10}=29.83) and there was a similar trend supported by Bayes Factor, when compared to the 'colour-colour' congruent condition (t_{20}=3.08, p=.006, d=.672, BF_{10}=7.83), though not when compared to the 'colour-location' congruent condition (t_{20}= 1.97, p=.063, d=.430, BF_{10}=1.14). There were no significant differences between the incongruent conditions (p>.0031, 0.23>BF_{10}< 1.66).

To investigate the asymmetry in the magnitude of the congruency effect as a function of cue and probe features, I computed six pairwise contrasts (Bonferroni corrected alpha level=0.008) comparing the size of the congruency benefit across the four cue-by-probe feature combinations. The congruency effect on 'location-location' trials (M=-.327, SD=.197) was significantly larger than on 'location-colour' trials (M=-.113, SD=208):

t_{20}=3.58, p=.002, d=.782, BF_{10}=21.04, and there was a similar trend for a greater congruency effect on 'colour-colour' trials (M=-.217, SD=.228) than 'location-colour' trials (M=-.113, SD=208), supported by Bayes Factor: t_{20}=-2.76, p=.012, d=-.602, BF_{10}=4.29. There were no significant differences between the congruency effect on 'colour-colour' trials and 'colour-location' trials (M=-.224, SD=.223; p=.930, BF_{10}=.228), between 'colour-colour' and 'location-location' trials (M=-.327, SD=.197; p=.107, BF_{10}=.764), between 'colour-location' and 'location-colour' trials (p=.121, BF_{10}=.698), or between 'colour-location' and 'location-location' trials (p=.073, BF_{10}=1.02). In summary, the size of the congruency benefit in the 'colour-location' condition was comparable to the conditions where the cue- and probe share the same feature characteristics, while the size of the congruency effect in the 'location-colour' condition was smaller in comparison.

Reaction time

RT for orientation recall was calculated as the time from the probe presentation until the first rotation key press. I analysed median RTs in a 2×2×2 repeated measures ANOVA. A main effect of congruency showed that RTs were smaller for congruent (M=.785, SD=.215) than incongruent (M=.940, SD=.287) trials (Figure 2B & D): $F_{1,20}$= 29.15, p< .001, η^2= .593, BF_{10}=5.19×10^{11}. There was a main effect of probe feature, showing faster responses when participants were probed by colour (M=.829, SD=.260) than when they were probed by location (M=.895, SD=.264): $F_{1,20}$=9.26, p=.006, η^2=.316, BF_{10}=12.31. In addition, there was a significant interaction between the cue feature (colour vs location) and probe feature (colour vs location): $F_{1,20}$=12.26, p=.002, η^2=.38, BF_{incl}=.505.

The interaction between cue and probe feature was mostly driven by slowed RTs on 'colour-location' trials. Six pairwise comparisons (Bonferroni corrected alpha value=.008) revealed that RTs were significantly faster when cue and probe were both in the colour dimension ('colour-colour': M=.822, SD=.247) than when the incidental task was cued by

colour and orientation recall was probed by location ('colour-location': M=.912, SD=.255): t_{20}=-3.31, p=.003, d=-.723, BF_{10}=12.30. RTs were also faster for 'location-colour' trials (M=.861, SD=0.277) than 'colour-location' trials (M=.912, SD=.255): t_{20}=2.40, p=.026, d=.525, BF_{10}=2.30, though the effect did not survive correction for multiple comparisons and Bayes Factor indicated only limited evidence. There were no other significant pairwise effects or interaction effects for RTs (p>.05, .28>BF>1.07).

Size of congruency benefit correlates for same vs different cue-probe features

Although there was a significant congruency effect when the cue and probe had distinct cue features (specifically 'colour-location'), as predicted if attention spreads between an object's features in memory, it remains possible that the congruency effects when the cue and probe shared a feature dimension relied on a different mechanism (e.g. increased accessibility of cue feature) than when the feature dimension of the cue and probe were distinct (i.e. 'colour cue-location probe' and 'location cue-colour probe'). However, for both MAE and RT there was a significant correlation between the magnitude of the congruency benefit when the incidental cue and probe shared the same characteristics and when they had different characteristics (MAE: r(19)=.495, p=.022, BF_{10}=2.91; RT: r(19)=.781, p<.001, BF_{10}=569.31), suggesting that overlapping mechanisms support the congruency effect in the two conditions.

Mixture Modelling

Memory errors can arise from multiple potential sources such as the precision of the memory representation, mistakenly reporting a different item in memory, or guessing. To delineate between these potential sources of errors and how they were affected by incidental cueing, I fit a mixture model to the orientation recall errors for each participant (Bays et al., 2009). I analysed each of the four model parameters (precision, target rate, guess rate, non-target rate "swap errors") in separate 2×2 repeated measures ANOVAs.

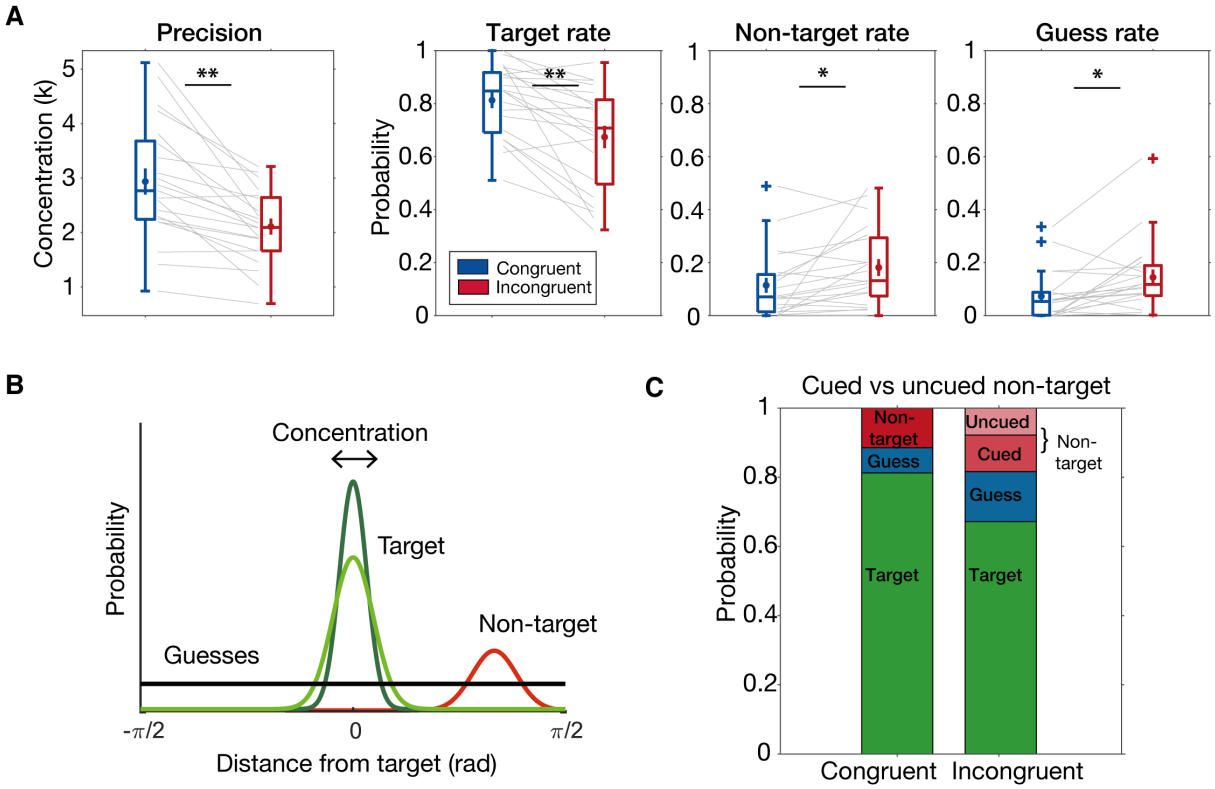

Figure 3. Mixture model fit. **A.** Boxplot of parameter estimates for each participant. Mean and standard error of mean overlaid. Grey bars represent each individual participant. **p<.001, *p<.05. **B**. Schematic of mixture model parameters. The uniform distribution represents the guess rate (blue), a Von Mises distribution centered around the target angle represents the target rate (green) and the width of the distribution represents the precision of the representation (concentration; dark green). A Von Mises distribution centered on the non-target represents the non-target rate (red). **C.** Mean target rate (green), guess rate (blue) and non-target rate (red) for congruent and incongruent trials. Cued (dark red) vs uncued (light red) non-targets represented separately for incongruent trials.

Congruency affected all sources of error (Figure 3). For each of the model parameters, there was a main effect of congruency, but no significant main effect or interaction with feature dimension (same vs different; p >.05, BF_{10}/BF_{incl} < 1; except congruency-cue dimension interaction for precision: BF_{incl}=1.56).

Precision was higher for congruent (M=3.09, SD=1.48) than incongruent trials (M=2.17, SD=.711): $F_{1,20}$=24.15, p<.001, η^2=.547, BF_{10}=312. Similarly, the target rate was higher for congruent (M=.808, SD=.157) than incongruent (M=.671, SD=.203) conditions: $F_{1,20}$=19.35, p<.001, η^2=.492, BF_{10}=3710. The probability of "swap errors" was lower for congruent (M=.123, SD=.135) than incongruent (M=.184, SD=.157) trials: $F_{1,20}$=5.82, p=.026, η^2=.225, BF_{10}=5.46. The guess rate (probability of responding with a random

orientation angle) was also lower for congruent (M=.069, SD=.096) than incongruent (M=.145, SD=.146) trials: $F_{1,20}$=13.04, p=.002, η^2=.395, BF_{10}=40.43.

To determine if swap errors in the incongruent condition were predominantly caused by participants incorrectly reporting the orientation of the *cued* item, I split the non-target parameter into two separate estimates, one centred on the invalidly cued item and one centred on the uncued item for each incongruent trial. While the probability of reporting the orientation of the *cued* non-target was higher (M=.106, SD=.114) than the *uncued* non-target (M=.078, SD=.057), this difference was not statistically significant (p=.255, BF_{10}=.416). However, the probability of reporting the invalidly *cued* non-target on incongruent trials (M=.106, SD=.114) was nevertheless greater than the probability of reporting either of the non-targets on congruent trials ($\frac{1}{2}$ NT rate on congruent trials=.057, SD=.064): t_{20}=-2.27, p=.035, d=-.495, BF_{10}=1.83, though evidence for this effect was small as indicated by Bayes Factor<2. The probability of reporting the *uncued* non-target on incongruent trials (M=.078, SD=.057) was also greater than the corresponding NT rate on congruent trials, but not significantly so: t_{20}=-1.88, p=.075, d=-.409, BF=.995. This indicates the congruency effect on errors was not merely caused by participants mistakenly reporting the cued item on incongruent trials.

Incidental task

I compared incidental task performance for colour vs location cues in paired samples t-tests. There was no significant difference in the number of errors of omission ("misses") for colour (M=20.00, SD=10.50) vs location (M=18.10, SD=10.57) cues (p=.349, BF_{10}=.342). However, participants made significantly more errors of commission ("false alarms" when the cue was a match) for colour cues (M=7.04, SD=4.49) than for location cues (M=2.98, SD=3.17): t_{20}= 4.56, p<.001, BF_{10}=156.26, likely because colours were more easily confused than locations. There was a trend towards faster RTs in response to

colour cues (M=.680, SD=.045) than location cues (M=.698, SD=.050; p=.057), but Bayes Factor indicated only limited evidence for this effect (BF_{10}=1.24).

Discussion (Experiment 1)

Experiment 1 showed that directing attention to one item-feature in WM brings associated features into a privileged state. The incidental task did not ask participants to retrieve any other features of the cued item. Nevertheless, when participants recalled the orientation of one of the objects, their precision was higher and responses faster when a feature of the same object had been cued during the maintenance delay (congruent) compared to when a different object had been cued (incongruent). The results are in line with previous studies of attention in WM that show behavioural benefits of non-predictive retro-cues such as 'refreshing cues' (Souza et al., 2015) and 'incidental cues' (Zokaei, Manohar, et al., 2014; Zokaei, Ning, et al., 2014).

While benefits of incidental cueing have previously been reported (Zokaei, Manohar, et al., 2014; Zokaei, Ning, et al., 2014), the current results demonstrate three new properties: First, the incidental task in the current study did not ask participants to report any other features of the cued item, ruling out the possibility that the congruency benefit is due to actively recalling the location of the incidentally cued item, as might have been the case in Zokaei, Ning, et al., (2014). Second, the congruency benefit was not merely explained by higher accessibility of the incidentally cued feature or strengthened cue-orientation associations as the congruency effect was similarly present when orientation recall was cued by a different feature than the incidental cue. The results suggest obligatory spread of attention between an object's features in WM as predicted by a recent computational model where retrieval is supported by auto-associative pattern completion (Manohar et al., 2019).

Finally, location cues, in addition to colour cues, were effective in bringing associated features into a privileged WM state. However, this depended on whether the incidental cue and the probe shared the same cue characteristics or not. While there was a robust congruency effect in the RTs independent of cue characteristic, for mean absolute error there was an asymmetry in the effect of congruency for colour and location cues when different features were cued: congruency improved performance when the cue was in the colour feature-dimension and orientation recall was probed by location, but not vice versa.

Asymmetry in the congruency benefit afforded by colour and location cues might be expected if there is a primacy of location in WM (Jiang et al., 2000; Pertzov & Husain, 2014; Treisman & Zhang, 2006; Wood, 2011). One model of feature-binding in WM suggests that non-spatial features are bound only through their shared location (Schneegans & Bays, 2017a). In line with this model, previous work has shown that memory errors for non-spatial features correlate with errors for the item's location, while memory errors for non-spatial features occur independently of each other (Bays et al., 2011; Fougnie & Alvarez, 2011). If non-spatial features are bound through their shared location, this might explain the discrepancy between congruency effects in 'colour-location' vs 'location-colour' conditions in the current design. Making a correct match/non-match judgement for a colour cue on the incidental task might require retrieving the location of the item and bringing it into an attended state, thereby facilitating performance when subsequently probed by location. If location can be accessed through either location×colour or location×orientation conjunctive "feature maps" (Schneegans & Bays, 2017a), the colour of the cued object is not necessarily refreshed by a location cue and may not benefit retrieval to the same extent when orientation recall is subsequently cued by colour.

"Swap errors", i.e. reporting the orientation of non-target items, were at least partly responsible for the reported congruency effect. Swap errors are consistent with theories proposing separate or parallel feature stores (Bays et al., 2009; Wheeler & Treisman, 2002) rather than models suggesting that objects are integrated in WM (Luck & Vogel, 1997). I did not find any difference in the likelihood of misreporting cued or uncued non-targets, indicating that participants did not merely report the orientation of the most recently attended item. Instead, cued items relative to uncued items were less susceptible to interference from other memory items. This may be expected if the cued item is maintained in the focus of attention until probe onset– a readily accessible state that does not require retrieval through an error-prone pattern completion process (Manohar et al., 2019; Wheeler & Treisman, 2002). Alternatively, the cue may strengthen connectivity between features belonging to the cued object, thereby increasing chances of retrieving the orientation of the correct item at probe, even in the absence of sustained attention (Manohar et al., 2019; Oberauer & Lin, 2017; Rerko & Oberauer, 2013; Souza et al., 2015).

In summary, Experiment 1 showed that a non-predictive cue, directing attention to an object-feature in memory, privileges associated features and makes them less susceptible to interference from other items. However, the underlying mechanisms of the incidental cueing effect remain unclear. Previous studies have investigated the neural correlates of the retro-cue benefit, but the neural mechanisms underlying facilitation by a non-predictive cue are not well understood. In Experiment 2, I measured EEG and eye-tracking to track the role of spatial- and feature-based attention during the delay in maintaining the cued object in a privileged state.

Experiment 2

Experiment 1 showed that selecting either a colour or location of an object stored in WM privileged other features belonging to the same object, even though these were currently redundant. In a second experiment, I measured electroencephalography (EEG) and eye-tracking while a new set of participants performed a modified version of the experimental task from Experiment 1 where the incidental cue was always in the colour dimension. I aimed to investigate the neural mechanisms supporting incidental cueing benefits, specifically the role of spatial attention and representational quality of object-features as decodable from patterns of EEG activity.

EEG has high temporal resolution and is therefore ideal for measuring rapid shifts of attention within WM. Desynchronization of oscillatory activity in the alpha frequency band is thought to reflect allocation of spatial attention (Jensen & Mazaheri, 2010), with relatively stronger desynchronization in the hemisphere contralateral to the attended spatial location (Worden et al., 2000). Previous studies have shown that predictive retro-cues are associated with alpha suppression contralateral to a cued location (Myers, Walther, et al., 2015; Poch et al., 2014; Wallis et al., 2015).

While most studies have used spatial retro-cues, a centrally presented colour retro-cue similarly modulated alpha lateralization, suggesting spatial attention is also involved in selecting non-spatial features from WM (Poch et al., 2017). This is further supported by a recent finding that colour retro-cues induces a small shift in gaze position toward the location at which the object was encoded, even though location information was never probed (van Ede, Chekroud, & Nobre, 2019). Covert shifts in spatial attention trigger micro-saccades toward the attended location (Corneil & Munoz, 2014; Engbert & Kliegl, 2003; Hafed & Clark, 2002). Selecting non-spatial content from WM may engage spatial attention in a way that "spills over" into the oculomotor system and can be tracked

through small shifts in eye-position (Van Ede et al., 2020; van Ede, Chekroud, & Nobre, 2019). If selecting a colour from WM in the current study similarly engages spatial attention, I expect to see contralateral alpha suppression as well as gaze bias toward the cued item's original position, even though the location of the item is irrelevant to completion of the incidental task during the delay.

Manipulating the focus of attention within WM also modulates the neural representational quality of WM content. Studies using multivariate analysis of fMRI and EEG initially indicated that only currently relevant information can be reliably decoded from neural activity (LaRocque et al., 2013; Lewis-Peacock et al., 2012; Serences et al., 2009; Sprague et al., 2016), leading to suggestions that stimulus-specific neural activity specifically codes for attended WM content (Larocque et al., 2014), while unattended but still relevant content may be maintained in an 'activity-silent' state, perhaps through short-term changes in synaptic connectivity (Kamiński et al., 2017; Mongillo et al., 2008; Rose et al., 2016; Stokes, 2015; Wolff et al., 2017). Although recent studies have found evidence for activity-based codes for unattended content in different brain areas (Christophel et al., 2018) or in a "inversed" neural code within the same area of cortex (van Loon et al., 2018; Q. Yu et al., 2020), it is nevertheless clear that the neural mechanisms supporting attended and unattended WM content show distinct signatures, particularly in sensory cortex.

If attentional selection spreads between features in WM, a non-predictive cue should be associated with greater EEG decoding quality for other features belonging to the cued item relative to features belonging to the uncued items. Alternatively, if individual features can be selectively attended without simultaneously retrieving currently irrelevant features belonging to the same object, there should be no difference in decoding strength between cued and uncued features. Given there is only limited evidence for multivariate decoding of colour WM in EEG (Hajonides, Nobre, Van Ede, &

Stokes, 2020), I modified the task from Experiment 1 so the incidental task always cued colour, boosting power to decode location and orientation features.

In summary, the behavioural results replicated those reported in Experiment 1. Furthermore, colour selection in WM engaged spatial attention and strengthened the neural representation of the cued item's location, in a way that predicted subsequent orientation judgements.

Methods

Participants

34 participants took part in Experiment 2. Four participants were excluded from analyses. One was excluded due to an EEG recording issue, two were excluded due to excessive eye movements during the task, and one was excluded due to low performance on the behavioural task. Of the 30 participants included in analyses, 18 were female and 12 were male. Mean age was 20.27 (SD=2.70). Three participants were left-handed and the remainder were right handed. All participants provided informed consent and were reimbursed at a rate of £15 per hour or with course credits. The study received ethical approval by the Oxford Central University Research Ethics Committee (R55073/RE003).

Procedure

I used the same experimental task from Experiment 1 (Figure 1) with a few modifications to suit EEG: 1) The incidental cue was always a centrally presented colour cue. This decision was made for two reasons: first, as there is limited evidence for multivariate colour decoding during the WM delay in EEG I wanted to improve power to detect location representations, and second, I wanted to make the task more manageable for participants. 2) Orientation recall responses were made with a dial (PowerMate,

Griffin Technology) instead of key presses. 3) The fixation period was lengthened and the duration jittered between 0.75-1.25 sec. 4) Visual feedback (sad/happy face) was provided for incidental task errors instead of audio feedback to reduce electrical noise and the feedback period was lengthened from 200 to 300 ms. 5) Finally, trial numbers were increased from 600 to 1080 total trials to improve power. As in Experiment 1, the 20% non-match incidental cue trials were skipped leaving 864 trials for the analyses. The trials were divided into 36 blocks of 30 trials and participants were encouraged to take breaks between trial blocks. In total the experimental task took about 2 hours to complete including breaks.

The task was presented on a 22 inch, 1680x1050 pixel monitor. Participants were seated 60 cm from the screen. The task was programmed and stimuli presented in Matlab R2015a (The MathWorks Inc., 2015) and Psychtoolbox 3.0.14 (Brainard, 1997a).

EEG was recorded at a sampling rate of 1kHz using Synamps amplifiers (Neuroscan, Inc., Albany, CA) and Curry Neuroimaging Suite (v7.0.12 X) software. Sixty-one electrodes were positioned according to the 10-10 system ("American Electroencephalographic Society guidelines for standard electrode position nomenclature.," 1991). Vertical eye-movements were recorded from EOGs above and below the right eye and horizontal eye-movements were recorded from EOGs next to each eye. A ground electrode was placed on the right elbow and the left mastoid was used as the active reference.

Participants' eye position was continuously recorded using an eye tracking camera (EyeLink) at a rate of 1kHz.

Behavioural analyses

Mean absolute error (MAE) and median reaction times (RT; time from probe onset until first movement of dial) as well as the parameters from the Mixture Model (precision,

target rate, non-target rate, and guess rate) were each analysed in a 2x2 repeated measures ANOVAs with factors congruency (congruent vs incongruent) and probe feature (colour vs location). Recall that the cue feature was always 'colour' in this experiment.

EEG pre-processing

ERPs were re-referenced offline to the mean of the right and left mastoid and down-sampled to 250 Hz. For the two EOG channels I applied a high-pass filter of 1 Hz and a low-pass filter of 40 Hz. For the 61 EEG channels, I applied a high-pass filter of .1 Hz and a low-pass filter of 45 Hz. The data were epoched around the onset of the WM array (-1.0 to 5.5 seconds). Visual inspection was performed for each trial to remove trials with nonstereotyped artefacts or saccades in time periods of interest (cue and delay periods). Independent component analysis (ICA) was performed to remove eye-blinks and noisy components from the data. Pre-processing steps were performed in Matlab R2017b (The MathWorks Inc., 2017) and EEGLab (Delorme & Makeig, 2004). Data were then converted into a format compatible with the Fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011) and a further pre-processing step was performed to exclude any trials that displayed a high degree of variance.

Eye-tracking pre-processing

The EyeLink EDF file was converted to ascii format using the edf2asc EyeLink program. Where both eyes were recorded, I took the mean of the left and right eye-traces. As for the EEG data, eye position data were epoched around the onset of the WM array (-1.0 to 5.5 seconds). Eye-blinks were interpolated using linear interpolation. Data were then converted into a format compatible with the Fieldtrip toolbox and down-sampled to 250 Hz to be compatible with the EEG data. Trials with missing data were excluded from eye-tracking analyses.

Time frequency decomposition and alpha lateralization

To track spatial attention following incidental cue onset, time-frequency decomposition was performed to compute an index of alpha lateralization. First, the surface Laplacian was computed (Perrin, Pernier, Bertrand, & Echallier, 1989), which is a high-pass spatial filter that sharpens the EEG topography by heightening local signals and filtering out more distant signals due to volume conduction (Kayser & Tenke, 2015). I estimated the surface Laplacian using the ft_scalpcurrentdensity fieldtrip function in Matlab, using the spline method (Oostenveld et al., 2011; Perrin et al., 1989) with parameters 10th-order Legendre polynomial and a lambda of 10^{-5} (de Vries, van Driel, & Olivers, 2017).

Surface Laplacian data were used to perform the time-frequency analysis. Spectral decomposition of the data was performed in Matlab with the FieldTrip toolbox (Oostenveld et al., 2011). I applied Hanning tapers with a time window width of five cycles per frequency, with frequencies of interest between 1 Hz and 45 Hz in steps of 1 Hz. The power at each frequency was log-transformed ($10\times\log_{10}$) and mean power was calculated over contralateral and ipsilateral posterior channels (P8/7, P6/5, P4/3, PO8/7, PO4/3, O2/1) relative to the original position of each of the items in memory (cued, uncued 1, uncued 2) on each trial. To get a lateralization index relative to each item in memory, I took the difference between contralateral and ipsilateral power relative to the hemifield where each item was originally presented. Alpha lateralization was calculated by taking the mean over frequency bands from 8-14 Hz. Finally, mean alpha lateralization was computed over a time-window of interest (TOI) from 0.2-1s from cue onset.

I tested for significant contralateral alpha suppression relative to the cued item using a one-sample t-test (one-tailed, <0) and tested for any lateralization relative to each of the uncued items (uncued 1, uncued 2; randomly assigned) using one-sample t-tests (two-

tailed). To test whether alpha lateralization depended on whether the item was cued or not, I compared alpha lateralization relative to the position of each of the three memory items with a one-way ANOVA. This was followed up by three paired samples t-tests (two-tailed) to test for significant differences between the cued item and each of the uncued items (one-tailed, cued<uncued) and between the two uncued items (two-tailed). Finally, cluster-corrected t-tests were performed across the post-cue time-window (0-2 s) to test for significant lateralization relative to the cued and uncued (averaged) memory items and for any differences between them (two-tailed, 10,000 permutations, cluster-forming threshold =.05).

Gaze bias

Microsaccades were detected using a Matlab function (Engbert & Mergenthaler, 2006) based on two-dimensional velocity vectors computed from eye-position data using moving averages (velocity type=2). I used standard settings of velocity threshold (λ) of 6 standard deviations of the velocity distribution, and minimum duration of 12 ms. Next, position data were down sampled to 250 Hz and baseline corrected by subtracting the mean position from 200-100ms before WM array onset from the signal. Eye-position data were normalized such that 100% equaled the radius of the position of the WM items on the screen (i.e. 7 degrees visual angle). Probability histograms for each participant were computed for the saccade amplitude and saccade onset over time before taking the mean across participants. Probability histograms were smoothed with a gaussian kernel for illustration purposes (SD=12.5 ms, SD=1% stimulus eccentricity). Heat maps of gaze density were created by computing the 2D histograms of eye-position from 0.4-1s from incidental cue onset (time-window of interest based on (van Ede, Chekroud, & Nobre, 2019)) separately for each location of the cued memory item. Histogram counts were converted to density by dividing by the trial number for each location condition before subtracting out density common to all locations. Finally, I took the mean across

participants. Density maps were smoothed by a 10×10% box car filter for illustration purposes.

Location decoding

To test the extent to which activity patterns in the EEG channels and eye-coordinates contained information about the locations in memory, I used Mahalanobis distance to calculate the trial-wise distances between the multidimensional activity at each time point for each of the six possible locations in memory distributed evenly on a circle (Wolff et al., 2017). I performed the below decoding procedure for the EEG voltage data and eye-position data. The data were mean-centred by subtracting the mean activity at -0.2:-0.1 s before WM array onset for each trial. For EEG decoding, I performed the analyses on the 17 posterior channels (P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2; Muhle-Karbe, Myers, & Stokes, 2020; Wolff, Jochim, Akyürek, Buschman, & Stokes, 2020; Wolff et al., 2017; Wolff, Kandemir, Stokes, & Akyürek, 2020). For the eye-position data, the X and Y coordinates were used as channels for the decoder. Decoding was done separately and independently for each participant.

The decoding procedure followed an 8-fold cross-validation approach to calculate the decoding strength of the location of interest for each trial. The pattern of activity of the trials in the testing fold was compared to the trials in the k-1 training folds at each time-point. Equal trial-numbers for each location was ensured by random subsampling. I computed the Mahalanobis distance between the test trial and the average response for each location in the training data using the covariance matrix estimated from trials in the k-1 training folds, using a shrinkage estimator (Ledoit & Wolf, 2004). To obtain a visual representation of a tuning curve, the six distances were centered relative to the item location on each test trial, mean-centered, and sign-reversed, so higher values represent greater similarity between the test trial and the training set. This procedure was repeated for all train and test fold combinations. To obtain reliable estimates, the

above procedure was repeated 50 times, using random folds each time, and averaged across the resulting 50 repetitions to obtain a single tuning curve with the six Mahalanobis distance values for each trial. To obtain a measure of decoding strength, I computed the cosine vector means of the tuning curves (Sprague et al., 2016; Wolff et al., 2017). Higher values reflect greater location tuning and thus greater decoding strength for that trial. I took the mean over trials, resulting in a single decoding value for each time-point for each participant. Finally, I took the mean over a time-window of interest from 0-1s following WM array onset and 0-1s following cue onset. Time-resolved decoding values were smoothed for visualisation and statistical testing using a gaussian kernel (SD=12.5 ms).

During the WM array epoch, I expected decoding quality to be significant for all three items and there should be no difference between them. During the cue epoch, I expected decoding quality to be significant for the cued memory item and decoding quality to be significantly greater for the cued than each of the uncued items. I tested for significant decoding for each of the items in memory for the WM array epoch and the cue epoch using one-sample t-tests (one-tailed, >0) and tested for any differences between them using a one-way ANOVA. For the cue epoch, the ANOVA was followed up with paired samples t-tests to test for pairwise differences in decoding strength between the cued item and each of the uncued items (one-tailed, cued>uncued) and between the two uncued items (two-tailed). As the two uncued items were arbitrarily assigned, there should be no difference between them. Finally, cluster-corrected t-tests were performed across the post-cue time-window (0-2 s) to test for significant location decoding for the cued and uncued (averaged) memory items and for any differences between them (two-tailed, 10,000 permutations, cluster-forming threshold =.05).

Eye-position may confound neural decoding from EEG, even when limiting analyses to posterior channels (Mostert et al., 2018; Quax, Dijkstra, van Staveren, Bosch, & van

Gerven, 2019). Consequently, I ran two control analyses: 1) regressing out eye-position and, 2) locking decoding analyses to microsaccade onsets (Appendix A; Supplementary Material) to delineate the extent to which the neural decoding results uncovered information beyond retinotopic position. The cued item's location could be decoded even after regressing out decoding from eye-position and the item's location was decodable from EEG voltages before the first post-cue micro-saccade onset (Appendix A; Figure S1).

Cross-temporal decoding of location

Cross-temporal decoding is a useful method for exploring how mental representations evolve over time (King & Dehaene, 2014). To investigate whether location representations that re-emerge following the cue, as decodable from eye-position and EEG, share the same coding format as location representations decodable during WM encoding and the pre-cue delay, I tested the extent to which location decoding cross-generalized across time points. I ran the same decoding procedure as described above, but instead of training and testing the decoder separately at each time point, I tested the decoder on all time-points for each training time point. To reduce computation time, I down sampled the data to 100Hz before running the analysis. I performed cluster-corrected t-tests across the full training×testing time matrix (0-3.75s from WM array onset; two-tailed, 10,000 permutations, cluster-forming threshold=.05) Decoding along the diagonal suggests dynamic coding, while off-diagonal decoding is indicative of a more stable coding scheme in WM (King & Dehaene, 2014). Finally, to quantify cross-temporal decoding strength, I took the mean decoding value when training the decoder on the WM array epoch (0-1 s from array onset) and testing the decoder on the cue epoch (1.75-2.75 s from array onset; Figure 6B red box) and tested for significant cross-generalization using a one-sample t-test (one-tailed, >0).

Orientation decoding

To test the extent to which activity patterns in the EEG channels contained information about the orientation angles in memory, I followed the same decoding procedure as described for location decoding above, with a few modifications. As the orientation stimuli were drawn at random between 1-180 degrees, orientations were first sorted into eight bins of 22.5 degrees. Previous studies have shown that orientation decoding does not cross-generalize across encoding locations (Bae & Luck, 2017; Wolff, Jochim, et al., 2020), so the decoding procedure was performed separately and independently for each of the six locations in memory. To obtain reliable estimates, the decoding procedure was repeated 50 times (with random folds each time), separately for two orientation spaces (bin centres: 0° to 157.5° in steps of 22.5° and 11.25° to 168.75° in steps of 22.5°) and averaged across the 100 resulting samples for each of the eight Mahalanobis distance values. Decoding values were smoothed for visualisation and statistical testing using a gaussian kernel (SD=10 ms). The remaining analysis steps and statistical tests were the same as for location decoding described above.

Given weak decoding results for orientation using the standard decoding approach described above, I added a complementary spatiotemporal decoding approach as an exploratory analysis. Recent literature has shown that decoding accuracy may be boosted by taking advantage of the dynamic nature of event-related potentials (ERPs)(Grootswagers, Wardle, & Carlson, 2016; Wolff, Jochim, et al., 2020; Wolff, Kandemir, et al., 2020b). This approach pools the relative voltages over time and space, thus leveraging information encoded in temporal as well as spatial patterns. I focused on a 400 ms time window from 100-500 ms following onset of the event of interest (WM array onset and cue onset). First, the data were down-sampled by taking the average every 16 ms. I then zero-centered the voltage fluctuations within the time-window by subtracting the mean voltage over the full time-window from the voltage at each time

point. This relative baseline is comparable to pre-stimulus baseline correction, common in ERP analyses, but further removes stable neural activity within the time-window and is optimised for decoding stimulus-evoked, dynamic activity. The resulting 25 values for each posterior channel were concatenated and used as the input to the multivariate decoder. The decoding procedure followed the same steps described above, but resulted in a single decoding value for each time-window of interest.

Statistical comparisons were the same as those described for location decoding above.

Relationship between alpha lateralization/gaze bias/decoding and behaviour (error/RT)

To test if alpha lateralization, gaze bias and decoding quality for the cued item was related to behavioural performance on the orientation recall task, two general linear models were fitted for each participant with trial-wise alpha lateralization/gaze bias/decoding values averaged over the time-window of interest as the predictor variable (0.2-1 s from cue onset for alpha lateralization; 0-1 s from cue onset for location/orientation decoding analyses). The outcome variable was log-transformed RTs in one model and absolute error in the other. The design matrices included a constant term. The models were fitted separately for each condition (congruent vs incongruent; colour vs location probe) and β weights were obtained for each participant.

I performed one-sample t-tests on the β values. For alpha lateralization, I expected contralateral alpha suppression to predict lower RT/error on congruent trials (one-tailed, >0) and higher RT/error on incongruent trials (one-tailed, >0). I expected greater gaze bias and decoding quality following the cue to predict lower RT/error on congruent trials (one-tailed, <0) and higher RT on incongruent trials (one-tailed, <0).

To test whether potential relationships with behaviour were specific to particular time-points, I fit the regression models over time to obtain β weights for each time point across the time-window of interest (0-2 s following cue onset) and performed cluster-

corrected t-tests (two-tailed, cluster-forming threshold=.05, 10,000 permutations). Time-resolved β values were smoothed with a Gaussian kernel (SD=12.5 ms) for visualisation purposes and statistical testing.

While I expect the regression analysis to be most powerful to detect trial-wise variance, it assumes a normal distribution across trials. However, the distribution of trial-wise RTs and absolute error tends to be positively skewed. Consequently, I performed a secondary analysis, less sensitive to potential skews in the distribution. For each participant, I split trial-wise measures based on the median RT/absolute error and calculated the mean alpha lateralization/gaze bias/decoding values for high and low RT/error trials respectively, separately for each condition. I performed a 2×2×2 repeated measures ANOVA with factors congruency (congruent vs incongruent), probe type (colour vs location) and RT/error (low vs high). Planned one-sample t-tests were performed to check for significant alpha lateralization (one-tailed, <0), gaze bias (one-tailed, >0) and decoding quality (one-tailed, >0) for each condition and paired-samples t-tests were performed to check for significant differences between high and low RT/error trials (two-tailed).

Results

Behaviour

- *Mean absolute error*

Figure 4A shows the distribution of orientation recall errors across participants. Descriptive statistics for mean absolute error (MAE) are shown in Table 2. As in Experiment 1, MAE was significantly lower for congruent (M=.662, SD=.138) than incongruent (M=.807, SD=.177) trials: $F_{1,29}$=61.50, p< .001, η^2=.680, BF_{10}=1.41×10^{12} (Figure 4B). In contrast to Experiment 1, there was a main effect of probe feature

showing that MAE was lower when orientation recall was probed by colour ('colour-colour': M=.724 , SD=.152) compared to location ('colour-location': M=.793, SD=.170): $F_{1,29}$=27.69, p=.001, η^2=.488, BF_{10}=11.61. Perhaps colour was a more salient feature in this experiment because the incidental task always cued colour and never location. The congruency effect was slightly larger when orientation recall was probed by location (M=-.162, SD=.132) than when probed by colour (M=-.128, SD=.087), but the interaction between congruency and probe feature on MAE was not significant ($F_{1,29}$=3.947, p=.059, η^2=.120, BF_{incl}=.608). These results match those of Experiment 1 where there was no significant difference in the size of the congruency benefit on 'colour-colour' probe trials and 'colour-location' probe trials.

Table 2. Descriptive statistics for mean absolute error (rad) and RT (s) in Experiment 2. N = 30.

	Mean absolute error				Reaction time			
	Congruent		Incongruent		Congruent		Incongruent	
Probe feature	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Colour	.639	. 126	.766	.171	.795	.156	.888	.171
Location	.685	.158	.846	.192	.863	.189	.928	.216

- *Reaction time*

Table 2 shows descriptive statistics for RTs. RTs were significantly lower for congruent (M=.829, SD=.175) than incongruent trials (M=.908, SD=.194), in line with results from Experiment 1: $F_{1,29}$=30.39, p<.001, η^2=.512, BF_{10}=1.69 ×10^6 (Figure 4C). There was a main effect of probe feature showing that RTs were significantly lower for colour probes (M=.842, SD=.169) than for location probes (M=.896, SD=.204): $F_{1,29}$=28.04, p<.001, η^2=.492, BF_{10}=140.24, matching the results of Experiment 1 where RTs were slowed on 'colour-location' trials compared to 'colour-colour' trials. Additionally, the congruency effect was larger when recall was probed by colour (M=-.094, SD=.090) than when probed by location (M=-.065, SD=.084), as indicated by a significant interaction between

congruency and probe feature: $F_{1,29}$=4.72, p=.038, η^2=.140, BF_{incl}=.507. A similar pattern was found in Experiment 1, though not significant. The interaction between congruency and probe-feature was followed up with four pairwise comparisons (corrected alpha level =.0125). RTs were faster for congruent than incongruent trials, whether orientation recall was probed by colour or location: ('colour-colour: t_{29}=-5.72, p<.001, d=-1.04, BF_{10}=5581, 'colour-location': t_{29}=-4.29, p<.001, d=-.783, BF_{10}=151). For both congruent and incongruent conditions, RTs were faster for colour-probe trials than location-probe trials (congruent: t_{29}=-5.39, p<.001 d=-.984, BF_{10}=2408.45; incongruent: t_{29}=-3.46, p=.002, d=-.632, BF_{10}=20.79).

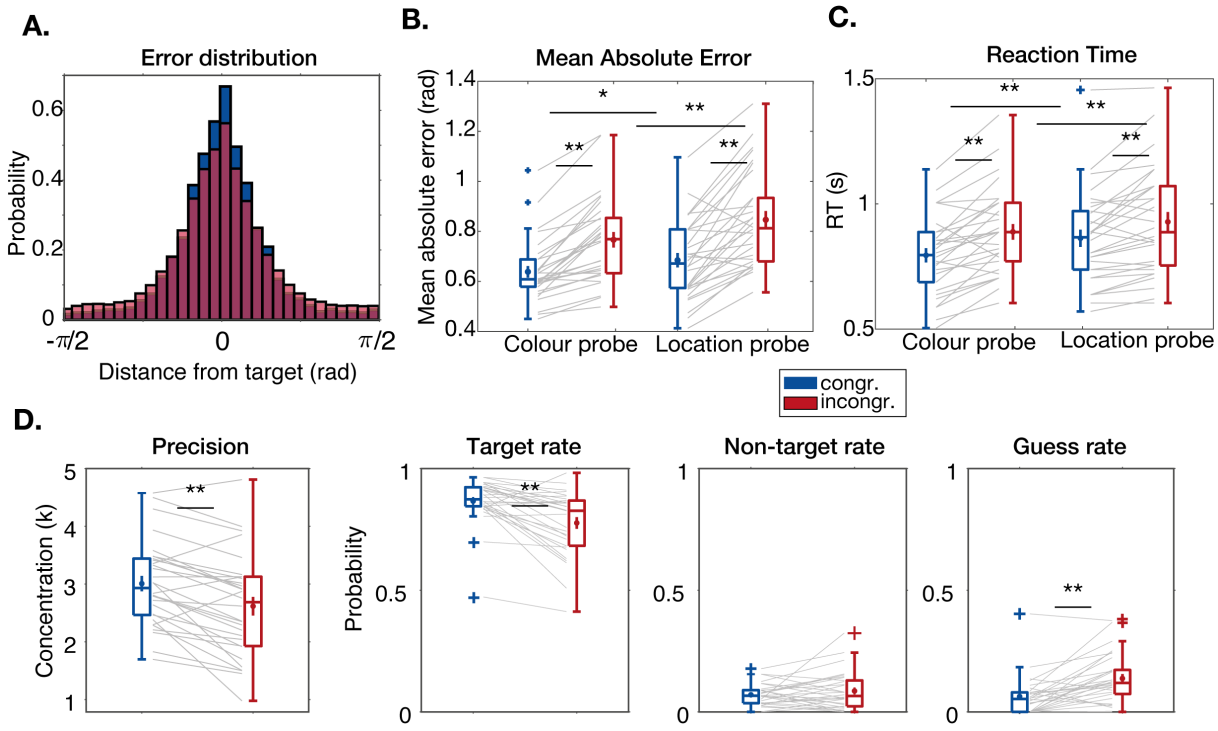

Figure 4. Behavioural results of Experiment 2. **A.** Distribution of errors for all participants relative to target angle for congruent (blue) and incongruent (red) trials. **B.** Boxplot of mean absolute error (rad) for congruent (blue) and incongruent (red) trials, separately for orientation recall probed by colour and location. **C.** Same as B panel but for reaction time (s). **D.** Boxplot of parameters of the mixture model fit for each participant. Mean and standard error across participants overlaid. Grey lines show individual participants. *p<.05, **p<.001.

- *Mixture Model*

Table 3 shows descriptive statistics for the parameters of the mixture model. As expected, orientation judgements were more precise and more likely to be centered on the target orientation for congruent than incongruent trials as reflected in a main effect

of congruency on precision ($F_{1,29}$=22.18, p<.001, η^2=.433, BF_{10}=162) and target rate ($F_{1,29}$=19.84, p<.001, η^2=.406, BF_{10}=3865). Additionally, there was a main effect of congruency on guess rate ($F_{1,29}$=21.95, p<.001, η^2=.431, BF_{10}=875), but no main effect of congruency on the non-target rate (p=.293, BF_{10}=.381), which differs from Experiment 1.

Table 3. Descriptive statistics for mixture model parameters in Experiment 2. N = 30.

	Precision				Target rate			
	Congruent		Incongruent		Congruent		Incongruent	
Probe feature	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Colour	3.177	.838	2.858	1.118	.879	.083	.799	.132
Location	2.973	.787	2.493	.790	.850	.118	.758	.169

	Non-target rate				Guess rate			
	Congruent		Incongruent		Congruent		Incongruent	
Probe feature	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Colour	.080	.057	.094	.064	.041	.061	.106	.103
Location	.065	.065	.077	.096	.084	.113	.165	.127

Orientation judgements were more precise and target rate higher when orientation recall was probed by colour than by location, revealed by a significant main effect of probe feature for precision ($F_{1,29}$=10.09, p=.004, η^2=.258, BF_{10}=4.99) and a trend for target rate ($F_{1,29}$=4.07, p=.053, η^2=.123, BF_{10}=.798). In addition, the guess rate was lower when participants were probed by colour than by location ($F_{1,29}$=10.70, p=.003, η^2=.270, BF_{10}=8.68). There was no main effect of probe feature on non-target rate (p=.105, BF_{10}=.518). This pattern of results reflects the lower mean absolute error for colour than location probes reported above and suggests it is attributable to differences in precision, target rate and guess rate. However, they diverge from the results of Experiment 1 where there was no main effect of probe feature on parameters of the mixture model. There were no interactions between congruency and probe feature for any of the model parameters (p>=.476, BF_{incl}=<.326).

Alpha lateralization relative to cued item location

If colour-selection in WM engages spatial attention, I would expect to see alpha suppression contralateral to the hemifield where the cued item originally appeared. Figure 5A shows lateralized (contra minus ipsilateral) oscillatory power in the frequency range 5-45 Hz relative to the original position of the cued item. Cluster-corrected one-sample t-tests over the post-cue window (0-2 s) and frequency range 5-45 Hz revealed significant contralateral suppression (contra < ipsi power) in the alpha-beta range (7-22 Hz) from 0.37 to 0.77 s (p=.004) after cue onset and significant contralateral enhancement (contra > ipsi power) in the alpha-beta range (7-33 Hz) from 0.77 to 1.69 s (p<.001) after cue onset.

I took the mean over the alpha frequency band (8-14 Hz; Fig 5B) relative to the location where the cued item was originally presented (blue) and relative to where each of the two uncued items were originally presented (pink; uncued averaged) and computed mean alpha lateralization over the time window of interest (TOI) from 0.2-1s following cue onset. As expected, there was significant alpha suppression contralateral to the original position of the cued item (t_{29}=-2.95, p=.003, d=-.538). Each of the uncued items instead showed the opposite pattern with greater contra- than ipsilateral alpha power relative to the item's original position, though this did not reach significance (uncued 1: p=.061, uncued 2: p=.059). Recall that the uncued items were arbitrarily assigned to uncued 1 or 2, so there should be no difference between them. Greater contra- than ipsilateral alpha power relative to the uncued memory items could be due to the fact that the cued item occupied one of the available locations such that uncued items were more likely to appear on the hemifield opposite of the cued item (3/5) compared to the same hemifield as the cued item (2/5).

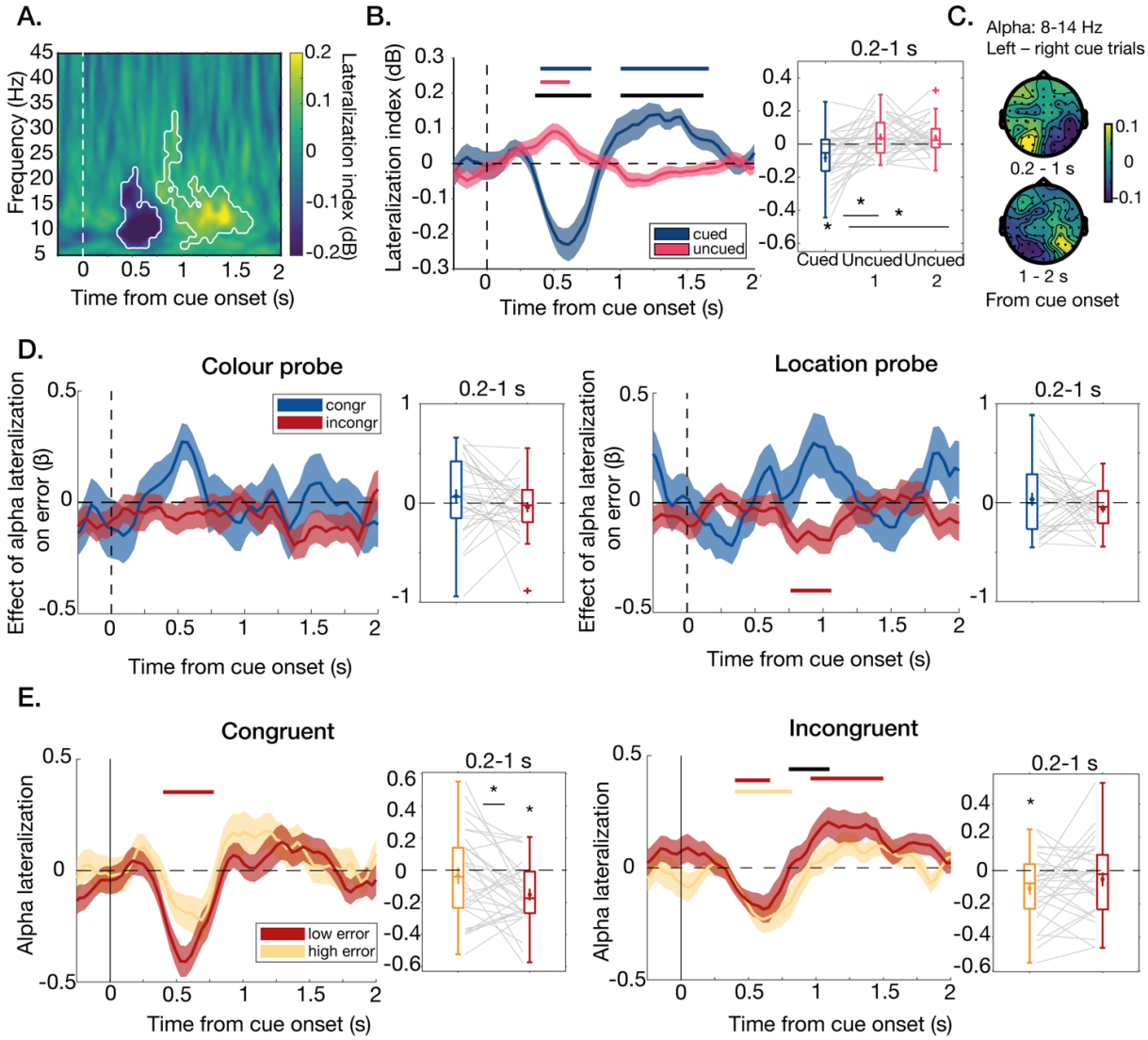

Figure 5. Alpha lateralization. **A.** Contralateral minus ipsilateral power relative to the hemifield where the cued item was presented at encoding. White borders show clusters significantly different from zero (corrected). **B.** Left: Lateralized alpha power (8-14 Hz) over time following cue onset relative to the memorized position of the cued object (blue) and two uncued objects (averaged; pink). Blue and pink bars show non-zero alpha lateralization (cluster-corrected) and black bars show significant differences between cued and uncued items. Shaded area shows standard error of mean (SEM). Right: Boxplot of mean alpha lateralization for the time window of interest (0.2-1 s from cue onset) relative to each memory item. Mean and SEM across participants overlaid. Grey lines show individual participants. **C.** Topography of alpha power for left minus right cue trials (top: 0.2-1s; bottom: 1-2s from cue onset). **D.** Beta weights from trial-wise regression with time-resolved alpha lateralization relative to cued item and absolute error for congruent (blue) and incongruent (red) colour-probe trials (left) and location-probe trials (right). Red/blue bars show non-zero beta weights (cluster-corrected). Boxplots shows beta weights from regression with mean alpha lateralization in across time of interest (0.2-1s) and absolute error. **E.** Alpha lateralization for high (yellow) vs low (red) error trials for congruent (left) and incongruent (right) trials averaged over colour and location probe trials. Red and yellow bars show significant alpha lateralization (corrected). Boxplots show mean alpha lateralization across time of interest separately for high (yellow) and low (red) error trials. *p<.05.

A one-way ANOVA comparing alpha lateralization relative to each of the items in memory showed a main effect of item: $F_{2,29}$=8.98, p<.001. Three planned pairwise

comparisons revealed a significant difference in alpha lateralization between the cued item and each of the uncued items (cued vs uncued 1: t_{29}=-3.84, p<.001, d=-.702; cued vs uncued 2: t_{29}=-3.17, p=.002, d=-.579), but no significant difference in lateralization between the two uncued items (p=.885) as expected. These results indicate that the incidental colour cue triggered alpha suppression contralateral to the position where the cued item was encoded, consistent with a role of spatial attention in selecting the cued colour from memory. As shown in Figure 5C, modulation in the alpha frequency band was greatest over posterior EEG channels covering occipital and parietal cortex.

Cluster-corrected t-tests across the cue epoch (0-2 s from cue onset) revealed a cluster of significant contralateral alpha suppression for the cued item initially following cue onset from 0.4 to 0.78 s (p=.002). This was not the case for uncued items (averaged), which instead showed greater contra- than ipsilateral alpha power from 0.4 to 0.62 s from cue onset (p=.037) and there was a significant difference between cued and uncued items from 0.36 to 0.78 s (p<.001). Surprisingly, later in the delay after the incidental task decision period was completed, lateralization relative to the cued item's original position reversed and instead showed greater contra- than ipsilateral alpha power from 1 to 1.66 s (p<.001). This resulted in a significant difference between alpha lateralization relative to the cued and uncued items from 1 to 1.62 s following cue onset (p<.001).

I speculated this could be due to the non-predictive nature of the cue and reflect re-orienting of spatial attention towards the original position of other items in memory to refresh these in preparation for the probe. To explore this possibility, trials were split according to the original position of the two uncued items in memory relative to the position of the cued item (Supplementary material, Figure S2). This resulted in three conditions: both uncued items were in the same hemifield as the cued item (load 2), one uncued item was in the same hemifield as the cued item (load 1), or the two uncued items were in the opposite hemifield (load 0). I computed mean alpha lateralization for

each of these conditions. Contralateral alpha suppression was significant only for the load 0 (t_{29}=-3.77, p<.001, d=-.689) and the load 1 conditions (t_{29}=-2.15, p=.040, d=-.393), while lateralization following the incidental cue for the load 2 condition was non-significant (p=.306; Figure S2B). Similarly, there was only a significant reversal in lateralization for the load 0 (t_{29}=-2.15, p=.015, d=.474) and load 1 conditions (t_{29}= 4.12, p<.001, d=.753), where the memorised location of at least one other item was in the opposite hemifield, but not in the load 2 condition (p=.383; Figure S2C). However, there was no significant difference in lateralization between any of the three conditions initially following the incidental cue or later in the delay (p>.05). I interpret these results with caution as the analyses were post-hoc and trial-numbers were not matched between conditions.

Post-cue alpha lateralization predicts behavioural error

If post-cue alpha lateralization plays a role in selecting the cued item and bringing it into a privileged state in WM, contralateral alpha suppression should be associated with faster and more precise performance on congruent trials ($\beta > 0$). If selection of the cued item comes at the expense of the uncued items in memory, alpha suppression might further be associated with slowed and less precise performance on incongruent trials ($\beta < 0$). The results of the regression analysis between alpha lateralization and absolute error are presented below. In addition, there was some evidence to suggest contralateral alpha suppression predicted slowed RTs on incongruent colour probe trials (Appendix A; Supplementary Figure S3).

While the relationship was in the predicted direction for congruent trials (Figure 5D), β weights were not significantly greater than zero for either congruent colour-probe trials (p=.171) or location-probe trials (p=.304). There was no relationship between alpha lateralization and absolute error for incongruent colour-probe trials (p=.214). On incongruent location-probe trials, contralateral alpha suppression was associated with

higher error, but this effect did not reach significance (t_{29}= -1.58, p=.062, d=-.289). Collapsing the data across location and colour probe trials, β weights were still not significantly greater than zero for congruent (t_{29}=1.35, p=.094, d=.246) or significantly below zero for incongruent trials (t_{29}=-1.37, p=.090, d=-.250), but the trends were in the expected direction. A 2×2 repeated measures ANOVA with congruency (congruent vs incongruent) and probe type (colour vs location) as factors revealed no significant main effect of congruency (p=.136), no main effect of probe type (p=.625) and no interaction (p=.890).

For completeness, the regression model was fit over time to obtain β weights for each time point during the delay following cue onset (Figure 5D). For both colour and location probe trials, positive β weights following cue onset indicated that contralateral alpha suppression relative to the cued item predicted lower error on congruent trials, though slightly earlier for colour-probe than location-probe trials. However this relationship was not significant (0-2 s). For incongruent location-probe trials, there was a significant cluster of negative β weights later in the delay from 0.76 to 1.06 s following cue onset (p=.023), indicating that contralateral alpha suppression relative to the cued item predicted lower error on incongruent trials.

Figure 5E shows alpha lateralization over trials with high and low error, separately for congruent (left) and incongruent trials (right). A 2×2×2 repeated measures ANOVA revealed no significant main effects and no significant interactions, although there was a trend for an interaction between congruency and error ($F_{1,29}$= 3.48, p=.072). On congruent trials, there was significant alpha suppression contralateral to the cued item on low error trials (t_{29}=-4.19, p<.001, d=-.765) but not on high error trials (p=.259) and suppression was greater on low vs high error trials (t_{29}= 2.39, p=.024, d=.436). Cluster-corrected t-tests across the post-cue delay (0-2 s post cue) for congruent trials revealed a significant cluster of contralateral alpha suppression for low error trials from 0.4 to 0.78

s after cue onset (p<.001), and no significant suppression for high error trials. Conversely, on incongruent trials, alpha suppression contralateral to the cued item was significant only on high error trials (t_{29}=-2.90, p=.003, d=-.529) and not on low error trials (p=.107), but these were not significantly different (p=.213). Cluster-corrected t-tests across the post-cue delay (0-2s) revealed significant contralateral alpha suppression for both low error (0.4-0.66 s; p=.039) and high error trials (0.4-0.82 s; p=.004) early in the delay. Interestingly, subsequent reversal of alpha lateralization later in the delay was significant only for low error incongruent trials (0.96 1.5 s; p=.003), not for high error incongruent trials and these were significantly different from 0.8 to 1.1 s (p=.049).

The results provide some evidence that contralateral alpha suppression supports selection of the cued item in a way that facilitates memory precision, but comes at a cost to other items in memory. However, reversal of alpha lateralization later in the delay may help remedy costs for uncued items as the strength of the reversal was associated with lower error on incongruent trials.

Decoding memorised location of cued item from EEG voltages

If selecting the cued colour from WM strengthens the neural representation of the item's location, EEG decoding quality for the cued item's location, relative to uncued items, should improve following the cue. Following the onset of the WM array (TOI: 0-1s from array onset; Figure 6A, left), the location of all three items in memory were decodable over posterior EEG channels (cued: t_{29}=9.99, p<.001, d=1.82; uncued 1: t_{29}= 9.99, p<.001, d=1.83; uncued 2: t_{29}=10.84, p<.001, d=1.98) and as expected there was no difference in decoding strength between the three items (one-way ANOVA: p=.638). Cluster-corrected t-tests across the WM array epoch (0-1.5 s) revealed significant decoding for the cued item from 0.015 to 1.11 s (p<.001) and for the uncued items (averaged) from 0.015 to 1.28 s (p<.001).

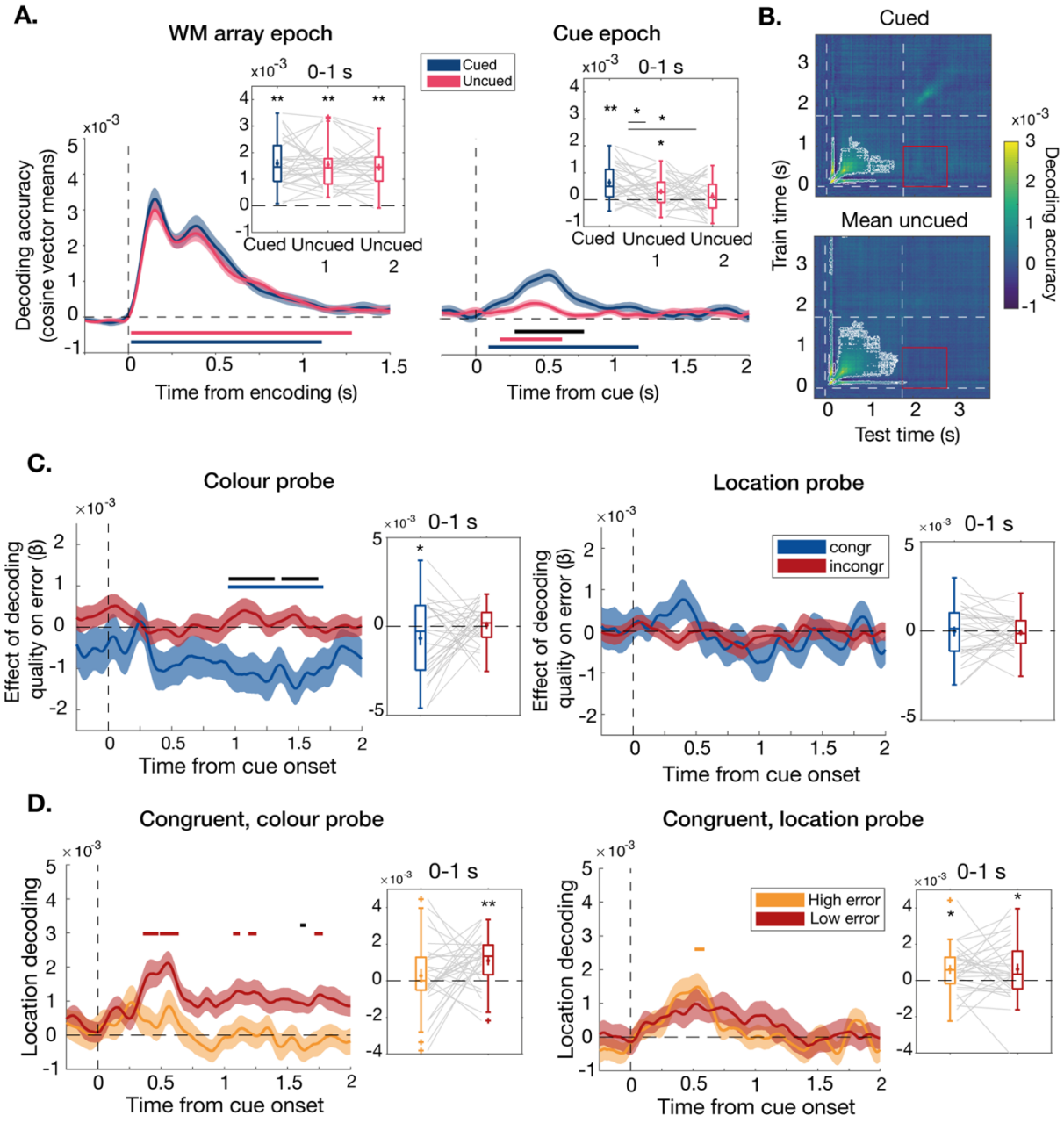

Figure 6. Location decoding from EEG voltages. **A.** Decoding quality for locations in memory following encoding (left) and cue onset (right) for cued item (blue) and uncued items (pink; averaged in time-resolved plots). Shaded area shows standard error of mean (SEM). Blue and pink bars show significant decoding. Black bar shows significant difference between cued and uncued items. Boxplots show mean decoding across TOI (0-1 s) following WM array onset (left) and cue onset (right). Mean and SEM overlaid. Grey lines show individual participants. **B.** Cross-temporal location decoding for cued (top) and uncued (bottom) items. Time-axes are relative to WM array onset. White dotted lines show onset of WM array (0 s) and cue (1.75 s). White contours show clusters of significant decoding. Red boxes show time-points for computation of cross-generalization between WM array and cue epoch. **C.** Trial-wise regression with time-resolved cued-item location decoding and absolute error for congruent (blue) and incongruent (red) colour-probe trials (left) and location-probe trials (right). Red/blue bars show non-zero beta-weights and black bars show significant differences. Boxplots show beta weights from regression with mean decoding across post-cue TOI (0-1s) and absolute error. **D.** Location decoding for high (yellow) vs low (red) error trials for congruent, colour-probe trials (left) and congruent, location-probe trials (right). Red/yellow bars show significant decoding. Boxplots show mean location decoding across post-cue TOI (0-1s). **p*<.05, ***p*<.001.

Following the onset of the incidental cue (TOI: 0-1 s post cue; Figure 6A, right), the memorised location of the cued item was decodable from EEG voltages as predicted (t_{29}= 5.29, p<.001, d=.965). The locations of the two uncued items were also decodable following cue onset, although only one reached significance (uncued 1: t_{29}=2.94, p=.003, d=.536, uncued 2: t_{29}=1.12, p=.120, d=.220). Consistent with the hypothesis, decoding strength for the memorised location of the cued item was significantly higher for than for the uncued items as revealed by a one-way ANOVA ($F_{2,29}$=4.85, p=.011) and planned pairwise comparisons between the cued and each of the uncued items (cued vs uncued 1: t_{29}=2.59, p=.007, d=.473; cued vs uncued 2: t_{29}=2.61, p=.007, d= .476). There was no difference in decoding between the two uncued items (p=.327). Cluster-corrected t-tests across the incidental cue epoch (0-2s) revealed clusters of significant decoding for the cued item from 0.09 to 1.19 s (p<.001) and for the uncued items (averaged) from 0.18 to 0.63 s (p=.013). There was a significant difference between cued and uncued item decoding from 0.29 to 0.79 s (p=.010).

Cross-temporal decoding of location from EEG

To compare the format of the neural code for location between the encoding and cue epochs, I tested the extent to which the location representation as decoded from EEG cross-generalized across time (Figure 6B). Decoding along the diagonal suggests dynamic coding, while off-diagonal decoding is indicative of a more stable coding scheme across time in WM (King & Dehaene, 2014). Initial decoding during WM encoding (0-0.25ms) cross-generalized to later time-points during delay. However, strong decoding along the diagonal was accompanied by a brief period of no off-diagonal decoding, suggesting the representation quickly re-configured after WM array offset into a new coding scheme that remained relatively stable during the first delay. Cluster-corrected t-tests across the full trial (0-3.75s from WM array onset) revealed significant clusters for both cued (p=.005) and uncued items (p=.003) following encoding and into the first delay (white

contours). Following cue onset (1.75 s from WM array onset), there was a re-emergence of strong location decoding for the cued item along the diagonal, that did not cross-generalize to other time points, as evident from the lower decoding values on the off-diagonal when on-diagonal decoding was strongest (~2-2.5 s). This indicates that the cued item's location was represented in a different code compared to at other time-points during the trial. Because cluster-correction was performed across the entire trial for this analysis, decoding of the cued item's location following cue onset did not reach significance in this analysis (p=.052).

Nevertheless, by visual inspection, cross-temporal decoding between the initial delay and the post-cue delay (Figure 6B; red box) was not at zero, even if it momentarily dissipated when within-temporal decoding is at its strongest. To quantify cross-temporal decoding strength, I took the mean decoding value when training the decoder on the WM array epoch (0-1 s from array onset) and tested the decoder on the cue epoch (1.75-2.75 s from array onset; Figure 6B red box). Cross-temporal decoding strength was significant for the cued item (t_{29}=2.02, p=.026, d=.369). For the uncued items, although within-temporal decoding was significant following cue onset as reported above (Figure 6A), cross-temporal decoding was not significant (p=.271). However, as decoding magnitude was worse overall for the uncued items, lower signal-to-noise may have reduced power to detect a significant effect.

Together, the results suggest the neural pattern coding for location following cue onset is distinct from the one coding for location during encoding and the initial delay. Thus, selecting the cued colour from WM strengthens the neural representation of the item's location, but in a transformed format. However it is complemented by a more stable coding scheme for location that appears to persist until probe onset and is specific for the cued item.

Decoding of cued item's location from EEG predicts error on congruent colour-probe trials

If EEG decoding quality of the cued item's location signals the extent to which the cued item is brought into a privileged state in memory, post-cue decoding quality should predict faster and more precise performance on congruent trials ($\beta < 0$). If there is an associated cost for uncued memory items, post-cue location decoding quality may predict slower RTs and greater error on incongruent trials ($\beta > 0$). The results of the regression analysis between EEG location decoding and absolute error are presented below, separately for congruency condition (congruent vs incongruent) and probe type (colour vs location). There was no relationship with RT, except for an unexpected negative relationship between location decoding and RT on incongruent colour probe trials (Appendix A; Supplementary Figure S4).

In support of the hypothesis, decoding quality for the cued item's location predicted lower errors on congruent colour-probe trials (Figure 6C; t_{29}=-1.71, p=.049, d=-.312), but there was no relationship between cued item decoding and error on incongruent colour-probe trials (p=.376). Cluster-corrected t-tests across the cue epoch (0-2 s post cue) revealed a significant cluster of negative β weights for congruent colour-probe trials from 0.95 to 1.70 s (p=.005), indicating that location decoding quality was predictive of error mostly later in the delay period. There were no significant clusters for incongruent colour-probe trials, but βs for congruent colour-probe trials were significantly lower than for incongruent colour-probe trials from 0.95 to 1.31 s (p=.035) and again from 1.37 to 1.66 s (p=.044). Surprisingly, on location-probe trials there was no significant relationship between location-decoding quality and error for either congruent (p=.465) or incongruent (p=.679) trials. Cluster-corrected t-tests across the cue epoch (0-2 s post cue) also showed no significant βs for location-probe trials, congruent or incongruent. A 2×2 repeated measures ANOVA showed no main effect of congruency (p=.248), no main effect of probe type (p=. 289) and no congruency by probe type interaction (p=.144).

Comparisons between high and low error congruent trials broadly confirmed the regression analyses (Figure 6D) with greater decoding quality on low compared to high error congruent trials when orientation recall was probed by colour, but not when probed by location. Mean decoding quality across the initial post-cue delay (0-1 s) was analysed in a 2×2×2 repeated measures ANOVA with congruency, probe-type, and error as factors. There were no significant main effects and no interactions. On congruent colour-probe trials, decoding quality for the cued item's location was significant for low error trials (t_{29}=4.45, p<.001, d=.813) and not for high error trials (p=.235), but these were not significantly different (p=.094). Cluster-corrected t-tests across the cue epoch (0-2s post cue) revealed clusters of significant decoding when behavioural error was low (0.36-0.48 s, p=.001; 0.49-0.64 s, p<.001, 1.07-1.12 s, p=.026; 1.19-1.26 s, p=.019; 1.71-1.78 s, p=.014), but not when behavioural error was high and location decoding quality was significantly better for low error trials than high error trials from 1.60 to 1.64 s (p=.044). By contrast, when orientation recall was probed by location, location decoding quality was significant for both low (t_{29}=2.02, p=.026, d=.370) and high error trials (t_{29}=2.47, p=.01, d=.451) and there was no significant difference between them (p=.986). There was a cluster of significant location decoding for high error trials (0.51-0.59 s, p=.006) but not for low error trials and there were no clusters of significant difference in location decoding between high and low error trials.

In summary, post-cue decoding quality for the cued item's location predicted precision of orientation recall on congruent trials, when recall was probed by colour.

Decoding memorised locations from eye-position

Eye-position has recently been shown to track spatial attention in WM (van Ede et al., 2018). Consequently, I expected the cued item's location to be decodable from participants' eye-position following cue onset. First, I confirmed that any decoding from eye-position was not driven by large shifts in eye-movements toward the positions where

the WM items were presented. Figure 7A shows the distribution of saccade amplitudes. The vast majority of saccades were within 20% of the distance of the WM stimuli from fixation (~1.4°) indicating these were mostly within the micro-saccade range (<1°) and constituted small deviations from fixation. Figure 7B shows saccade frequency across the trial. Saccade frequency increased following the onset of the WM array (t=0) and again following cue onset (t=1.75), suggesting these were related to trial events.

Figure 7C shows two-dimensional histograms of gaze position from 0.4-1 s following cue onset, separately for when cued item had been presented at each of the six locations. For each location, gaze density was higher in the direction of the memorised location of the cued item. While small saccades were directed towards the cued item's location, practically none went the full way to the position where it was presented. To quantify gaze bias, the location decoding procedure was applied to the eye-position data (Figure 7D). The locations of each memory item was decodable from eye-position following the onset of the WM array (TOI: 0-1s; cued: t_{29}= 4.63, p<.001, d= .845; uncued 1: t_{29}= 5.07, p<.001, d=.926; uncued 2: t_{29} = 6.12, p<.001, d= 1.117) and there was no significant difference in decoding strength between them at this stage as expected (one-way ANOVA, p=.649). Cluster-corrected t-tests over the WM array epoch (0-1.5 s) revealed significant decoding quality for the cued item's location from 0.3 to 1.5 s (p<.001) and for the locations of the uncued items (averaged) from 0.25 to 1.5 s (p<.001).

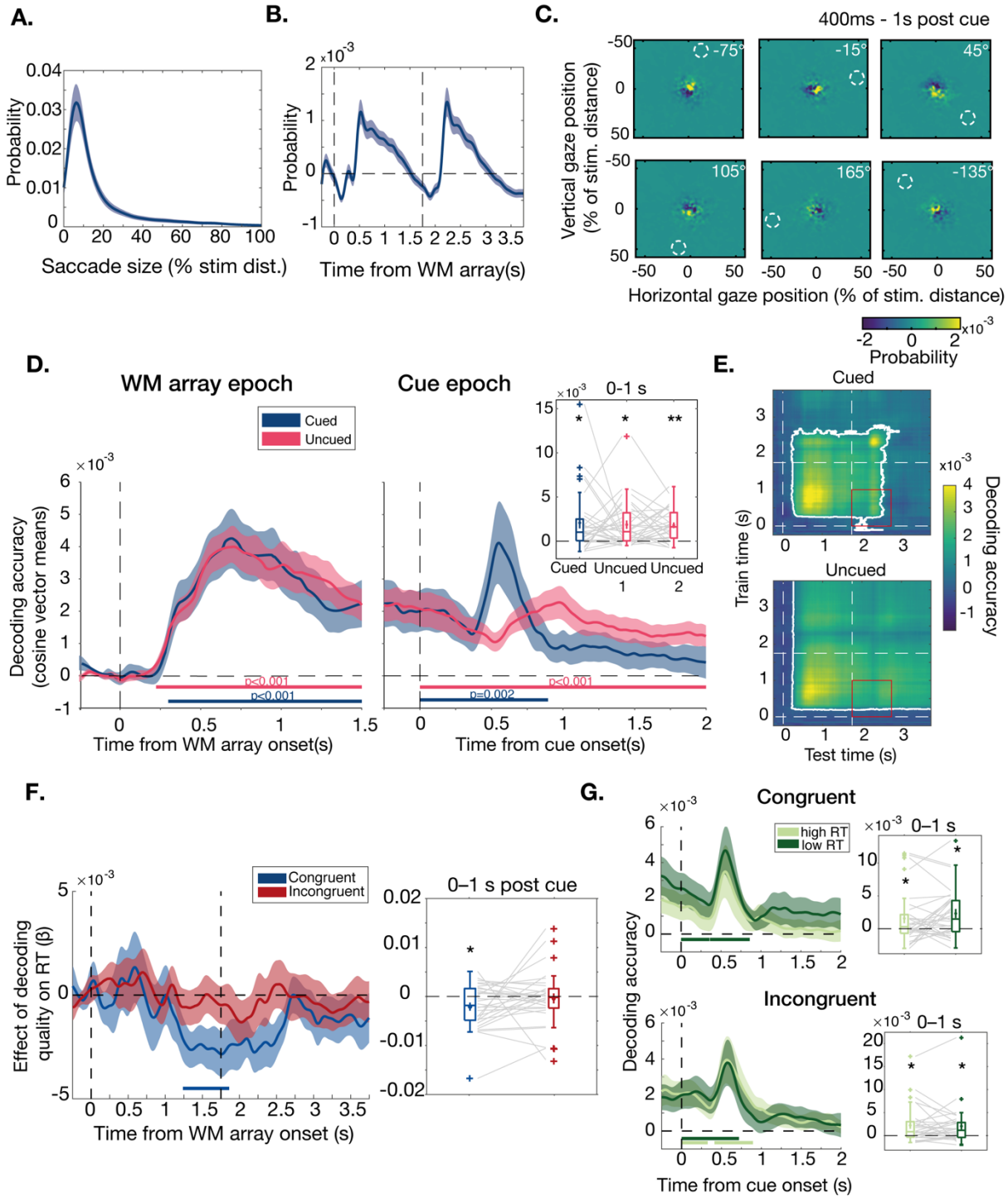

Figure 7. Gaze bias. **A.** Micro-saccade amplitude (% of stimulus distance, 100% = 7°). Shaded area shows standard error of mean (SEM). **B.** Baseline-corrected probability of microsaccades over time from WM array onset (0 s). **C.** Two-dimensional histograms of gaze position following cue onset (0.4-1s) for each cued-item location. White circles show direction of memorized location of cued item (for illustration purposes only). **D.** Location decoding of items in memory following WM array onset (left) and cue onset (right) for cued (blue) and uncued (pink; averaged in time-resolved plots). Pink/blue bars show significant location decoding. Boxplot shows mean location decoding in post-cue TOI (0-1s). Mean and SEM overlaid. Grey lines show individual participants. **E.** Cross-temporal location decoding for cued (top) and uncued (bottom) items. Time axes relative to WM array onset. White dashed lines show onset of WM array (0 s) and cue (1.75 s). White contours show significant decoding. Red boxes show time-points for computation of cross-generalization between WM array and cue epoch and cue epoch decoding along the diagonal. **C.** Beta weights from trial-wise regression with time-resolved cued-item location decoding and RT for congruent (blue) and

incongruent (red) trials. Red/blue bars show non-zero beta-weights. Boxplot shows beta weights from regression with mean decoding across post-cue TOI (0-1s) and RT. **D.** Location decoding for high (light green) vs low (dark green) RT trials for congruent (top) and incongruent (bottom) trials. Dark/light green bars show significant decoding. Boxplots show mean location decoding across post-cue TOI (0-1s). *p<.05, **p<.001.

Following the onset of the cue (TOI: 0-1s from cue onset), the memorised location of all three items in memory remained decodable from eye-position (cued: t_{29}= 3.21, p=.002, d=.587; uncued 1: t_{29}= 3.93, p<.001, d= .717; uncued 2: t_{29}= 5.44, p<.001, d=.994) and there was no significant difference in decoding strength between items (one-way ANOVA, p=.902). However, inspecting the time-course of eye-position decoding (Figure 7D) revealed a peak in decoding magnitude for the cued item's memorized location following cue onset, which was not present for the uncued items. Cluster-corrected t-tests across the incidental cue epoch (0-2s post cue) showed significant clusters of decoding for the cued item from 0 to 0.91 s (p<.001), after which decoding was no longer significant, while the uncued items (averaged) remained decodable across the full epoch (p<.001). Thus, eye-position tracked all three items in memory and although there was a peak in decoding quality following cue-onset specifically for the cued item's location, this did not reach significance.

Cross-temporal decoding of location from eye-position

Next, I tested the extent to which the location representation as decoded from eye-position cross-generalized between WM array and cue epochs. Recall that for location decoding from patterns of EEG activity, location was represented in distinct patterns of activity between the two epochs, complemented by a more stable code (Figure 6B). Cross-temporal decoding results for eye-position are presented in Figure 7E. Cluster-corrected t-tests across the full trial (0-3.75s from WM array onset) revealed significant clusters of location decoding for both cued (p=.011) and uncued items (p<.001) that cross-generalized from WM array offset (0.25 s) to the rest of the pre-cue delay. Following cue

onset, there was an increase in location decoding for the cued item that cross-generalized to previous time points, after which decoding was no longer significant. Decoding of the uncued item positions was stable across both the pre-cue and post-cue delays, unlike the corresponding EEG decoding.

To quantify cross-temporal decoding strength, I took the mean decoding value when training the decoder on the WM array epoch (0-1 s from array onset) and testing the decoder on the cue epoch (1.75-2.75 s from array onset; Figure 7E red boxes). Cross-temporal decoding strength was significant for both the cued (t_{29}=2.98, p=.003, d= .544) and uncued items (t_{29}=5.87, p<.001, d= 1.07). Thus, eye-position tracked locations in memory in a stable manner across the entire trial, though decoding for the cued item's position dissipates ~1 s after cue onset when the incidental task is completed. Note that a stable coding scheme for eye-position decoding is expected given only two coordinates contribute to the code.

Eye-position decoding of cued item's location predicts RTs on congruent trials

To the extent eye-position tracks spatial attention in WM and spatial attention is involved in privileging the cued item in WM, decoding quality for the cued item's location may predict faster and more precise performance on congruent trials ($\beta < 0$). If spatial attention toward the cued item's position comes at a cost to uncued memory items, decoding quality for the cued item's location may predict slower RTs and greater error on incongruent trials ($\beta > 0$). There was no relationship between eye-position location decoding and absolute error (Appendix A; Supplementary Figure S5). The results of the regression analysis between eye-position location decoding and RT are presented below, separately for congruency condition (congruent vs incongruent) and probe type (colour vs location).

As expected, on congruent trials there was a negative relationship between gaze bias toward the cued item's position following cue onset (0-1 s) and RTs, that was significant when recall was probed by location (t_{29}=-1.78, p=.043, d=-.325), but not when recall was probed by colour (t_{29}=-1.47, p=.077, d=-.268). When boosting power by averaging over probe-types, this relationship was more robust (Figure 7F; t_{29}=-2.484, p=.010, d=-.453). There was no relationship between cued item decoding and RT on incongruent trials for either probe type (colour-probe: p=.657; location-probe: p=.442; mean: p=.648). A 2×2 repeated measures ANOVA with factors congruency (congruent vs incongruent) and probe-type (colour vs location) showed a non-significant trend for a main effect of congruency ($F_{1,29}$=3.80, p=.061), suggesting β weights were lower on congruent than on incongruent trials. There was no main effect of probe type (p=.663) and no congruency by probe type interaction (p=.933). Cluster-corrected t-tests across the incidental cue epoch (0-2 s post cue) revealed no significant clusters for either condition. However, an exploratory analysis across the entire trial period (0-3.75 s from WM array onset) revealed a significant cluster of negative β weights before cue onset for congruent trials (averaged over probe type) from 1.24 to 1.86 s from WM array onset (p=.034). This suggests that the extent to which participants' gaze was already directed toward the *to-be-cued* item's location prior to cue onset predicted the speed at which orientation recall was initiated on congruent trials.

Comparisons between high and low RT trials broadly confirmed the regression analyses (Figure 7G) as decoding quality was better on low compared to high RT congruent trials, though this effect was not significant. A 2×2×2 repeated measures ANOVA with congruency, probe-type, and RT as factors revealed no significant main effects and no significant interactions, although there was a trend for an interaction between congruency and RT condition ($F_{1,29}$= 3.30, p=.080). On congruent trials, the cued item's location was decodable both when RTs were fast (t_{29}=3.38, p=.001, d= .618) and slow

(t_{29}=2.14, p=.020, d=.391) and there was no difference between them (p=.198). Cluster-corrected t-tests across the cue epoch (0-2 s post-cue) revealed significant decoding for congruent, fast RT trials from 0 to 0.36 s (p=.030) and again from 0.36 to 0.86 s (p=.010) and no significant decoding for slow RT trials. On incongruent trials the cued item's location was decodable both when RTs were fast (t_{29}= 2.62, p=.007, d=.478) and slow (t_{29}= 2.99, p=.003, d=.546) and there was no difference between them (p=.725). There were clusters of significant location decoding for fast RT trials from 0 to 0.72 s (p=.003) and for slow RT trials from 0 to 0.33 s (p=.032) and again from 0.41 to 0.9 s (p=.013).

In summary, the extent to which participants' gaze was directed toward the cued item's location predicted response initiation on congruent trials, but this relationship was not specific to the post-cue period as expected if gaze-position was specifically involved in privileging the cued item in WM.

Memorised orientation of cued item could not be decoded from EEG voltages

If selecting the cued colour from WM strengthens the neural representation of the item's orientation, EEG orientation decoding quality for the cued item, relative to uncued items, should improve following the cue. Figure 8 shows orientation decoding from EEG voltages. Following WM array onset (TOI: 0-1s), item-orientations could, to some extent, be decoded from posterior EEG channels (Figure 8A; cued: p=.206, uncued1: t_{29}=2.85, p=.004, d=.521, uncued2: t_{29}=2.23, p=.017, d=.408). There was no significant difference in decoding strength between items (one-way ANOVA: p=.307). Cluster-corrected t-tests over the WM array epoch revealed clusters of significant decoding for the cued item from 0.07 to 0.37 s (p=.021) and for the uncued items (averaged) from 0.06 to 0.58 s (p=.002). Following cue onset, orientation decoding was not significant for any of the three items in memory (Figure 8B; cued: p=.352, uncued1: p=.454, uncued2: p=.261) and there was no significant difference in decoding strength between them (one-way ANOVA: p=.946). Cluster-corrected t-tests over the cue epoch (0-2 s post-cue) also did not show any

significant decoding for any of the orientations in memory and showed no difference between cued and uncued items.

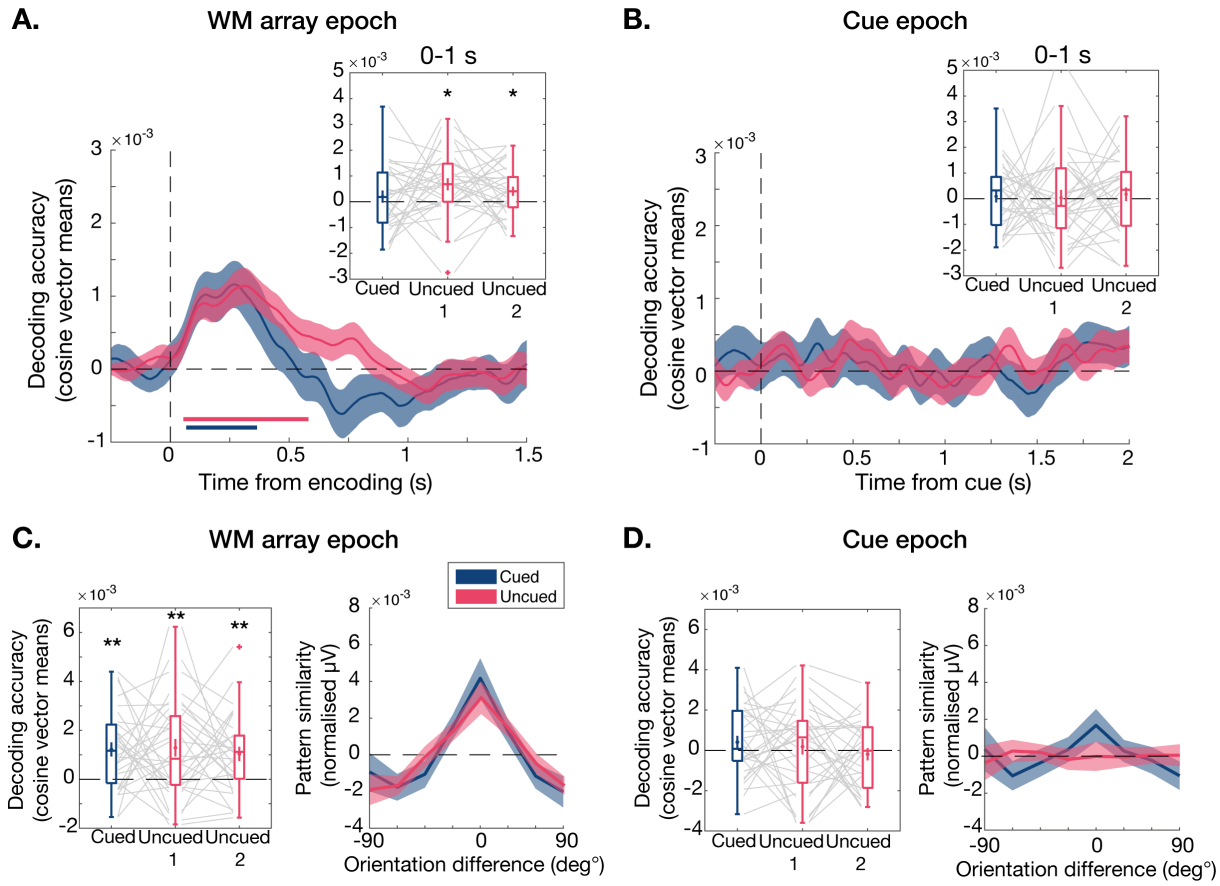

Figure 8. Orientation decoding from EEG. **A.** Orientation decoding of items in memory following WM array onset for to-be-cued (blue) and uncued (pink; averaged in time-resolved plots) memory items. Shaded area shows standard error of mean (SEM). Blue/pink bars show significant decoding. Boxplot shows mean decoding from 0-1 s following WM array. Mean and SEM overlaid. Grey lines show individual participants. **B.** Same as A but following cue onset. **C.** Left: boxplot of spatio-temporal decoding from 100-500ms following WM array onset for each item in memory. Right: orientation tuning from 100-500 ms following WM onset for items in memory. **D.** Same as C, but following cue onset. *p<.05, **p<.001.

While orientation angles in memory could, to some extent, be decoded immediately following encoding, decoding strength was lower than previous studies using similar decoding procedures (Muhle-Karbe et al., 2020; Wolff et al., 2017). One reason may be that the present study included six possible locations in memory, while previous studies have tended to only have one or two. This may have limited power to detect orientation representations in WM. Given the weak decoding results for orientation using the standard approach, I added a complementary spatiotemporal decoding approach, as an

exploratory analysis, that leverages information encoded in temporal as well as spatial patterns (Grootswagers et al., 2016; Wolff, Jochim, et al., 2020; Wolff, Kandemir, et al., 2020b).

Using the spatiotemporal decoding procedure (Figure 8C), all three items were decodable from posterior EEG channels following WM array onset (cued: t_{29}=4.14, p<.001, d=.756, uncued1: t_{29}=3.53, p<.001, d= .645, uncued2: t_{29}=3.47, p<.001, d=.634). There was no difference in decoding strength between items (one-way ANOVA: p=.843). Following cue onset, none of the three item orientations could be decoded (cued: p= 0.108, uncued1: p=.322, uncued2: p=.728), although the tuning function for the cued item does show a small peak at zero (Figure 8D). There was no significant difference between the three items (one-way ANOVA: p=.433).

Discussion (Experiment 2)

Experiment 2 replicated the behavioural results from Experiment 1 showing that orientation judgements were faster for congruent relative to incongruent trials. Contralateral alpha suppression indicated that a non-predictive, centrally presented, colour cue triggered spatial attention toward the cued object's original position. Furthermore, there was a relative boost in the representation of the cued object's memorised location as decoded from EEG voltages. The strength of the location representation following the cue predicted behavioural error when the same item was subsequently probed for orientation recall.

Predictive retro-cues have previously been shown to elicit contralateral alpha suppression and gaze bias relative to the position where a cued item was originally presented (Myers, Walther, et al., 2015; van Ede, Chekroud, & Nobre, 2019; Wallis et al.,

2015; Wolff et al., 2017). In this experiment, a non-predictive colour cue similarly triggered transient contralateral alpha suppression, as expected if spatial attention was directed toward the cued item's position. Although there was a peak in gaze bias toward the cued item following the cue, this did not reach significance. Importantly, the incidental delay task did not require retrieval of the item's location, suggesting spatial attention may play a role in accessing non-spatial features, such as colour, in WM. This is consistent with previous reports of engagement of spatial attention when retrieval from WM is cued by colour, even when location is irrelevant to the task (Poch et al., 2017; van Ede, Chekroud, & Nobre, 2019).

Initial post-cue contralateral alpha suppression was associated with lower error on congruent trials, suggesting contralateral alpha suppression may signal the extent to which the cued item is successfully selected and brought into a privileged memory state. Following initial contralateral alpha suppression, there was a reversal in lateralization one second after cue onset. The non-predictive nature of the cue may encourage re-orienting of spatial attention once the incidental task-component is completed–either to fixation in preparation for the probe (Jin, Nobre, & van Ede, 2020) or toward the opposite hemifield to refresh other items in memory. Supporting this interpretation, the strength of the reversal was associated with behavioural error on incongruent trials when one of the uncued items was probed for recall. However, note that the reversal of alpha lateralization did not replicate in Chapter 3.

While gaze bias and alpha lateralization are both thought to reflect orienting of spatial attention, they seem to play distinct roles in WM. Gaze bias, in contrast to alpha lateralization, was *not* associated with behavioural error. Instead, location decoding from eye-position predicted how long participants took to initiate a response on congruent trials. Van Ede (2019) showed that the degree of post retro-cue gaze bias toward the cued item predicted RTs. In the current study, the relationship between eye-position decoding

and RT was not specific to the post-cue time window. Rather, RTs were faster on congruent trials when eye-position happened to be biased toward the to-be-cued item even prior to cue onset. Consistent with a proposed tight coupling between spatial attention, eye-movements and motor planning (A. Heuer et al., 2020; Olivers & Roelfsema, 2020; Rizzolatti, Riggio, Dascola, & Umiltá, 1987; van Ede, Chekroud, & Nobre, 2019), gaze bias may indicate the readiness to respond to certain content in memory (i.e. "accessibility") rather than the strength of the WM representation itself. However, while the relationship between eye-position and RT looked similar in Chapter 3, it was not significant, and instead gaze bias toward the cued item predicted worse performance on incongruent trials. Possible sources of these discrepancies are discussed further in Chapter 3.

Internal attention toward retro-cued content in WM has been shown to boost neural decodability of relevant WM content from fMRI and EEG, particularly in sensory cortex (Christophel et al., 2018; LaRocque et al., 2013; Lepsien & Nobre, 2007; Lewis-Peacock et al., 2012; Sprague et al., 2016; Wolff et al., 2017). Here, selecting a cued colour in memory for an orthogonal change-detection task strengthened decoding quality of the cued item's location, relative to uncued items. Previous studies have focused on the representational state of content cued either as relevant or irrelevant for subsequent decision-making. The findings reported here are novel in showing that the neural representation for location, which was irrelevant to the incidental task, was nevertheless strengthened when attention was directed to the cued colour. To the extent that the neural activity state reflects the focus of attention in WM (Larocque et al., 2014), selecting the cued colour in WM may activate the location context of the cued item and bring it into an attended state in memory.

The strength of the neural location representation was functionally relevant to WM precision, as EEG decoding of the cued item's location predicted orientation judgements

on congruent trials, when recall was probed by colour. Surprisingly, there was no evidence of a relationship when recall was probed by location. While this null effect should be interpreted with caution, such asymmetry between probe features might be expected if colour and orientation are bound only through their shared location (Schneegans & Bays, 2017a). That is because the strength of the location representation should index the strength of the colour-location connection, which is a necessary route to retrieve orientation on colour-probe trials, but not on location-probe trials where the location context is provided and orientation can be accessed more directly. Consequently, the findings are consistent with a model of WM where location plays a special role in binding non-spatial features (Schneegans & Bays, 2017; Treisman & Zhang, 2006).

The neural pattern coding for item location following the cue was dynamic as indicated by strong decoding along the diagonal of the cross-temporal decoding matrix (King & Dehaene, 2014; Stokes, 2015). This coding scheme was qualitatively different from the neural pattern coding for the same item's location during the initial delay. For the cued item, dynamic coding was complemented by a more stable coding scheme that was present in both delays, but reduced when decoding strength was greatest following the cue. Previous research has shown dynamic coding during encoding into WM that settles into a distinct, stable coding scheme during the delay period (Spaak et al., 2017; Stokes, 2015; Stokes et al., 2013; Wolff et al., 2017). Stable coding during the delay may transition into a distinct coding scheme following presentation of distractors (Cavanagh, Towers, Wallis, Hunt, & Kennerley, 2018; Parthasarathy et al., 2017) or a strong visual impulse designed to reveal latent states in memory (Wolff et al., 2015). Distinct coding schemes might be expected if encoding into WM changes the functional connectivity of the network, such that patterns of activity depends on the interaction between the memory trace and new sensory input (Stokes, 2015). In this experiment, strengthened decodability following the cue did not purely reflect a *re-activation* of the coding scheme

established after encoding, as would be expected if the WM representation was supported exclusively by persistent activity in selective neural populations. Rather location was temporarily represented in a dynamic neural code, perhaps reflecting a cascade of processing steps associated with cued retrieval from memory.

I found no evidence for a boost in the representation of the cued item's orientation following the incidental cue, but low power may limit interpretability of this null effect. While trial numbers and participant numbers were similar to other studies successfully decoding orientation from WM using EEG (Wolff et al., 2015, 2017), previous studies have shown that orientation decoding does not cross-generalize well across encoding locations (Bae & Luck, 2017; Wolff, Jochim, et al., 2020). In this study, there were six possible locations, reducing trial numbers and thus power within each location. A high-powered study with fewer locations should be able to test whether attention spreads to other object-features in WM or if location is a special feature in this regard. This will be the focus of Chapter 3.

General conclusion

In this study, I tested the prediction that attention spreads between object-features in working memory. In both experiments, orientation judgements of items in WM were faster and more precise when another feature of the same item was selected for an orthogonal task during the delay (congruent), compared to when a feature from a different item was selected (incongruent). This suggests that directing attention to one item-feature in WM privileges other features belonging to the same object. I further showed that retrieval of spatial context is obligatory when attending a non-spatial feature in WM. The spatial component may serve to strengthen the WM representation, boosting WM performance.

Obligatory retrieval of the location context of the cued colour is equally predicted by flexible conjunctive coding models, with short-term synaptic plasticity, where object-features are retrieved through auto-associative pattern completion (Manohar et al., 2019) as well as activity-based fixed conjunctive coding models where non-spatial features are stored conjunctively with location (Schneegans & Bays, 2017a, 2017b). The former model, but not the latter, further predicts a boost in the orientation representation following colour-selection, which was not supported by the current study. However, low power limited the interpretability of this null finding and Chapter 3 will aim to test whether colour-selection strengthens orientation representation in WM.

Chapter 3. Colour-selection in working memory engages spatial attention and strengthens neural representation of associated features

In Chapter 2, I showed that selecting a cued colour or location from WM behaviourally privileges other features belonging to the same object. Alpha lateralization indicated that colour-selection engages spatial attention towards the item's previous location and decoding from EEG showed that the neural representation of the item's location was strengthened following the non-predictive colour cue. The strength of the location representation predicted WM precision when the same item was subsequently probed for orientation recall, but only when probed by colour. Together, the results indicated that spatial attention is involved in selecting non-spatial features in WM and that location plays a special role in binding features in WM.

There were several questions that could not be answered with the experimental design in Chapter 2. First, although the cued item's location was redundant at the time of the cue in the EEG and eye-tracking experiment (Experiment 2), participants did need to encode locations as orientation recall was probed either by colour or location. In the current experiment, I aimed to test whether spatial attention is involved in selecting a non-spatial feature from WM when location is a fully redundant feature. Second, the power to detect neural representations of orientation with EEG was low in Chapter 2 (Experiment 2), perhaps because WM items were presented at six possible locations. Previous studies using a similar analysis pipeline have shown that orientation decoding from EEG is most successful when training and testing the classifier within the same location (Bae & Luck, 2017; Wolff, Jochim, et al., 2020; Wolff et al., 2017). In this study I reduced the number of locations at which WM items were presented to two and increased trial numbers from 1080 to 1400 to improve power to detect potential orientation representations during the WM delay. The hypotheses and planned analyses were pre-registered on the open science framework after data collection had begun, but

before I started to analyse the data (https://osf.io/pgdfj). The pre-processing pipeline was tested on a pilot participant not included in the analyses. Below I include the pre-registered hypotheses and methods. I note any deviations from the pre-registered methods in the footnotes.

Pre-registered hypotheses and methods

WM flexibility allows humans to select and prioritise information depending on its current relevance (Bundesen, 1990) and to hold novel items in mind by combining arbitrary features into coherent representations ("feature binding") (Wheeler & Treisman, 2002). In this study, I will use behavioural measures, electroencephalography (EEG) and eye-tracking to investigate whether selecting one feature of a multi-feature item in WM automatically privileges associated item features, as predicted by recent computational models of WM flexibility (Manohar et al., 2019; Oberauer & Lin, 2017).

According to recent models of WM flexibility, features in WM are retrieved through pattern completion, perhaps via strengthened connectivity between features belonging to the same objects (Manohar et al., 2019; Oberauer & Lin, 2017). If so, a retrieval cue that directs attention to one item-feature in memory should automatically privilege all features of the same object, even if these are not currently relevant. To test this prediction, I presented participants with two oriented, coloured bars on each trial. After a delay, a probe prompts them to report the orientation angle of one of the remembered items. During the delay, participants are additionally presented with a non-predictive retro-cue (i.e. 'incidental' cueing) (Zokaei, Manohar, et al., 2014; Zokaei, Ning, et al., 2014) and asked to make a speeded response if the cued colour did *not* match the colour of either of the remembered items. On congruent trials, the item that was incidentally

cued was also probed during memory recall; on incongruent trials, the other item was probed.

Previous studies have shown that selecting information in WM is associated with faster and more accurate behavioural performance (Griffin & Nobre, 2003; Pertzov, Bays, et al., 2013; Souza & Oberauer, 2016) and improved neural decoding of relevant features from fMRI and EEG (LaRocque et al., 2013; Lepsien & Nobre, 2007; Lewis-Peacock et al., 2012; Sprague et al., 2016). Here I will test if this is similarly true for object-features that are not currently relevant. First, if selecting the cued colour for the incidental task automatically privileges associated, but currently redundant item-features, orientation judgements should be faster and more precise on congruent than on incongruent trials. This would replicate previous behavioural findings (Zokaei, Manohar, et al., 2014). Alternatively, it may be possible to selectively attend the cued colour without associated benefits to orientation judgements for the same item. If so, there should be no difference in performance between congruent and incongruent trials.

Second, using alpha lateralization and gaze bias as indices of spatial attention in WM (van Ede, Chekroud, & Nobre, 2019; Wallis et al., 2015), I will test if spatial attention is directed toward the original position of the cued item following the non-predictive colour cue, despite location being a fully redundant feature in this task. If so, I expect contralateral minus ipsilateral alpha power to be reduced relative to the original location of the cued item. I also expect gaze position to be biased toward the cued item's original position. Alternatively, there may be no significant alpha lateralization or gaze bias relative to the cued item's original location, which would suggest a mechanism for colour retrieval in WM without involvement of spatial attention.

Finally, if selecting the incidentally cued colour by extension strengthens the representation of the orientation of the cued item, neural decodability should improve for

the cued item's orientation relative to the orientation of the uncued item following the cue, even though the item's orientation is irrelevant to the 'incidental' change-detection judgement. Alternatively, it may be possible to selectively access a single feature in memory, similar to selective coding for only relevant item features observed during encoding (Bocincova & Johnson, 2019; Serences et al., 2009). I expect the orientation of the cued item to be decodable from EEG activity patterns following incidental cue onset and decoding accuracy to be greater for the cued than the uncued item. Alternatively, patterns of activity in the EEG channels may contain no information about the cued item and/or may contain equal information about the orientations of both items following the incidental cue.

Methods

Participants

In the pre-registration (https://osf.io/pgdfj), I planned to recruit a total of 35 participants. However, due to Covid-19 I only managed to test a total of 32 participants. One participant was excluded from analyses due to colour-blindness. A further five participants were excluded as they had less than 80% of usable epochs remaining after artefact rejection, in line with the pre-registered exclusion criteria, leaving a total of 26 participants for analyses (14 female, 12 male). Participants were aged between 18 and 34 (M=23.23, SD=5.02). Twenty-two were right handed and four were left handed. All participants had normal or corrected-to-normal vision (including normal colour vision), and did not have any history of neurological or neuropsychiatric disorders. Participants provided informed consent prior to participating in the study. They were reimbursed for their time at a rate of £15/hour or with course credit. The study has been approved by the Oxford Central University Research Ethics Committee (R55073/RE008).

Procedure

Participants performed a computerised experimental task (Figure 1). It consisted of a continuous report WM task, in which participants reported the orientation angle of one of two items in memory at the end of a delay. During the delay, participants performed a secondary "incidental task", where a non-predictive colour cue, presented centrally, required a speeded mismatch response: does the colour differ from the colours of the two memory items? At the end of the trial, a probe indicated which of the two items participants should report. This was either the same item as the one that was cued on the incidental task (congruent condition, 50% of trials) or the other item in memory (incongruent condition, 50% of trials).

Following an initial fixation period of 300 ms, participants were presented with a WM display consisting of two oriented, coloured bars. The WM items were presented on screen for 250 ms followed by a delay period of 750 ms. For the incidental task, participants were presented with a colour cue for 200 ms. On mismatch trials (20%), participants pressed the space bar within 1000 ms of incidental cue onset, after which the rest of the trial was skipped. On match trials (80%), participants made no response and the trial continued (unless they incorrectly pressed the space bar, in which case the trial was also discontinued). Participants received feedback for correct responses or incorrect omissions. After another delay of 500 ms, the probe appeared on the screen. The colour of the probe indicated which of the two items participants should report. Participants used the arrow keys[2] to turn the on-screen dial to match the orientation of the probed item and pressed the space bar to respond. There was no maximum time to respond, but once participants initiated a response the trial would time-out 2500 ms

[2] The response dial in Chapter 2 was initially adopted to make orientation reports more intuitive for participants. However, it was associated with software issues and piloting suggested participants found the arrow keys easy to use.

after response initiation. They received visual feedback of the correct orientation for 100 ms. The inter-trial-interval was jittered between 400-800 ms. At the end of each block, they received feedback of their incidental colour change-detection rate and on their mean absolute error.

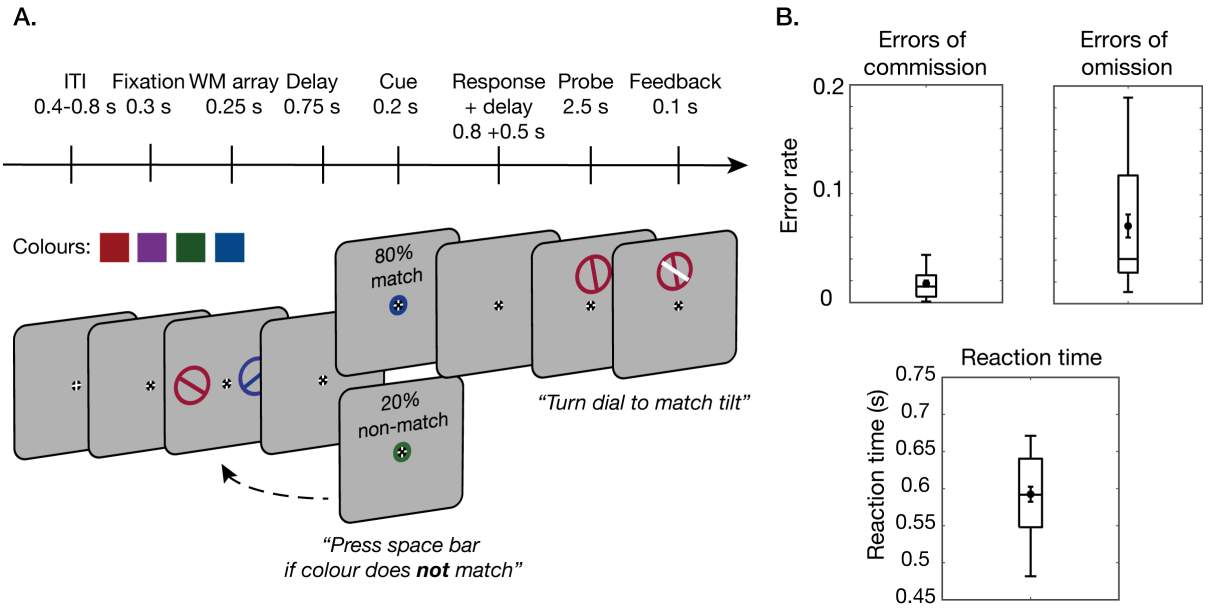

Figure 1. Task design and incidental task performance. **A.** Task design. Participants remembered two oriented, coloured bars on each trial, presented laterally. During the delay, they were presented with a non-predictive colour cue that asked them to make a speeded mismatch-response (i.e., does it differ from the colours in memory?; 'incidental task'). At the end of the trial, they reported the orientation of one of the items, cued by colour. This item may be the same as was also cued on the incidental task (congruent condition, 50%), or the other item (incongruent condition, 50%). **B.** Boxplots of incidental task performance. Top: errors of commission (i.e. key press when colour was a match; left) and errors of omission (i.e. no key press when colour was a mismatch; right). Bottom: reaction time from cue onset on correct mismatch trials. Mean and standard error of mean (SEM) overlaid.

Participants completed a total of 1400 trials. The 20% non-match incidental trials were excluded, leaving a maximum of 1120 trials per participant for analyses. Participants practiced the task beforehand until they reached a criterion of >80% correct incidental task change-detection rate and <25 degrees mean absolute error relative to the target orientation.

Apparatus and stimuli

The task was programmed and stimuli presented in Matlab with Psychtoolbox (Brainard, 1997a; The MathWorks Inc., 2017). The task was presented on a 24 inch, 1920x1080 pixel monitor running at 100 Hz.

Stimuli were presented on a grey background (RGB=128, 128, 128[3]). A fixation target was presented at the centre of the screen throughout the trial. It consisted of a filled black circle (0.5° diameter) with a white cross inside (width: 0.15°). To signal trial onset and to encourage fixation, a black dot (0.15° diameter) appeared inside the fixation target 300 ms before the onset of the memory array. The memory array consisted of two oriented, coloured bars inside circles (diameter: 6°, width: 0.3°). The colours were picked at random without replacement out of four possible colours on each trial: red (RGB=153, 0, 27), purple (RGB=122, 39, 150), green (RGB=16, 83, 0), and blue (RGB=0, 72, 136), chosen to be isoluminant in CieLab space. On every trial, the orientation of each stimulus was randomly selected without replacement from a uniform distribution of 160 orientation angles between 1.125 and 180 degrees. The two WM items were presented laterally 6° from the fixation target. For the incidental task, participants were presented with a colour cue appearing around the fixation target (diameter: 1°, width: 0.5°). The probe was a tilted bar inside a circle (6°; same size as memory items) and matched the colour of the probed memory item. The orientation angle of the probe was selected at random on each trial between 1 and 180 degrees. The probe was presented 6° above or below the fixation target, with equal probability, pseudorandomised for the full experimental session. This type of probe display has been reported to reduce the risk of stimulus-specific eye-movements (Mostert et al., 2018). Feedback of the correct target

[3] The pre-registration mistakenly said RGB = 0.5, 0.5, 0.5.

orientation was presented as a white line on top of their reported orientation for comparison.

EEG data acquisition

Electroencephalography (EEG) was recorded at a sampling rate of 1kHz using Synamps amplifiers (Neuroscan, Inc., Albany, CA) and Curry Neuroimaging Suite (v7.0.12 X) software. Sixty-one electrodes were positioned according to the 10-10 system ("American Electroencephalographic Society guidelines for standard electrode position nomenclature.," 1991). Eye movements were recorded from bipolar EOG derivations from electrodes above and below the right eye (vertical) and lateral to each eye (horizontal). A ground electrode was placed on the left elbow and the left mastoid was used as the active reference.

In addition, participants' eye movements were recorded using an eye tracking camera (EyeLink) tracking at least one eye at a rate of 1kHz. Eye-movements from six participants were accidentally recorded at a sampling rate of 500 Hz instead of 1000 Hz. These participants were not excluded as the eye-tracking data were down-sampled to 500 Hz for all participants, in line with the pre-registered pre-processing pipeline.

Analyses

Behaviour

The behavioural measures of interest were mean absolute error from the target angle and median response initiation times (henceforth RT). I performed paired samples t-tests (one-tailed) to test for a main effect of congruency for both measures. I predicted that mean absolute error and median RTs would be significantly lower for congruent relative to incongruent trials.

To estimate sources of errors, I also fit a swap model to the data (Bays et al., 2009). The model is an extension of the Zhang & Luck, 2008 model and can be described in the following way:

$$P(\theta') = (1 - \gamma - \beta)\phi_\sigma(\theta' - \theta) + \gamma\frac{1}{2\pi} + \beta\frac{1}{m}\sum_{i}^{m}\phi\,(\theta' - \theta_i^*)$$

Where θ denotes the target orientation in radians, θ' denotes the reported orientation, γ is the proportion of trials where the participant responds at random and ϕ_σ denotes the circular equivalent of the normal distribution (Von Mises distribution) with mean = 0 and standard deviation = σ. The swap model extension also includes the β parameter, which is the probability of responding with one of the non-target values: $\{\theta_1^*, \theta_2^*, \theta_m^*\}$ for m non-targets. As such, the model estimates the precision, target rate, non-target rate and guess rate of responses. As above, I performed paired samples t-tests (one-tailed; congruent>incongruent for target rate and precision, and incongruent>congruent for guess rate and non-target rate[4]) for each measure to test for a main effect of congruency.

Pre-processing of eye-movement data

The EyeLink EDF file was converted to ASCII format and read into Matlab using the Fieldtrip toolbox (Oostenveld et al., 2011). The data were epoched around the onset of the WM array (-1 to 4 seconds). I took the mean of the right and the left eye (or used a single eye, where it was only possible to record from one eye[5]) and down-sampled the data to 500 Hz. Eye-blinks were detected (zeros in the eye-position data) and interpolated from -100 to +100 ms around the blink onset and offset using linear interpolation. Saccades (including microsaccades) were detected using a Matlab function

[4] The pre-registration mistakenly said 'two-tailed', but I specifically predicted performance to be better for congruent than incongruent trials.

[5] Elaboration on pre-registration: Due to some participants wearing glasses, it was only possible to record from one eye. I did not exclude these participants.

(Engbert & Mergenthaler, 2006) based on two-dimensional velocity vectors computed from eye-position data using moving averages (velocity type=2). I used standard settings of velocity threshold (λ) of 6 standard deviations of the velocity distribution, and minimum duration of 6 ms.

EEG Pre-processing

Pre-processing of the EEG data was performed in Matlab using the Fieldtrip toolbox (Oostenveld et al., 2011). The ERPs were re-referenced offline to the mean of the right and left mastoid and down-sampled to 500 Hz. A high-pass filter of 0.1 Hz and a low-pass filter of 40 Hz was applied to all channels. The data were epoched around the onset of the WM array (-1 to 4 seconds). I performed ICA to remove any heartbeat artefacts and/or artefacts due to eye-blinks. Artefact detection and rejection was performed separately for the time-window following WM array onset (0-1 s) and following cue onset (0-1.5 s). In addition, for an exploratory analysis not included in the pre-registration, I also performed artefact rejection separately for a time-window following probe onset (0-1.5s)[6]. When an artefact was detected, I rejected the epoch containing the artefact from analyses, not the entire trial. I marked epochs with a high degree of variance using the Fieldtrip summary plot for visual artefact rejection and further marked epochs that exceed a cut-off threshold of 50 of the z-transformed value of the pre-processed data to detect any large artefacts due to muscle movements. I further excluded epochs with saccades >1° visual angle identified from the eye-tracking data. In case eye-tracking data were missing, I visually inspected EOGs to reject saccade epochs. This was generally only the case for the occasional trial, but two participants had missing eye-tracking data for a larger number of trials.

[6] Exploratory analysis not included in pre-registered methods.

I excluded participants with fewer than 80% of epochs remaining following artefact rejection from further analyses. On average, I rejected 7.22% of trials from the encoding epoch, 10.20% of trials from the cue epoch and 2.13% of trials from the probe epoch (exploratory analysis).

Time frequency decomposition and alpha lateralization

To test whether a centrally presented, non-predictive colour cue drew spatial attention to the original position of the cued item, I computed the alpha lateralization index relative to the cued item in the time window after incidental cue onset (0.2–0.8 s; to avoid carry over from probe onset). Spectral decomposition of the data was performed in Matlab with the FieldTrip toolbox (Oostenveld et al., 2011). I applied Hanning tapers with a time window width of five cycles per frequency, with frequencies of interest between 8 Hz and 14 Hz in steps of 1 Hz. I log-transformed the power at each frequency $10 \times log_{10}$ and calculated the mean power over contralateral and ipsilateral posterior channels (P8/7, P6/5, P4/3, PO8/7, PO4/3, O2/1) relative to the cued and uncued item. To get a lateralization index, I took the difference between contralateral and ipsilateral alpha power averaged over the specified time-window and tested for significant lateralization using a one-sample t-test (one-tailed, <0)[7]. In an exploratory analysis, I also tested for significant alpha lateralization at each time-point across the time-window from 0-1 (one-sample t-test, two-tailed) using a cluster-based permutation test (10,000 permutations) to correct for multiple comparisons over time. I used a cluster-forming and cluster significance threshold of p<.05.

[7] The pre-registration mistakenly said (one-tailed, >0) instead of (one-tailed, <0). I was specifically interested in contralateral alpha suppression, so I expected contra–ipsilateral alpha power to be negative.

Gaze bias analysis

A recent study has shown that human gaze may be biased toward attended spatial locations in WM (van Ede, Chekroud, & Nobre, 2019). Here, I similarly used gaze bias as an index of spatial attention towards the original position of the cued item. If spatial attention is drawn to the original position of the incidentally cued item, I expected horizontal gaze position to be biased toward the left (i.e. negative values) for left cue trials and toward the right (i.e. positive values) for right cue trials. First, I centred the data to the mean position during fixation (300–100 ms before stimulus onset). I then normalized the data according to the eccentricity of the memory items, so +/− 100% corresponds to the eyes being focused on the original position of the memory items (i.e. 6° visual angle) and 0% corresponds to fixation. I plotted the gaze position following incidental cue onset separately for left and right cue trials. Heat maps of gaze density were created by computing the 2D histograms of eye-position from 0.2-0.8 s from cue onset separately for the two possible locations of the cued memory item. I converted histogram counts to density by dividing by the trial number for each location condition before subtracting out density common to all locations. Finally, I took the mean across participants[8].

To obtain an aggregate measure of 'towardness', I took the mean horizontal gaze position for right cue trials minus the mean gaze position for left cue trials, divided by two, averaged across the time-window following incidental cue onset (0.2–0.8) (van Ede, Chekroud, & Nobre, 2019). I tested for significant gaze bias toward the cued item using a one-sample t-test (one-tailed, >0). In an exploratory analysis, I also tested for significant towardness at each time-point across the time-window from 0-1 (one-sample t-test, two-tailed) using a cluster-based permutation test (10,000 permutations) to correct for multiple comparisons over time. I used a cluster-forming and cluster significance

[8] The precise methods for plotting the eye-position heat maps were not included in the pre-registration.

threshold of $p<.05$. Time-resolved towardness values were smoothed with a Gaussian kernel (SD=10 ms) for visualisation and statistical tests.

Orientation decoding

To test whether the patterns of activity in the EEG channels contained information about the orientation angles of the items in memory, I used Mahalanobis distance to calculate the trial-wise distances between the multidimensional activity at each time point for each of the 160 possible orientation angles in memory, grouped into 16 angle bins distributed around a 180 degree angle space. The left and right memory items were decoded separately and independently for each participant. As in Wolff et al. (2017), I performed the analyses on the 17 posterior channels (P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2).

The decoding procedure followed an 8-fold cross-validation approach to calculate the decoding accuracy of the orientation of interest for each trial. The activity pattern of the trials of the testing fold at a particular time-point were compared to the trials of k-1 training folds. The trials of the training folds were averaged into 16 orientation bins relative to the test trial orientation, each bin containing trials with orientations within a range of 11.25°. As trials were rejected due to artefacts, I ensured equal trial numbers in each bin by random subsampling. I computed the pairwise Mahalanobis distances between the test trials and each of the orientation bins using the covariance matrix estimated from trials in the k-1 training folds, using a shrinkage estimator (Ledoit & Wolf, 2004). To obtain a visual representation of a tuning curve, the 16 distances were ordered as a function of orientation difference, mean-centered, and sign-reversed, so higher values represented greater similarity between the test trial and the training set. This was repeated for all train and test fold combinations. To obtain reliable estimates, the above procedure was repeated 100 times (with random folds each time), separately for eight orientation spaces used for binning the training trials with respect to the test

trial (bin centres: 1.41°-170.16°, 2.81°-171.56°, 4.22°-172.97°, 5.63°-174.38°, 7.03°-175.78°, 8.44°-177.19°, 9.84°-178.59° and 11.25°-180°, each in steps of 11.25°). The resulting 800 samples (100 repetitions×8 orientation spaces) were averaged for each of the 16 Mahalanobis distance values at each time point.

Finally, to obtain a summary measure of decoding accuracy, I computed the cosine-weighted means of the tuning curves (Sprague et al., 2016; Wolff et al., 2017). Higher values reflect greater orientation tuning and thereby greater decoding quality for that trial. I took the mean over trials, resulting in a single decoding value for each time-point for each participant.

To test for significant orientation decoding of the two items, I took the mean decoding accuracy within the time-window following array onset (0–1 s) and following cue onset (0–1 s). During the WM array window, I expected significant orientation decoding for both items. I used one sample t-tests, to test for significant decoding of each item (one-tailed, >0). During the cue window, I expected orientation decoding to be significant for the cued item and decoding for the uncued item to be significantly lower than the cued item. Again, I used one sample t-tests, to test for significant decoding of each item (one-tailed, >0) and a paired sample t-test (one-tailed, cued>uncued) to test for any difference in decoding for the two items.

For completeness, I also tested for significant decoding of each memory item and for any difference between them at each time-point (one-sample t-test, one-tailed, >0, paired samples t-test, two-tailed) using a cluster-based permutation test (10,000 permutations) to correct for multiple comparisons over time. I used a cluster-forming and cluster significance threshold of $p<.05$. Time-resolved decoding values were smoothed with a Gaussian kernel (SD=10 ms) for visualisation and statistical tests.

Exploratory analyses

Spatiotemporal decoding of orientation from EEG

Recent literature has shown that decoding accuracy may be improved by taking advantage of the dynamic nature of ERPs (Grootswagers et al., 2016; Wolff, Jochim, et al., 2020; Wolff, Kandemir, et al., 2020b). This approach pools the relative voltages over time and space, thus leveraging information encoded in temporal as well as spatial patterns. I complement the standard decoding with this approach focusing on a 400 ms time window from 100-500 ms following onset of the event of interest (WM array onset and cue onset) as described in Chapter 2. First, the data were down-sampled by taking the average every 10 ms. I then zero-centered the voltage fluctuations within the time-window by subtracting the mean voltage over the full time-window from the voltage at each time point. This relative baseline is comparable to pre-stimulus baseline correction, common in ERP analyses, but further removes stable neural activity within the time-window and is optimised for decoding stimulus-evoked, dynamic activity. The resulting 40 values for each posterior channel were concatenated and used as the input to the multivariate decoder. The decoding procedure followed the same steps described above, but resulted in a single decoding value for each time-window of interest. As above, the procedure was repeated 100 times (with random folds each time), separately for eight orientation spaces and averaged across the resulting 800 samples for each of the 16 Mahalanobis distance values at each time-point.

As above, I used one sample t-tests, to test for significant decoding of each item (one-tailed, >0). During encoding, I expected decoding strength to be significant for both items. During the cue window, I expected orientation decoding to be significant for the cued item and decoding strength to be greater for the cued than uncued item (cued>uncued).

Relationship between alpha lateralization/gaze bias/decoding and behaviour (error/RT)

To test if alpha lateralization, gaze bias and orientation decoding quality for the cued item was related to behavioural performance on the orientation recall task, two general linear models were fitted for each participant with trial-wise alpha lateralization/gaze bias/decoding averaged over the time-window of interest (0.2-0.8 s from cue onset for alpha lateralization and gaze bias; 0-1 s from cue onset for decoding analyses) as the predictor variable. The outcome variable was log-transformed RTs in one model and absolute error in the other. The design matrices included a constant term. I fit the models separately for congruent and incongruent trials and obtained β weights for each participant.

I performed one-sample t-tests on the β values. I expected contralateral alpha suppression to predict lower RT/error on congruent trials (one-tailed, >0) and higher RT/error on incongruent trials (one-tailed, >0). I expected greater gaze bias and decoding quality following the cue to predict lower RT/error on congruent trials (one-tailed, <0) and higher RT on incongruent trials (one-tailed, <0).

To test whether potential relationships to behaviour were specific to particular time-points in the post-cue time-window, I fit the regression models over time to obtain β weights for each time point during the delay following cue onset and performed cluster-corrected t-tests across a time-window from 0-1 s following cue onset (two-tailed, cluster-forming threshold=.05, 10,000 permutations). Time-resolved β values were smoothed with a Gaussian kernel (SD=10 ms) for visualisation and statistical tests.

While I expect the regression analysis to be most powerful to detect trial-wise variance, it assumes a normal distribution across trials. However, the distribution of trial-wise RTs and absolute error tends to be positively skewed. Consequently, I performed a secondary analysis, less sensitive to potential skews in the distribution. For each

participant, I split trial-wise measures based on the median RT/absolute error and calculated the mean alpha lateralization/gaze bias/decoding for high and low RT/error trials respectively, separately for congruent and incongruent trials. I performed a 2×2 repeated measures ANOVA with factors congruency (congruent vs incongruent) and RT/error (low vs high). Planned one-sample t-tests were performed to check for significant alpha lateralization (one-tailed, <0), gaze bias (one-tailed, >0) and decoding (one-tailed, >0) for each condition and paired-samples t-tests were performed to check for significant differences between high and low RT/error trials (two-tailed).

Probe processing

As participants were required to turn the dial from the random orientation of the probe to the appropriate orientation of the probed item (target), the absolute distance between the random probe orientation and the item in memory quantifies the response magnitude. Recent theoretical perspectives predict that prioritised content in WM is configured in an action-oriented format optimized for computing the appropriate response at retrieval (A. Heuer et al., 2020; Myers et al., 2017; Olivers & Roelfsema, 2020). In this analysis, I aimed to test whether the computation of the probe-target distance was modulated by whether the probed item in memory was previously cued (congruent) or not (incongruent).

- *Decoding probe-target distance before response onset*

I performed the orientation decoding procedure on the absolute angular distance between the target item in memory and the randomly oriented probe presented at orientation recall (probe-target distance). The analysis was performed on the time-window from –0.5-0 s before the response onset to avoid any motor confounds (response-locked). Since reaction times were mostly longer than 0.5 s, this epoch generally began after the onset of the probe (Figure 6B). Training and testing was done across all trials, after which trials were sorted and averaged according to congruency condition

(congruent vs incongruent). One-sample t-tests were performed to test for significant probe-target distance decoding for congruent and incongruent trials within in time-window of interest (–0.5-0s; one-tailed, >0) and a paired-samples t-test was performed to test for a significant difference between them (two-tailed). Additionally, cluster-corrected t-tests were performed across the full time-window prior to response onset (–1.5-0s; two-tailed, cluster-forming threshold=.05, 10,000 permutations). Time-resolved decoding values were smoothed with a Gaussian kernel (SD=10 ms) for visualisation and statistical tests.

- *Decoding probe-cue distance following probe onset*

Next, I asked whether the probe stimulus automatically triggers a computation of the response magnitude relative to the *cued* memory item. Using the same decoding procedure described above, I decoded the absolute distance between the orientation of the *cued* item and the orientation of the random probe (probe-cue distance). Note that on congruent trials, the cued item is the same as the probed item, whereas on incongruent trials, the cued item is the *non-probed* item. Thus, decoding strength of the probe-cue distance on incongruent trials may indicate the extent to which the cued item is incorrectly used as a template for comparison at probe onset.

The aim of this analysis was to determine the extent to which the probe stimulus triggers computation of a comparison signal for the cued item in memory, even when it is irrelevant on incongruent trials. To capture the probe-evoked neural response, the analysis was time-locked to the probe-onset (0-0.5s) rather than to response onset as above. One-sample t-tests were performed to test for significant comparison-signal decoding for congruent and incongruent trials within in time-window of interest (0-0.5s; one-tailed, >0) and a paired-samples t-test was performed to test for a significant difference between them (two-tailed). Additionally, cluster-corrected t-tests were performed across the full time-window prior to response onset (0-1.45s; two-tailed,

cluster-forming threshold=.05, 10,000 permutations). Time-resolved decoding values were smoothed with a Gaussian kernel (SD=10 ms) for visualisation and statistical tests.

Relationship between probe processing and behavioural error

To test if decoding strength of the probe-target / probe-cue distance was related to behavioural performance on the orientation recall task I fit a general linear model with EEG decoding values averaged over the time-window of interest (response-locked probe-*target* distance: –0.5-0s; probe-locked probe-*cue* distance: 0-0.5s) as the predictor variable and absolute error as the outcome variable. The design matrix included a constant term. I fit the models separately for congruent and incongruent trials. I performed one-sample t-tests on the β weights to test whether the quality of comparison-signal decoding was associated with absolute error. I expected comparison-signal decoding relative to the *probed* item to predict lower error on both congruent and incongruent trials (one-tailed, <0). I expected comparison-signal decoding relative to the *cued* item to predict lower error on congruent trials (one-tailed, <0), but higher error on incongruent trials (one-tailed, >0).

For completeness, I fit the regression model over time to obtain β weights for each time point (response-locked probe-*target* distance: –1.5-0 s; probe-locked probe-*cue* distance: 0-1.45 s) and performed cluster-corrected t-tests (two-tailed, cluster-forming threshold=.05, 10,000 permutations) to test for β weights significantly different from zero and for significant differences between congruent and incongruent trials. Time-resolved β values were smoothed with a Gaussian kernel (SD=10 ms) for visualisation and statistical tests.

I next split trial-wise decoding values for each participant based on median absolute error and calculated mean decision-variable decoding across the time-window of interest over trials with high error and trials with low error, separately for congruent and incongruent trials. I performed a 2×2 repeated measures ANOVA with congruency

(congruent vs incongruent) and error (low vs high) as factors. Planned one-sample t-tests were performed to check for significant decoding (one-tailed, >0) for each condition and paired-samples t-tests were performed to check for significant differences between high and low error trials (two-tailed).

Results (pre-registered analyses)

Incidental task performance

Participants performed well on the incidental task (Figure 1B) with mean accuracy of .972 (SD=.021). Mean commission error rate was .017 (SD=.014) and mean omission error rate was .071 (SD= .054). The mean median RT on correct trials was .592 s (SD=.052).

Responses are faster and more precise for congruent trials

The raw behavioural data for orientation recall are shown in Figure 2A. As predicted, mean absolute error was significantly lower for congruent (M=13.75, SD=3.22) compared to incongruent trials (M=20.67, SD=6.08): t_{25}=-7.01, p<.001, d=-1.38. Similarly, RTs were significantly lower for congruent (M=.55, SD=.12) compared to incongruent trials (M=.67, SD=.15): t_{25}=-8.35, p<.001, d=-1.64. Consequently, selecting the cued colour privileges memory for the orientation belonging to the same item.

Congruency affects all sources of error

To investigate the source of reduced error on congruent trials relative to incongruent trials, I fit a swap model to the trial-wise distribution of errors within each participant. The parameters of the swap model are shown in Figure 2B. Precision (K) was significantly higher for congruent (M=4.50, SD=1.30) than incongruent (M=3.27, SD=1.27) trials: t_{25}=6.31, p<.001, d=1.24. The same was true for the target rate

(congruent: M=.94, SD=.06; incongruent: M=.79, SD=.14): t_{25}=5.74, p<.001, d=1.13. Guess rates were significantly lower for congruent (M=.05 , SD=.05) than incongruent (M=.17 , SD=.13) trials: t_{25}=-4.73, p<.001, d=-.927, and the same was true for non-target rates (congruent: M=.02, SD =.02; incongruent: M=.04, SD=.06): t_{25}=-2.20, p=.019, d=-.431. Broadly confirming the results of Chapter 2, incidental cueing affects all sources of error in WM.

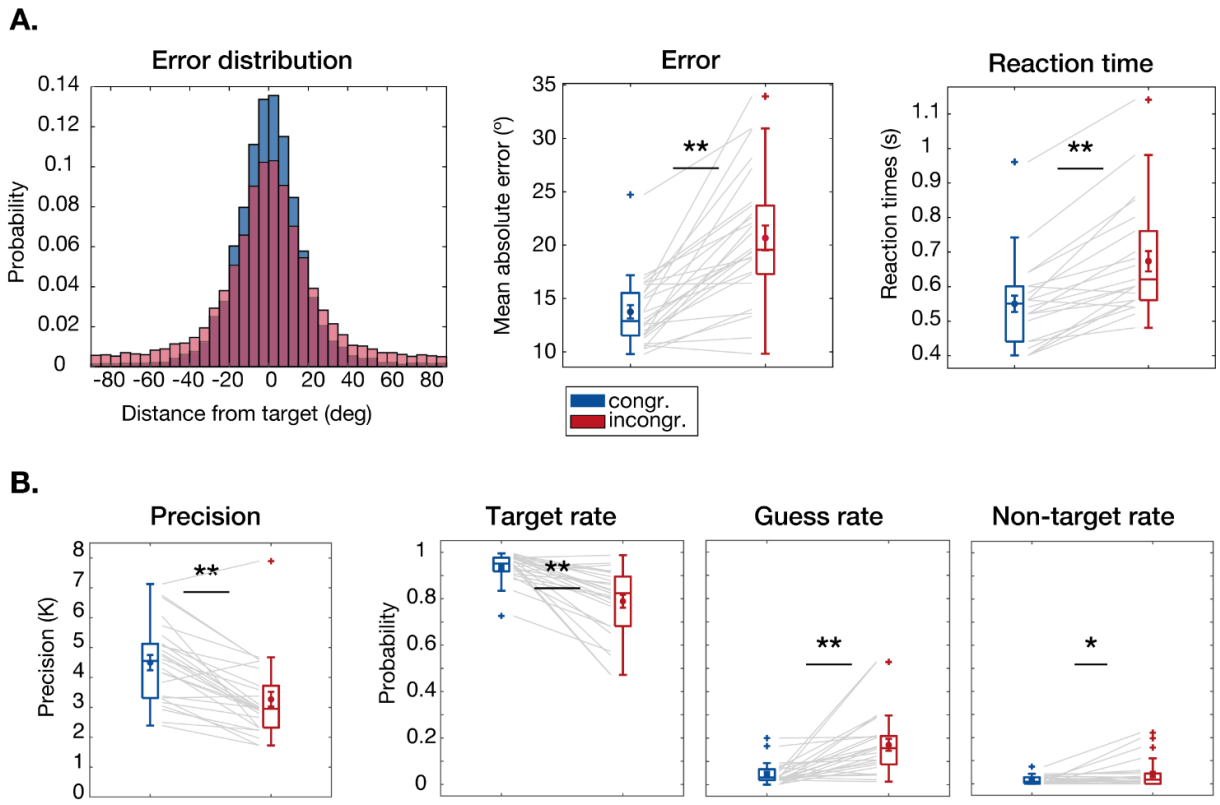

Figure 2. Behavioural results. **A.** Raw behavioural data. Left panel: histogram of error distribution relative to target orientation across all participants for congruent (blue) and incongruent (red) trials. Middle panel: boxplot of mean absolute error. Right panel: boxplot of median reaction times (time until response initiation). **B.** Swap model parameters. Boxplots of precision (K), target rate, guess rate and non-target rate. Grey lines show individual participants. Mean and standard error of mean (SEM) overlaid. *p<.05, **p<.001.

Alpha suppression contralateral to the original position of cued item

If spatial attention was directed toward the cued colour's original position, I expected to see contralateral alpha suppression. Figure 3A shows contralateral minus ipsilateral power following cue onset. In line with the hypothesis, contralateral alpha power (8-14

Hz) was significantly below zero in the time-window following incidental cue onset (0.2-0.8 s): t_{25}=-4.38, p<.001, d=-.877. Time-resolved alpha lateralization following cue onset (0-1s) showed a significant cluster from 0.2 to 0.71 s (p<.001). Thus, colour-selection in WM is associated with transient suppression in contralateral alpha power, even when memory for location is not demanded by the task.

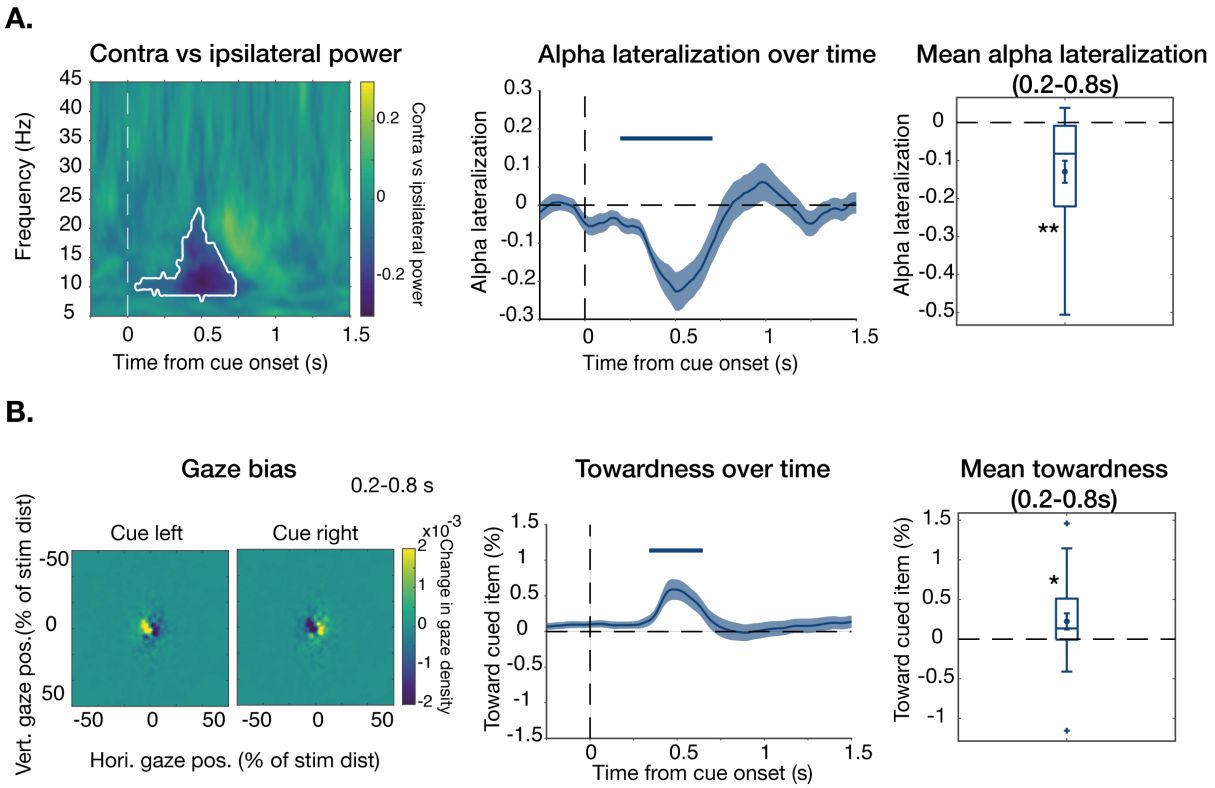

Figure 3. Alpha lateralization and gaze bias **A.** Alpha lateralization. Left: Contra minus ipsilateral power relative to original position of cued item over time from cue onset. White contours show cluster of significant contralateral suppression. Middle: Mean contra minus ipsilateral alpha power (8-14 Hz) over time from cue onset. Shaded area shows SEM. Right: Boxplot of mean alpha lateralization across TOI 0.2-0.8 s after cue onset. Mean and SEM overlaid. **B.** Gaze bias. Left: 2D histogram of gaze density for cue left and cue right trials (subtracting out mean gaze across conditions). Middle: mean horizontal gaze position for right cue trials minus the mean gaze position for left cue trials, divided by two ("towardness") over time from cue onset. Positive values indicate gaze bias toward cued item. Right: Boxplot of mean towardness across TOI 0.2-0.8 s after cue onset. *p<.05, **p<.001.

Gaze bias toward original position of cued item

The results from the gaze bias analysis are shown in Figure 3B. As expected, mean gaze position was significantly biased toward the original position of the cued item (M=.223%, SD=.529%) following the incidental cue (0.2-0.8s): t_{25}=2.15, p=.021, d=.423. It is worth

noting that this bias was tiny (0.223% of 6°=0.013° visual angle) and as evident from the 2D histogram (3B; left) did not arise from overt saccades to the item's location on a proportion of trials, but rather small shifts in gaze position. Time-resolved towardness following incidental cue onset (0.2-0.8s), showed a significant cluster from 0.37 to 0.63 s (p=.038), suggesting colour-selection triggers a transient shift in gaze position toward the cued item's position.

Decoding the cued item's orientation from EEG

The orientations of both items were decodable following encoding during the WM array epoch (0-1s; Figure 4A). Even though the cue does not appear until later in the trial, I use the 'cued' and 'uncued' labels to keep consistent with other analyses. As expected, orientation decoding for the cued item (M=.0013, SD=.0011) and the uncued item (M=.0013, SD=.0015) were both significantly above zero (cued: t_{25}=5.66, p<.001, d=1.11; uncued: t_{25}=4.37, p<.001, d=.856) and there was no significant difference between them (p=.966). The time-resolved analysis over the WM epoch (0-1 s) showed a cluster of significant decoding strength for the cued item from 0.07 to 0.76 s following WM array onset (p<.001) and from 0.06 to 0.80 s for the uncued item (p<.001) and there was no significant difference in decoding strength between them.

In the cue epoch (TOI: 0-1 s from cue onset; Figure 4B), decoding strength of the cued item's orientation was significantly greater than zero (M=4.90×10^{-4}, SD=.0014): t_{25}=1.85, p=.038, d=.362. This was not the case for the uncued item (p=.417). However, there was no significant difference between cued and uncued item decoding (p=.155). The time-resolved analysis revealed a cluster of significant decoding of the cued item's orientation from 0.30 to 0.54 s following cue onset (p=.025). There was no cluster of significant decoding for the uncued item, but again there was no difference in decoding strength between the cued and uncued items. As predicted, the results show that the non-predictive colour cue strengthens neural decoding for the cued item's orientation, even

though orientation is irrelevant at the time of the cue. However, orientation decoding strength was not significantly different between the cued and uncued item.

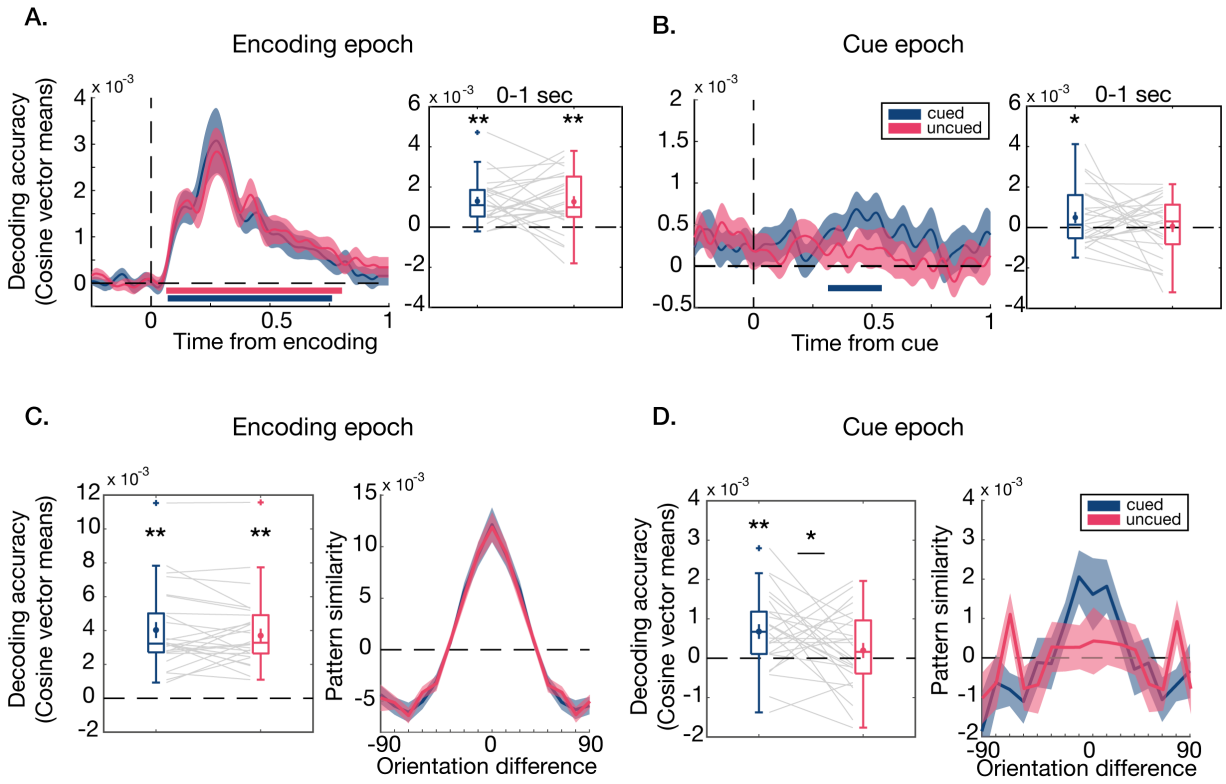

Figure 4. Orientation decoding from EEG. **A.** Left: Decoding accuracy over time following encoding. Shaded area shows SEM. Bars show clusters of significant orientation decoding. Boxplot shows mean decoding across TOI (0-1s from WM array onset). Overlaid mean and SEM. Grey lines show individual participants. **B.** Same as A but for cue epoch. **C.** Exploratory analysis: Spatiotemporal orientation decoding from EEG. Left: Boxplot of decoding accuracy following encoding (100-500 ms). Right: Mahalanobis distance pattern similarity as a function of orientation difference. Shaded area shows SEM. **D.** Same as C but for cue epoch (100-500 ms). *p<.05, **p<.001.

Decoding from M/EEG may be confounded by fixational eye-movements (Mostert et al., 2018; Quax et al., 2019). As a control, I repeated the same orientation decoding analysis for the eye-position data (Appendix B; Supplementary Figure S1). While orientation decoding from eye-position was not significant following cue onset, two participants were outliers and showed strong orientation decoding for the cued item. Thus, I repeated the EEG and eye-position decoding analyses excluding these participants to show that orientation decoding remained significant for posterior EEG channels, but was at zero for eye-position.

Results (exploratory analyses)

Spatiotemporal decoding of the orientation of the cued item

I applied a spatiotemporal decoding approach to boost power to detect neural patterns coding for memorised orientations using EEG. As expected, the orientations of both items could be decoded following the onset of the WM array (100-500 ms): cued (M=.0040, SD=.0024): t_{25}=8.45, p<.001, d=1.69; uncued (M=.0040, SD=.0022): t_{25}=9.07, p<.001, d=1.81, and there was no difference in decoding strength between the two items (p=.756). After cue-onset, the orientation of the cued item could be decoded (M=6.79×10^{-4}, SD= 9.46×10$^{-4)}$: t_{25}=3.59, p<.001, d=.703, but not the orientation of the uncued item (M=1.97×10^{-4}, SD=9.68×10$^{-4)}$: p=.159, and decoding strength was significantly higher for the cued item compared to the uncued item (t_{25}=1.80, p=.042, d=.354). The results of the spatiotemporal decoding analysis broadly confirm the standard decoding results and show that colour-selection strengthens neural decoding of the cued item's orientation. Further, orientation decoding following the cue was shown to be significantly stronger for the cued than the uncued item, in line with the pre-registered hypothesis.

Contralateral alpha suppression predicts RTs on congruent trials and behavioural error on incongruent trials

If lateralized alpha following the cue plays a functional role in selecting the cued item in a way that privileges the selected content in WM, contralateral alpha suppression should be associated with faster and more precise performance on congruent trials. If selection of the cued item comes at the expense of the uncued item in memory, alpha suppression might also be associated with slowed and less precise performance on incongruent trials.

- *Reaction times*

In support of the hypothesis, contralateral alpha suppression predicted faster RTs on congruent trials (Figure 5A; right; t_{25}=2.20, p=.019, d=.432). While β weights were mostly negative on incongruent trials, contralateral alpha suppression did not significantly predict slowed RTs on those trials (p=.207) and there was a significant difference in βs between congruent and incongruent trials (t_{25}=2.18, p=.039, d=.257). Fitting the regression model over time (Figure 5A; left) revealed a cluster of significant positive β weights for congruent trials from 0.56 to 0.81 s (p=.025) and βs were significantly different between congruent and incongruent trials from 0.52 to 0.73 s (p=.030). This indicates that the more the colour cue triggered contralateral alpha suppression, the faster participants were to initiate a response to that item when it was subsequently probed. However, this benefit did not come at the expense of slowed responses when the other item was probed.

Figure 5B shows alpha lateralization over congruent trials with fast and slow RTs. A 2×2 repeated measures ANOVA with factors congruency (congruent vs incongruent) and RT (fast vs slow) revealed no main effect of congruency (p=.279), no main effect of RT (p=.208) and no interaction (p=.378). On congruent trials, alpha lateralization was significant for both slow (t_{25}=-2.73, p=.006, d=-.536) and fast (t_{25}=-5.14, p<.001, d=-1.01) RT trials, and although there was a trend for greater lateralization for fast than slow RT trials, they were not significantly different (t_{25}= 1.87, p=.073, d=.367). On incongruent trials, lateralization was again significant for both slow (t_{25}=-3.10, p=.002, d=-.607) and fast (t_{25}=-2.22, p=.018, d=-.435) RT trials, and there was no difference between them (p=.769).

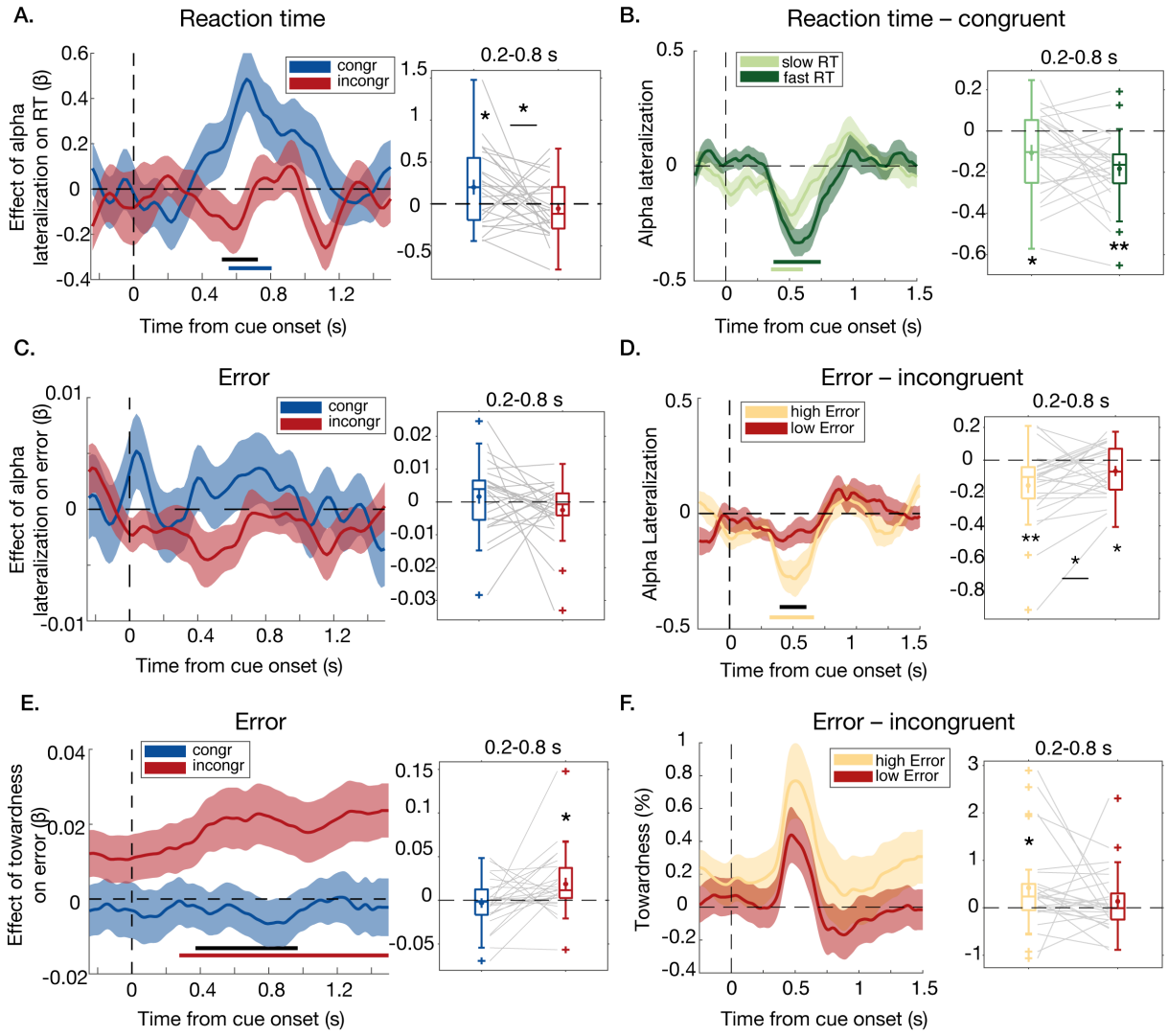

Figure 5. Relationship between alpha lateralization and reaction times (A-B), alpha lateralization and recall error (C-D) and gaze bias and recall error (E-F). **A.** Beta weights from trial-wise regression with time-resolved alpha lateralization and reaction times for congruent (blue) and incongruent (red) trials. Blue bar shows non-zero beta-weights for congruent trials and black bar shows significant difference between congruent and incongruent trials. Shaded area shows SEM. Boxplots show beta weights from regression with mean alpha lateralization across post-cue TOI (0.2-0.8 s) and reaction times. Mean and SEM overlaid. Grey lines show individual participants. **B.** Alpha lateralization for high (light green) vs low (dark green) congruent RT trials. Bars show significant lateralization. Boxplots show mean lateralization decoding across post-cue TOI (0.2-0.8 s). **C.** Same as A for regression between alpha lateralization and absolute error. **D.** Same as B, but with alpha lateralization for high (yellow) and low (red) error incongruent trials. **E.** Same as A for regression between gaze bias and absolute error. Red bar shows non-zero beta-weights for incongruent trials and black bar shows significant difference between congruent and incongruent trials. **F.** Same as B, but with gaze bias for high (yellow) and low (red) incongruent error trials. *p<.05, **p<.001.

- *Absolute error*

Figure 5C shows β weights for the regression analysis between contralateral alpha suppression and absolute error. βs were not significantly above zero on congruent trials

(p=.231) and although there was a trend for negative beta weights on incongruent trials, this did not reach significance (t_{25}=-1.42, p=.084, d=-.278). There was no difference between conditions (p=.202). Fitting the regression for each time-point also revealed no significant clusters for either congruent or incongruent trials. These results contrast with Chapter 2 as they show no evidence that alpha lateralization improves memory precision and instead suggest alpha lateralization is important for making the selected item more accessible.

However, the median split analyses did suggest a relationship between alpha suppression contralateral to the cued item and higher error on incongruent trials. Alpha lateralization on incongruent trials with high and low error is shown in Figure 5D. There was no main effect of congruency (p=.286), no main effect of error (p=.325) but a trend for a congruency by error interaction ($F_{1,25}$=4.16, p=.052). On congruent trials, alpha lateralization was significant for both low (t_{25}= -6.13, p<.001, d=-1.20) and high (t_{25}=-2.80, p=.005, d=-0.549) error trials, and there was no significant difference between conditions (p=.452). On incongruent trials, lateralization was again significant for both low (t_{25}=-2.12, p=.022, d=-.416) and high (t_{25}=-3.34, p=.001, d=-.656) error trials, but suppression was significantly greater on high than low error trials (t_{25}=-2.23, p=.035, d=-.438), suggesting alpha suppression contralateral to the cued item's original position may promote errors when the other item is probed.

Gaze bias predicts behavioural error on incongruent trials

The results of the regression analysis between gaze bias and absolute error are presented below. I did not find any relationship between gaze bias and RT (Appendix B; Supplementary Figure S3).

There was no significant relationship between gaze bias and absolute error for congruent trials (Figure 5E; right; p=.286). On incongruent trials, gaze bias toward the cued item

was associated with higher error (t_{25}= 2.56, p=.008, d=.503) and β weights were significantly greater for incongruent than congruent trials (t_{25}=-2.38, p=.025, d=-.467). The results indicate that the more participants' gaze was biased toward the cued item in memory, the higher the error when the other memory item was probed. However, inspecting the time-resolved gaze position, this was a general property of gaze bias throughout the trial rather than specific to cued selection. Fitting the regression model over time (Figure 5E; left) revealed a cluster of significant positive beta weights for incongruent trials from 0.28 s following cue onset until onset of the probe (1.5 s; p=.010) and there was a significant difference in β weights between congruent and incongruent trials from 0.37 to 0.97 s from cue onset (p=.034).

Towardness for incongruent trials with high and low absolute error is shown in Figure 5F. A 2×2 repeated measures ANOVA revealed no main effect of congruency (p=.373), no main effect of error (p=.452) and no interaction (p=.138). On congruent trials, although gaze bias toward the cued item's location was slightly higher for low error than high error trials, neither was significant (low: p=.056; high: p=.183) and there was no difference between them (p=.334). On incongruent trials (Figure 5F), gaze bias was significant only for high error trials (t_{25}= 2.26, p=.017, d=.442) and not for low error trials (p=.157), but there was no significant difference between them (p=.211).

Together, the results suggest gaze bias toward to the cued item's original position comes at the expense of the uncued item, resulting in greater error on incongruent trials.

No evidence that post-cue orientation decoding predicts behavioural performance

If neural orientation decoding signals successful selection of the cued item, decoding strength should be associated with faster and more precise performance on congruent trials and potentially slower and less precise performance on incongruent trials. However, I did not find any evidence for a relationship between post-cue orientation

decoding and RT on either congruent (p=.752) or incongruent (p=.788) trials and there was no difference between them (p=.344; supplementary material Figure S4). Similarly, there was no relationship between post-cue orientation decoding and absolute error on congruent (p=.825) or incongruent trials (p= .612) and no difference between them (p=.395). I fit the regression model over time to obtain beta weights for each time during the delay following cue onset and found no clusters of significant β weights for either congruent or incongruent trials.

For completeness, the same regression model was fitted using the trial-wise spatiotemporal orientation decoding results instead. Again, there was no evidence for a relationship between orientation decoding and RT (congruent: p=.345; incongruent: p=.752; difference: p=.858) or absolute error (congruent: p=.621; incongruent: p=.669; difference: p=.603).

Decoding probe-target distance from EEG prior to response onset

The absolute distance between the random probe and the probed memory item (probe-target distance) could be decoded prior to response onset on both congruent and incongruent trials (TOI:-0.5-0 s; congruent: t_{25}=4.83, p<.001, d=.947; incongruent: t_{25}=3.76, p<.001, d=.737; Figure 6A). As reaction times were mostly longer than 0.5 s, this epoch generally began after the onset of the probe (Figure 6B). Although decoding strength was better on congruent than incongruent trials there was no difference between the conditions (p=.151). Cluster-corrected t-tests across the time window prior to response onset (-1.5-0 s) revealed clusters of significant decoding on congruent trials from -0.61 to 0 s before response onset (p<.001) and on incongruent trials from -0.44 to 0 s (p<.001). Thus, leading up to response initiation, the probe-target distance could be decoded from posterior channels in EEG and this signal was not modulated by whether the item was previously cued or not.

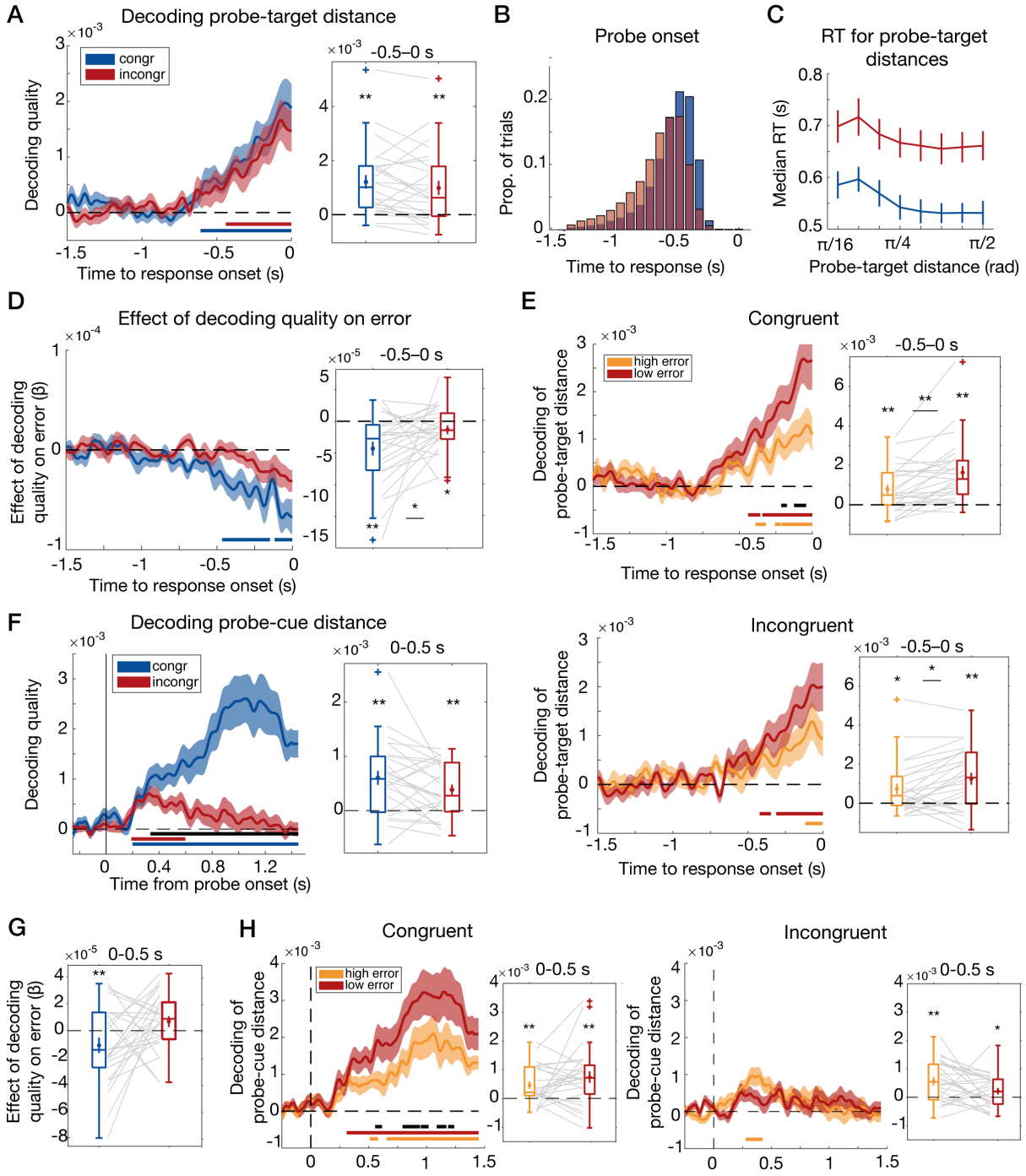

Figure 6. Probe-target distance decoding from EEG voltages. **A.** Decoding of probe-target distance (absolute distance between random probe orientation and orientation of *probed* item in memory) for congruent (blue) and incongruent (red) trials time-locked to response onset. Shaded area shows SEM. Bars show significant decoding. Boxplot shows mean decoding strength before response onset (-0.5-0 s). Mean and SEM overlaid. Grey lines show individual participants. **B.** Histogram of probe onset relative to response onset across participants for congruent (blue) and incongruent (red) trials. **C.** Median RT as a function of probe-target distance for congruent (blue) and incongruent (red) trials. Error bars show SEM. **D.** Beta weights from trial-wise regression with time-resolved probe-target distance decoding and absolute error for congruent (blue) and incongruent (red) trials. Bars show non-zero beta weights. Boxplots shows beta weights from regression with mean decoding in TOI (-0.5-0s) and absolute error. **E.** Probe-target distance decoding locked to response onset for high (yellow) vs low (red) error trials on congruent (top) and incongruent (bottom) trials. Red/yellow bars show significant decoding. Black bar shows significant differences between high and low error trials. Boxplots show mean decoding across TOI (-0.5-0s). **F.** Decoding probe-*cue* distance (absolute

distance between random probe orientation and orientation of *cued* item in memory) for congruent (blue) and incongruent (red) trials time-locked to probe onset. Red/blue bars show significant decoding. Black bar shows significant difference in decoding between conditions. Boxplot shows mean decoding strength across TOI after probe onset (0-0.5 s). **G.** Boxplot shows beta weights from regression with mean probe-cue distance decoding in TOI (0-0.5 s from probe onset) and absolute error. **F.** Probe-cue distance decoding for high (yellow) vs low (red) error trials on congruent (left) and incongruent (right) trials. Red/yellow bars show significant decoding. Black bar shows significant differences between high and low error trials. Boxplots show mean decoding across TOI (0-0.5 s).*p<.05, **p<.001.

However, RTs varied as a function of probe-target distance with slower RTs when the probe-target distance was small (Figure 6C). Consequently, probe-target distance decoding may be confounded by a temporal signal of how recently the probe appeared. To control for this possibility, I computed the residual decoding values after regressing out trial-wise RT (Appendix B; Supplementary Figure S5A). Although residual decoding values were reduced, they remained significant for both congruent and incongruent trials prior to response onset.

Probe-target distance decoding quality predicts error to a greater extent on congruent than incongruent trials

Figure 6D shows β weights from the regression analysis between the strength of probe-target distance decoding and absolute error. In support of the hypothesis, β weights were significantly below zero for both congruent (t_{25}=-3.77, p<.001, d=-.740) and incongruent (t_{25}=-1.97, p=.030, d=-0.387) trials (Figure 6D), but decoding-strength predicted error to a greater extent on congruent than incongruent trials (t_{25}=-2.20, p=.037, d=-.432). Cluster-corrected t-tests across the time window prior to response onset (-1.5-0 s) revealed significant clusters of negative beta weights for congruent trials prior to response onset from -0.47 to -0.15 s (p=.004) and again from -0.12 to 0 (p=.028). There were no clusters of significant correlations between probe-target distance decoding and error on incongruent trials and no significant differences between congruent and incongruent trials. As above, the regression analysis was repeated with residual probe-

target distance decoding values after regressing out RTs (supplementary material; Figure S5B). The results were similar, except the relationship between decoding quality and error was no longer significant on incongruent trials.

The median-split analyses broadly confirm the regression results. Probe-target distance decoding strength for high and low error trials is shown in Figure 6E, separately for congruent (left) and incongruent (right) trials. There was a main effect of error, with greater decision-variable decoding on low than high error trials ($F_{1,25}$=19.56, p<.001). There was no significant main effect of congruency (p=.151) and no congruency by error interaction (p=.321). On congruent trials, decision-variable decoding was significant for both low error trials (t_{25}=5.13, p=<.001, d=1.01), and high error trials (t_{25}=3.63 , p<.001, d=.712) and decoding was significantly greater on low than high error trials (t_{25}=-3.81, p<.001, d=-.747). On incongruent trials, decision-variable decoding was significant for both low error trials (t_{25}= 4.11, p<.001, d=.807) and high error trials (t_{25}= 2.81, p=.005, d=.550), and decoding was significantly greater on low than high error trials (t_{25}=-2.51, p=.019, d=-.493).

In summary, while the strength of the probe-target distance decoding was similar across the two conditions, it was a better predictor of response precision on congruent than on incongruent trials.

Decoding probe-cue distance on incongruent trials

To test if the probe stimulus triggers computation of the distance between the probe and the *cued* memory item, I decoded the absolute distance between the cued item and the random probe stimulus (probe-cue distance) from posterior EEG channels. Interestingly, the probe-cue distance could be decoded on both congruent and incongruent trials following probe onset (TOI:0-0.5 s; congruent: t_{25}=4.52, p<.001, d=.887; incongruent: t_{25}=3.95, p<.001, d=.775; Figure 6F), even though it was the *other* item in memory that

was probed for recall on incongruent trials. Furthermore, there was no difference in decoding strength between conditions in the initial time-window after probe onset (0-0.5s; t_{25}=2.03, p=.053, d=.398), although it trended toward stronger decoding on congruent than incongruent trials. This suggests that the probe stimulus initially triggers computation of the relative distance of the random probe to the cued memory item, regardless of whether that item was the target for orientation recall or not. Cluster-corrected t-tests across the time window after probe onset (0–1.45 s) revealed a cluster of significant decoding on congruent trials from 0.2 to 1.45s (p<.001) and on incongruent trials from 0.19 to 0.6 s (p=.005). There was a significant difference in decoding between congruent and incongruent trials from 0.33 to 1.45 s (p<.001). While the onset of probe-*cue* distance decoding was similar for congruent and incongruent trials, decoding of the probe-*target* distance following probe onset (probe-locked) was delayed on incongruent compared to congruent trials, suggesting the probe may first have been compared to the cued memory item before it was compared to the appropriate target (Appendix B; Supplementary Figure S6).

As for probe-target distance decoding above, I computed the residual decoding values after regressing out trial-wise RT (Appendix B; Supplementary Figure S5C). Again, while decoding strength was reduced, the probe-cue distance could still be decoded in the initial 500 ms after probe onset on both congruent and incongruent trials and there was no difference between conditions.

Probe-cue distance decoding supported memory precision on congruent trials and interfered with memory precision on incongruent trials

If the probe-dependent signal relative to the cued memory item is relevant to behaviour, it may support orientation recall on congruent trials, but interfere with recall on incongruent trials. Consequently, I expected probe-cue distance decoding strength to predict lower error on congruent trials ($\beta < 0$) but greater error on incongruent trials

($\beta > 0$). In support of the hypothesis, there was a negative relationship between probe-cue distance decoding strength on congruent trials (t_{25}=-1.87, p=.037, d=-.367) and a trend for a positive relationship between probe-cue distance decoding strength and error on incongruent trials (t_{25}=1.63, p=.057, d=.320). There was a significant difference between congruent and incongruent trials (t_{25}=-2.29, p=.031, d=-.449; Figure 6G). Cluster-corrected t-tests across the time window after probe onset (0–1.45 s) revealed no significant clusters for incongruent trials. There was a significant cluster of negative beta weights for congruent trials from 0.35 to 1.32 s (p=.001) and significant differences between beta weights for congruent and incongruent trials from 0.33 to 1.31 s (p<.001).

As above, the regression analysis was repeated with residual probe-cue distance decoding values after regressing out RTs (supplementary material; Figure S5D). While the pattern of results was similar as those reported above, the negative relationship between residual probe-cue distance decoding and error from 0-0.5 s after probe onset on *congruent* trials was not significant. Instead, the positive relationship between residual probe-cue distance decoding strength and error on *incongruent* trials was significant in this analysis.

Figure 6H shows mean probe-cue distance decoding strength relative to the cued memory item for high and low error trials, separately for congruent (left) and incongruent (right) trials. A 2×2 repeated measures ANOVA revealed a trend for a main effect of congruency ($F_{1,25}$=4.04, p=.055) but no main effect of error (p=.779). There was a significant congruency by error interaction ($F_{1,25}$=4.26, p=.049). On congruent trials, probe-cue distance decoding was significant for both low error trials (t_{25}=5.13, p=<.001, d=1.01), and high error trials (t_{25}=3.63 , p<.001, d=.712) and decoding was significantly greater on low than high error trials (t_{25}=-3.81, p<.001, d=-.747). On incongruent trials, probe-cue distance could be decoded on both low error trials (t_{25}= 1.80, p=.042, d=.353)

and high error trials (t_{25}= 3.97, p<.001, d=.779), but more so on high than low error trials, supported by a trend-level difference (t_{25}= 2.03, p=.053, d=.399).

This suggests the probe-cue comparison signal promotes memory precision when the cued item is also probed, but interferes with precision when the uncued item is probed.

Discussion

Confirming the pre-registered hypotheses, colour-selection in WM engaged spatial attention and follow-up analyses showed that trial-wise measures of spatial attention predicted subsequent memory performance. As expected, the non-predictive colour cue strengthened the neural representation of the cued object's orientation and subsequent orientation recall was faster and more precise for cued than non-cued items. Together the results provide behavioural and neural evidence that attentional selection spreads between an object's features in WM. Exploratory analyses further revealed that the cued item was automatically compared with the probe at recall, even when the other memory item was probed, suggesting the cued item may be prioritised for decision-making (Myers et al., 2017; Souza & Oberauer, 2016).

Consistent with results from Chapter 2 and previous studies showing benefits of non-predictive retro-cues in WM (Souza et al., 2015; Zokaei, Manohar, et al., 2014; Zokaei, Ning, et al., 2014), orientation recall was faster and more precise when the colour of the same object was previously selected for the incidental task, compared to when the colour of the other memory item was selected. Thus, selecting the cued feature in memory for the incidental task facilitates subsequent retrieval of associated features belonging to the same WM object. In Chapter 2, I showed that the congruency benefit was not exclusively explained by higher accessibility of the cued feature (colour), as the benefit

was also seen when the cue and the probe did not share the same feature dimension.

The neural representation of the cued item's orientation re-emerged following cue, decoded from posterior channels in EEG. Previous work has shown that stimulus-selective neural areas respond more strongly to attended perceptual stimuli than unattended stimuli (Desimone & Duncan, 1995; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999). Similarly, WM content inside and outside the focus of attention show distinct neural representational formats, with attended WM content being represented more strongly in neural activity patterns compared to unattended content (Christophel et al., 2018; LaRocque et al., 2013; Larocque et al., 2014; Lewis-Peacock et al., 2012; Rose et al., 2016; Sprague et al., 2016; Wolff et al., 2017). In the present study, selecting the cued colour temporarily strengthened the neural representation of the orientation belonging to the same object relative to the orientation of another object in memory, even though orientation was irrelevant at the time of the cue. This builds a strong case for obligatory spread of attention, even between non-spatial object-features in WM, predicted by computational models with flexible conjunctive coding where retrieval is supported by automatic, associative pattern completion (Manohar et al., 2019). In contrast, automatic retrieval of the orientation in response to a colour cue is not predicted by models with fixed conjunctive coding between spatial and non-spatial features (Schneegans & Bays, 2017a, 2017b).

There was no relationship between post-cue orientation decoding magnitude and orientation recall. This was surprising given orientation was the relevant feature for read-out. However, the sample size may have been insufficient to detect individual differences, as this analysis was exploratory. Additionally, the weaker magnitude of post-cue orientation decoding in Chapter 3, compared to location decoding in Chapter 2 might have reduced power to detect significant correlations with behaviour. This should be followed up in high-powered study designed to test for individual differences to

confirm whether re-emergence of stimulus decoding reflects the extent to which the cued object is brought into a privileged memory state.

Object-based selection in WM can be contrasted with the alternative hypothesis that it is possible to selectively attend to a single object feature of a multi-feature object encoded into WM. Several studies have shown that WM selectively encodes and represents only object-features that are relevant to the current task (Bocincova & Johnson, 2019; Serences et al., 2009; Woodman & Vogel, 2008). This principle might have extended to selection and retrieval in WM, such that it would be possible to selectively access the cued colour, without simultaneously privileging associated object-features. For example, research has shown behavioural benefits for feature-based retro-cues that are of a similar magnitude to spatial or object-based retro-cues (Hajonides et al., 2019; A. Heuer & Schubö, 2016; Niklaus et al., 2017). While successful feature-based selection initially appears at odds with the current findings, it is possible that predictive feature-based retro-cues allow irrelevant features to be dropped or removed from memory, thereby freeing up WM resources and improving performance for relevant features (Souza et al., 2016). The present task did not allow participants to remove any features from memory as colour-orientation bindings had to be maintained until orientation recall. It is possible that there would not be the same spread of attention if certain object-features could be strategically dropped ahead of the cue (although location may be a special case, as discussed next).

Previous studies have shown that indices of spatial attention, such as alpha lateralization and gaze position, are biased towards the memorised location of retro-cued objects in WM (Myers, Walther, et al., 2015; van Ede, Chekroud, & Nobre, 2019; Wallis et al., 2015). Here, a centrally presented colour cue triggered transient contralateral alpha suppression and gaze bias relative to the memorised position of the cued item, even though the cue was non-predictive of the to-be-probed item and despite location

being a fully redundant feature in this task. This replicates results from Chapter 2 in a situation where location information is fully redundant, consistent with previous research showing a special status of location in binding and retrieving non-spatial features in WM (Cai, Sheldon, Yu, & Postle, 2019; Pertzov & Husain, 2014; Poch et al., 2017; Schneegans & Bays, 2017a; Treisman & Zhang, 2006; van Ede, Chekroud, & Nobre, 2019; B. Wang et al., 2016; Wood, 2011). The current results provide strong support for a role of spatial attention in selecting non-spatial content from WM. A question for future research is whether spatial attention is engaged in WM selection only when location is a distinguishing feature as it was here or also when memory items share overlapping spatial locations and location information may be counterproductive to performance (Pertzov & Husain, 2014).

The extent to which participants directed spatial attention to the cued colour's original position was predictive of orientation recall performance, suggesting spatial attention plays a functional role in WM maintenance, even of non-spatial features (Cai et al., 2019). Contralateral alpha suppression following cue onset predicted faster orientation recall on congruent trials, suggesting alpha plays a role in selecting the cued content, making it more accessible for retrieval. Alpha lateralization and gaze bias relative to the cued item's position were both associated with greater error on incongruent trials, suggesting the extent to which spatial attention was directed toward the cued item negatively affected the response precision for the other memory item. This pattern of correlations was different from Chapter 2, where contralateral alpha suppression was associated with lower error and gaze bias predicted faster RTs on congruent trials, so I hesitate to draw strong conclusions from these results until further replicated.

Discrepancies in the task designs between the two experiments may help account for the different pattern of results. In the current experiment, participants only maintained two items in memory, making the task easier and perhaps introducing ceiling effects for

orientation judgements, specifically on congruent trials. This could explain why there was no relationship with error on congruent trials. The two items were presented laterally which may have caused more direct trade-off in the distribution of spatial attention between memory items and impaired performance on incongruent trials. I did not find evidence of a switch in the alpha lateralization profile later in the delay in this experiment, perhaps because there were fewer non-cued items or because there was limited time to re-orient attention before probe onset, as the post-cue delay was reduced from 2 s to 1.5 s. Future studies should further disentangle the relationship between alpha lateralization and gaze bias as measures of spatial attention and the ways in which spatial attention facilitates WM for non-spatial features.

While the results converge to support a model where attentional selection spreads between an object's features in WM, spatial attention and the relative boost in neural decoding were transient rather than sustained following the cue. If these measures indeed track attentional states within WM, corroborated by a wealth of literature (LaRocque et al., 2013; Larocque et al., 2014; Lewis-Peacock et al., 2012; Myers, Walther, et al., 2015; Poch et al., 2014, 2017; Rose et al., 2016; Van Ede et al., 2020; van Ede, Chekroud, & Nobre, 2019; Wallis et al., 2015; Wolff et al., 2017), it raises the question of how behavioural benefits arise if not through maintaining the cued item in a sustained focus of attention.

Several studies have shown retro-cue benefits in the absence of sustained attention and instead proposed cued selection may boost an object's traces in WM, perhaps through strengthened connectivity between object-features (Rerko & Oberauer, 2013; Rerko et al., 2014; Souza et al., 2015). Enhanced memory traces might make a retro-cued item more accessible in memory and robust to interference (Makovski et al., 2008; Y. A. Tabi et al., 2019). Recent theoretical models of WM predict strengthened connectivity between co-activated features through short-term Hebbian plasticity (Fiebig & Lansner, 2017;

Manohar et al., 2019; Oberauer & Lin, 2017). It is possible that attentional selection of the cued object enhances synaptic connectivity between the object's features through short-term plasticity thereby increasing memory robustness and accessibility even in the absence of sustained attention.

An alternative, though not mutually exclusive, account suggests retro-cues may support behaviour in the absence of sustained attention if selected content is re-configured into an action-oriented state (Myers et al., 2017). Analyses of neural processing at the probe provided support for an account where there cued item is placed in a privileged state for interacting with the probe at decision-making (Olivers et al., 2011; Olivers & Roelfsema, 2020). The probe-target distance could be decoded immediately prior to response onset and decoding quality predicted memory precision to a greater extent on congruent than incongruent trials. The probe-target distance indicates the magnitude of the required response so the strength of decoding may signal how effectively the probe is compared to the relevant item in memory.

Interestingly, a comparison signal with respect to the *cued* item's orientation angle (probe-cue distance) was additionally decodable immediately after probe onset. On congruent trials, the cued item was the same that was probed for recall, but on incongruent trials this signalled an inappropriate comparison with the unprobed memory item. Nevertheless, initially after probe onset decoding of the probe-cue distance followed the same trajectory for congruent and incongruent trials. Interestingly, this appeared to come at the cost of decoding the actual *probe-target* distance on incongruent trials, which was delayed by ~100 ms. The probe may be compared to the cued item first, postponing decision-making when the other memory item is probed. This hypothesis may be tested formally by comparing the non-decision time parameter of a drift-diffusion model of decision-making between congruent and incongruent trials in a follow-up analysis. Finally, probe-cue distance decoding accuracy within the first 500 ms was a

positive predictor of memory precision on congruent trials, but a negative predictor of memory precision on incongruent trials. Consequently, the probe may initially be compared to the cued item, boosting orientation recall on congruent trials, but interfering with retrieval of the appropriate memory item on incongruent trials, contributing to the congruency effect observed here.

In conclusion, the study reproduced results of Chapter 2 showing that attentional selection of a single object-feature in WM engages spatial attention and improves WM performance for the cued object. The current study further showed that spatial attention was engaged in colour-selection in WM when spatial position was a fully irrelevant feature, consistent with literature suggesting a special role of spatial attention in selecting non-spatial features in WM (Cai et al., 2019; Pertzov & Husain, 2014; Poch et al., 2017; Schneegans & Bays, 2017a; Treisman & Zhang, 2006; van Ede, Chekroud, & Nobre, 2019; B. Wang et al., 2016; Wood, 2011). Colour-selection in WM further boosted the representation of the cued object's orientation, as decoded from EEG. Together, the results provide evidence for automatic spread of attention between an object's features in WM, consistent with computational models predicting retrieval through auto-associative pattern completion between features (Manohar et al., 2019). This can be contrasted with the alternative hypothesis that feature-based attention enables selection of a single feature of a multi-feature item in WM.

A key outstanding question remains how selected content in WM is prioritised for subsequent recall and decision-making. The current results are in line with previous reports that sustained attention is not always necessary for an item to remain in a privileged state in WM (Hollingworth & Maxcey-Richard, 2013; Rerko & Oberauer, 2013; Rerko et al., 2014; Souza et al., 2015). In the absence of sustained internal attention, attentional selection of the cued item may strengthen functional connectivity between selected features and place the cued item in a privileged state for subsequent retrieval

and decision-making (see Chapter 6 for discussion)(Manohar et al., 2019; Oberauer & Lin, 2017; Olivers & Roelfsema, 2020; Rerko & Oberauer, 2013). Prioritisation for action in WM will be the focus of Chapter 4-5.

Chapter 4. Prospective working memory: Is working memory supported by task-specific representations?[9]

Traditionally, WM has been conceived as a retrospective mechanism by which the brain represents sensory information that is no longer present in the environment (D'Esposito & Postle, 2015; Funahashi et al., 1989; Harrison & Tong, 2009; Pasternak & Greenlee, 2005; Serences, 2016). However, a crucial function of WM is to maintain information *for* a forthcoming task. This prospective aspect of WM has received relatively less attention in the literature (J. A. Miller et al., 2020; Nobre & Stokes, 2019; Oberauer, 2010; Olivers & Roelfsema, 2020; Rainer et al., 1999). In two behavioural experiments, I ask whether WM is supported by a transformation of sensory memoranda into prospective task-sets, optimised for decision-making.

Previous work has shown that WM is flexible: humans can use knowledge about a current goal to selectively encode and maintain relevant information, thereby increasing the odds of successful behaviour despite WM capacity limits (Luck & Vogel, 1997). This is highlighted by studies showing that information already encoded in WM can be selected and prioritised temporarily to improve performance, for example by a retrospective cue ("retro-cue") (Griffin & Nobre, 2003; Landman et al., 2003; Souza & Oberauer, 2016). Retro-cue benefits in visual WM have been shown at the level of objects (Griffin & Nobre, 2003; Landman et al., 2003) and object-features such as colour (A. Heuer & Schubö, 2016).

Studies using functional magnetic resonance imaging have found that brain activity preferentially represents task-relevant features or objects in WM (Lee et al., 2013;

[9] The majority of Experiment 2 has been published as a pre-print: Printzlau, F., Myers, N. E., Muhle-Karbe, P. S., Manohar, S. G., & Stokes, M. G. (2019). Prospective task knowledge improves working memory-guided behaviour. *PsyArxiv*. https://doi.org/10.31234/OSF.IO/QHKU6

Serences et al., 2009). For example, the colour but not orientation of a WM item can be decoded from BOLD activity in visual cortex during the delay when participants are instructed to report colour, but not when they were instructed to report orientation and vice versa (Serences et al., 2009). Similarly, object-identity is decodable from BOLD signals in sensory cortex or lateral prefrontal cortex depending on whether participants are asked to make decisions based on fine-grained visual details or abstract, category-level information respectively (Lee et al., 2013). In both cases, the pattern of BOLD activity carried information about the task-relevant feature in WM that was needed to perform the task. It is clear that anticipated task demands influence *what* information is stored in WM. Here I ask whether it also affects *how* the relevant information is stored.

Initial evidence suggests that the WM system modulates how the same features are stored depending on the expected context at retrieval (Geng et al., 2017; Navalpakkam & Itti, 2007; X. Yu & Geng, 2019). For example, colour representations in WM were systematically shifted when distractors were linearly separable from the target and asymmetrically sharpened when target-distractor similarity was high. Templates stored in WM were not veridical representations of the sensory input, but served to increase the distinctiveness of the template by shifting the representation away from predictable distractors (X. Yu & Geng, 2019). Together, these results suggest that the nature of WM representations may depend on their future use, in a way that optimises performance for the upcoming task.

Another way WM may be optimised for anticipated task demands is by establishing tight links between the sensory representation and the associated motor plan (van Ede, Chekroud, Stokes, et al., 2019). Using EEG, van Ede et al. (2019) showed that when WM contents were associated with specific actions, the selection of task-relevant sensory representations and appropriate motor plans occurred in parallel and not serially as would be expected if the WM code was primarily sensory in nature. Instead, the results

are consistent with a model of WM in which the brain, when possible, will link sensory representations to their appropriate motor plan to guide future behaviour (Gallivan, Bowman, Chapman, Wolpert, & Flanagan, 2016; Gallivan, Logan, Wolpert, & Flanagan, 2016; J. A. Miller et al., 2020). This contrasts with typical WM tasks, such as continuous report tasks (Bays et al., 2009) or change-detection tasks (Luck & Vogel, 1997), which do not allow for motor preparation at encoding, but instead require a comparison between a WM stimulus and an unpredictable post-delay probe target. Prospective coding in such tasks would require preparation of a more abstract task-set for items in WM, by linking each stimulus with the context-depending action (Monsell, 2003).

Oberauer (2010) distinguishes between declarative and procedural WM, with the declarative portion representing information about the world and the procedural portion holding stimulus-response (S-R) mappings or task-sets for guiding action (Oberauer, 2010). A task-set refers to the prospective configuration of cognitive processes, such as perception, attention, memory and motor planning for a specific task (Monsell, 2003; Sakai, 2008). Task-sets are traditionally studied in task-switching paradigms, where participants are asked to switch between tasks involving the same stimulus set and the same response outputs, for example, to classify numbers based on their magnitude or parity (Monsell et al., 2003). Performance costs following task switches relative to repeats are termed "switch-costs" (Monsell, 2003), and may be reduced (but not abolished) by cueing the task before stimulus onset, perhaps by allowing top-down re-configuration of the neural state for the new task ahead of time (Meiran, 1996; Rogers & Monsell, 1995), but c.f. Allport et al. (1994). Configuration of a task-set can be conceptualized as retrieval of the relevant rule-dependent S-R mappings ahead of time (e.g. for the magnitude task rule: 'press LEFT if the number is lower than 5 and press RIGHT if the number is greater than 5'), allowing for a faster and more accurate translation to a motor output once the target appears.

Recent theoretical and experimental work has suggested that WM may similarly be supported by a translation of the relevant sensory representation into a prospective S-R mapping in preparation for an upcoming target (Muhle-Karbe et al., 2017, 2020; Myers et al., 2017; Olivers & Roelfsema, 2020). For example, in a change-detection task, a red memory item may be translated into the task-set 'press RIGHT if the target is red, press LEFT for any other colour'. Encoding WM content in such a use-specific format might be adaptive as it reconfigures the memory representation for context-dependent processing, thereby enabling faster and more accurate decision-making at the time of the probe. In two behavioural experiments, I test whether WM is supported by a transformation of retrospective sensory memoranda into such prospective task-dependent decision circuits.

Experiment 1

In Experiment 1, I asked whether parametric WM delay representations are configured as prospective S-R mappings (task-sets), in preparation for the upcoming probe. I developed a task modelled after experiments used to measure task-rule congruency effects in the task switching and cognitive control literature (Braem et al., 2019; González-García et al., 2020; Liefooghe et al., 2013, 2012; Monsell, 2003). Task-rule congruency effects refer to the finding that when participants perform two different tasks sharing an overlapping set of stimuli and responses, reaction times are faster and accuracy higher when the two task rules point to the same response (congruent) compared to when they point to conflicting responses (incongruent). Studies have consistently shown that when participants encode explicit S-R mappings to be executed after a delay, a nested secondary task during the delay period may activate these task-rules and drive responses, even when irrelevant (Braem et al., 2019; González-García et al., 2020; Liefooghe et al., 2013).

In one study participants were shown a set of S-R mappings on each trial (e.g. if 'N' press left, if 'P' press right) to be executed following a delay ('inducer task') (Liefooghe et al., 2012). Between the presentation and probe, participants performed a 'diagnostic' task where the same two letters 'N' and 'P' were shown, but in this task participants had to respond to the orientation of the letters while ignoring their identity. Even so, participants responded faster and more accurately for congruent than incongruent diagnostic targets. However, this was not the case when participants merely memorized S-R mappings for later recognition, suggesting task-rule congruency effects are specific to task-sets that are being prepared for execution (Liefooghe et al., 2013).

In this study, I used a similar paradigm to investigate whether parametric stimuli encoded into WM are similarly reconfigured as prospective task-sets, such that a probe quickly triggers the appropriate response at recall. If so, I would expect WM content to drive responses even when irrelevant in a nested task during the WM delay. Alternatively, if WM delay representations are predominately retrospective representations of past sensory input and the full reconfiguration from stimulus to response occurs only after probe onset, I would not expect WM content to drive responses on the nested secondary task.

Participants performed a classic rotation-discrimination WM task ('inducer task'). They encoded two orientations on every trial and after a delay were asked to judge whether a probe was tilted clockwise or counter-clockwise relative to one of the two WM items, indicated by a valid retro-cue during the delay. Between the retro-cue and memory probe, participants were sequentially presented with two, four or six targets for a nested secondary task ('diagnostic task'). Each target was either blue or pink and had a random orientation. Participants indicated the colour of each target while ignoring its orientation. The unpredictable number of targets meant that participants could not anticipate when the memory probe would appear and should be ready to make a speeded

response to the relative tilt of the memory probe with respect to the retro-cued WM item. The response keys to the inducer task (clockwise vs counter-clockwise) and the diagnostic task (blue or pink) were overlapping so each diagnostic target could point to the same response, depending on its colour and relative tilt compared to each WM item (congruent), or could point to two different responses (incongruent). Congruency was assessed either with respect to the cued item in memory (prioritised) or with respect to the uncued WM item.

If prioritised items in WM are stored as prospective task-sets, such that a probe triggers the appropriate response in a reflex-like manner, the diagnostic targets (with irrelevant orientations) may themselves activate these task rules. I compared participants' speed and accuracy in responding to the colour of diagnostic targets when these were either congruent or incongruent with respect to the cued or uncued memory item. To preview the results, I found that responses were faster and more accurate for congruent than incongruent diagnostic targets, specifically when assessed relative to cued, but not uncued WM items, in line with the hypothesis that prioritised WM contents are stored as prospective templates for decision-making or task-sets.

Methods

Participants

Thirty-one participants took part in the study (17 female). Participants' ages ranged from 19 to 27 (M=20.65, SD=1.49). Four participants were left-handed and the remainder were right handed. All participants had normal or corrected-to-normal vision (including colour vision). The study was approved by the Oxford Central University Research Ethics Committee (CUREC) and participants provided informed consent before

taking part in the study. Participants were reimbursed for their time at a rate of £10 per hour.

Procedure

The computerised experimental task consisted of an inducer WM task and a concurrent diagnostic task (Figure 1A). The inducer task was a classic rotation-discrimination WM task requiring participants to encode two oriented bars on each trial. Following a short delay, an arrow indicated which item (left or right) would be probed with 100% reliability (retro-cue). At the end of each trial, participants made a rotation-discrimination judgement about whether the presented probe was tilted clockwise ('J' key) or counter-clockwise ('F' key) relative to the cued item in memory. During the post retro-cue delay period of the inducer task, the diagnostic task was presented. Blue or pink diagnostic targets with random orientations were presented (two, four or six on each trial) and participants indicated the *colour* of each target using the 'F' and 'J' keys (counterbalanced across participants), irrespective of the orientation of the target.

Each trial started with a fixation dot (800-1200 ms) followed by the presentation of the WM items (250 ms). After an initial delay period (750 ms), a 100% informative retro-cue appeared (200 ms) indicating which item would be probed at the end of the trial, followed by another delay (800 ms). Two, four or six diagnostic targets were presented on each trial (counterbalanced within each block of 18 inducer trials). Each target appeared on screen for 100 ms and participants were required to respond within one second. If no response was detected within the response window it was recorded as an incorrect response. The response period was followed by an inter-stimulus-interval (ISI) of 500 ms before the next target would appear. Following the last ISI, the inducer task probe appeared in the centre of the screen (250 ms) and participants were required to respond whether it was tilted clockwise or counter-clockwise relative to the cued memory item. If no response was detected within a response window of two seconds the trial would time-

out and was recorded as incorrect. Participants received feedback on the inducer task accuracy following the response.

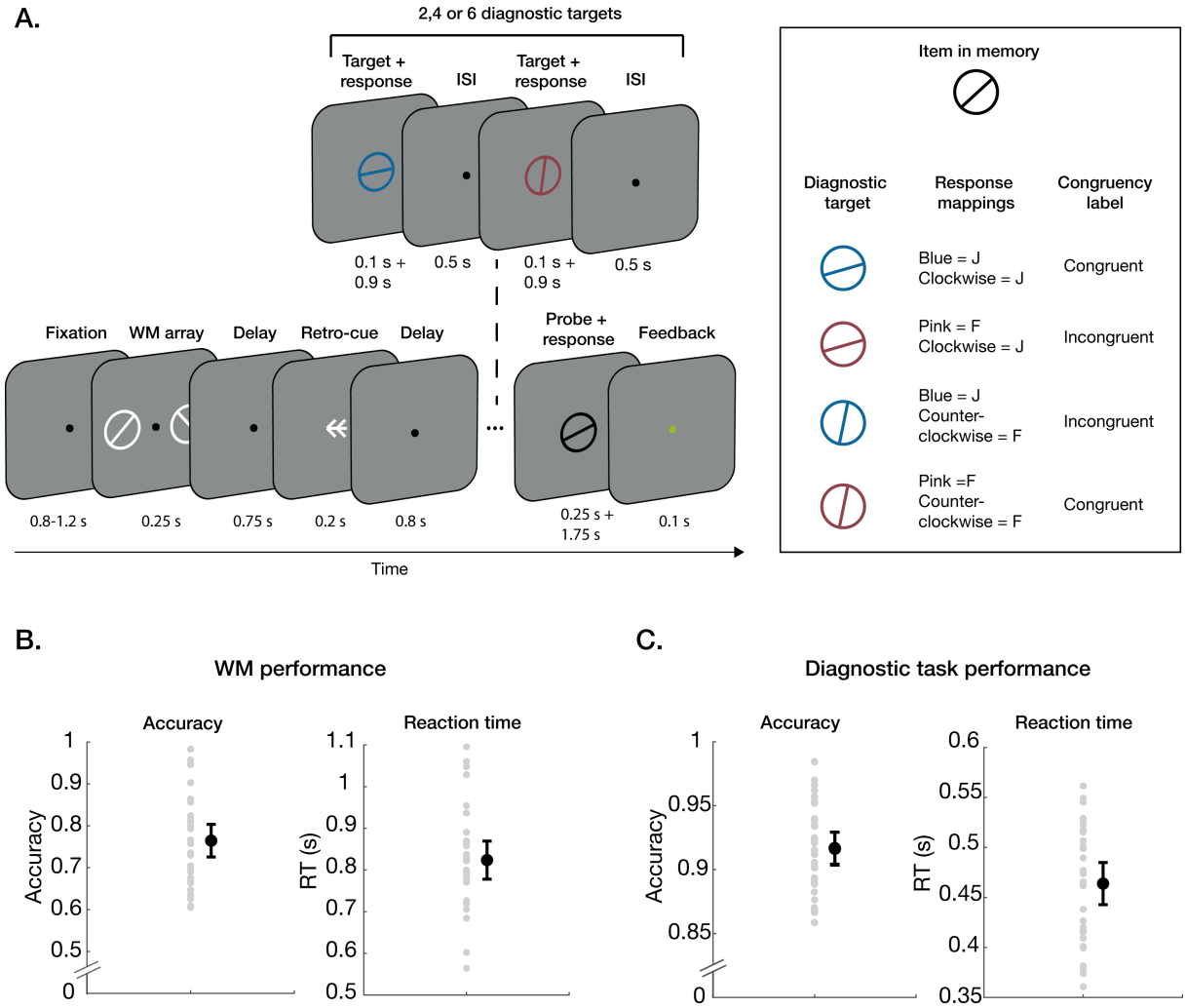

Figure 1. Task design and overall performance for Experiment 1. **A.** Left panel: Participants performed a standard WM task, where they remembered two orientations and after a delay decided if a probe was tilted clockwise ('J' key) or counterclockwise ('F' key) relative to the item cued by the retro-cue (100% valid). During the delay following the retro-cue, participants performed a nested "diagnostic" task where they responded only to the colour of the targets (blue or pink, 'J' or 'F' keys, counterbalanced across participants), while ignoring orientation. Right panel: To quantify whether the diagnostic targets activate the WM task rules, I labelled each target according to its colour and orientation and the stimulus-response mappings with respect to the cued and uncued memory items separately. If the key to respond according to the target's colour was the same as for responding according to its relative orientation, I labelled a target congruent and if stimulus-response mappings conflicted, I labelled it incongruent. **B.** Overall performance on the WM task. Left: accuracy (proportion correct). Right: median reaction times (RT). Grey dots show individual participants, black dot shows mean, and error bars show 95% confidence intervals. **C.** Overall performance on the nested diagnostic task. Otherwise same as B.

Participants performed a mean of 351.84 inducer trials (min=198, max= 450), divided into blocks of 18 trials, corresponding to a mean of 1407.4 diagnostic trials in total

(min=792, max=1800). Individual differences in the requirement for initial practice of the task to threshold left different lengths of time for the actual task and variability in trial numbers across participants. At the end of each block, participants received feedback about their average performance on the inducer and diagnostic task for the latest block and overall. Participants practiced the task before starting the experiment until they achieved a performance threshold of 70% correct on both tasks.

Apparatus and stimuli

The task was programmed and stimuli presented in Matlab R2017b (The MathWorks Inc., 2017) and Psychtoolbox 3.0.1431 (Brainard, 1997b). It was presented on a 24 inch, 1920×1080 pixel monitor running at 100 Hz and participants were seated 60 cm from the screen.

Stimuli were presented on a grey background (RGB = 128, 128, 128). A black fixation dot was presented at the centre of the screen (0.3° visual angle in diameter) throughout each trial. For the inducer task, the memory stimuli were tilted white bars within circles (6°), presented laterally at a distance of 6° from the fixation dot. On every trial, the orientation of each stimulus was selected at random between 1 and 180° with replacement. The retro-cue was a white two-headed arrow (0.5°) pointing either left or right depending of the to-be-probed memory item. Each memory item was cued with equal probability, counterbalanced within each block. The inducer probe was a black tilted bar within a circle (6°, same size memory items), presented centrally. The angular offset between the cued memory item and the probe was +/- 30°, with equal probability, counterbalanced within each block. Feedback for the inducer task was provided at the end of each trial. The fixation dot turned green for a correct response and red for an incorrect response.

Two, four or six diagnostic targets were presented during the delay of each inducer trial. The number of diagnostic targets on each inducer trial was chosen with equal probability, counterbalanced within each block of 18 inducer trials. Diagnostic targets were either blue (RGB = 0, 91, 149) or pink (RGB = 125, 77, 80) tilted bars within circles (6°), presented centrally. The colour of the diagnostic targets was chosen with equal probability, counterbalanced across the full experiment.

Participants responded using the 'F' and 'J keys on a standard (QWERTY) keyboard. For the inducer task, 'J' indicated clockwise and 'F' indicated counter-clockwise. For the diagnostic task, response-mappings were counter-balanced across participants.

Analyses

For the WM task, I calculated the proportion of correct responses and median reaction times for correct WM trials. For the diagnostic task, I calculated the proportion of correct responses on diagnostic targets, after excluding WM error trials. I computed median reaction times, again excluding WM error trials as well as diagnostic targets with incorrect responses and subsequent diagnostic targets due to post-error slowing. I also computed the composite measure of inverse efficiency (IES; RT/accuracy) as previous similar studies suggested this might be more sensitive to detect subtle task-rule congruency effects (Braem et al., 2019; Muhle-Karbe et al., 2017).

If relevant items in WM are stored as prospective templates for decision-making, such that a probe triggers the appropriate response, the diagnostic targets (with irrelevant orientations) may themselves activate these task rules. To quantify whether this was the case, I labelled each diagnostic target as congruent or incongruent depending on its colour and relative orientation compared to the cued and uncued memory item. If the appropriate key for responding according to a target's colour was the same as it would be for responding to the target's irrelevant tilt (clockwise vs counter-clockwise), I labelled a

target congruent, and if they were conflicting, I labelled it incongruent (Figure 1A, right panel). Consequently, congruency is a measure of the extent to which the target's irrelevant orientation is processed in line with the currently irrelevant inducer task rule.

This resulted in four conditions: congruent relative to the cued WM item, incongruent relative to the cued WM item, congruent relative to the uncued WM item, and incongruent relative to the uncued WM item. I performed a 2×2 repeated measures ANOVA with factors cueing (cued vs uncued) and congruency (congruent vs incongruent) separately for accuracy, reaction time and inverse efficiency. Also, I ran two planned paired samples t-tests (one-tailed) to test for a significant congruency benefit separately when defined relative to the cued and uncued WM items. I expected a significant congruency×cueing interaction, with a significant congruency benefit for the cued WM item, but not the uncued WM item.

Next, to explore whether a task-specific memory code, as indexed by the congruency benefit, might facilitate WM task performance, I ran a between-subjects Pearson linear correlation between the congruency benefit (congruent – incongruent) with respect to the cued memory item and WM task performance. For all measures, I expected a positive correlation.

I complement the frequentist statistics with Bayes Factor Analysis, performed using JASP (v 0.9.0.1) with default priors (JASPTeam, 2018). These analyses indicate the likelihood of the data given H_1 relative to the null hypothesis H_0 (reported here as BF_{10}; values > 1 suggest evidence in favour of the H_1, while values <1 suggest evidence in favour of the null). For the Bayesian repeated measures ANOVA analyses interaction terms I report the inclusion Bayes Factor across matched models (BF_{incl}). This compares the models that contain that effect to the equivalent models stripped of this effect by dividing the sum of the P(M | data) of the former by the P(M | data) of the latter.

Results

Good overall performance on both tasks

Figure 1C shows participants' overall performance on the WM task and Figure 1D shows overall performance on the nested diagnostic task. Performance was good on both tasks, with a mean accuracy of .765 (SD=.107) and a mean median RT of .824 (SD=.125) on the WM task and a mean accuracy of .917 (SD=.034) and a mean median RT of .464 (SD=.057) on the nested diagnostic task.

Task rule congruency benefit for prioritised WM content

- *Accuracy*

I performed a Shapiro-Wilk test to confirm the accuracy data were normally distributed (W=.973, p=.591). For accuracy, there was no main effect of cueing condition ($F_{1,30}$=.204, p=.654, BF_{10}=.185) and no main effect of congruency ($F_{1,30}$=3.20, p=.084, BF_{10}= .842). The cueing condition×congruency interaction was also not significant: $F_{1,30}$ =2.88, p=.100, n_p^2=.088, BF_{incl}=2.11, though Bayes Factor indicates weak evidence for this interaction. When diagnostic targets were labelled relative to the cued item in memory, accuracy was significantly higher for congruent (M=.920, SD=.034) than incongruent (M=.913, SD=.036) diagnostic targets (Figure 2A and 2B, left panel): t_{30}=2.22, p=.034, d=.399, BF_{10}=1.61, though Bayes Factor suggests only weak evidence for this effect. This was not the case when targets were labelled relative to the uncued WM item (congruent: M=.915, SD=.036, incongruent: M=.917, SD=.036: t_{30}=-.373, p=.712, BF_{10}=.204) and Bayes Factor suggests evidence in favour of the null hypothesis.

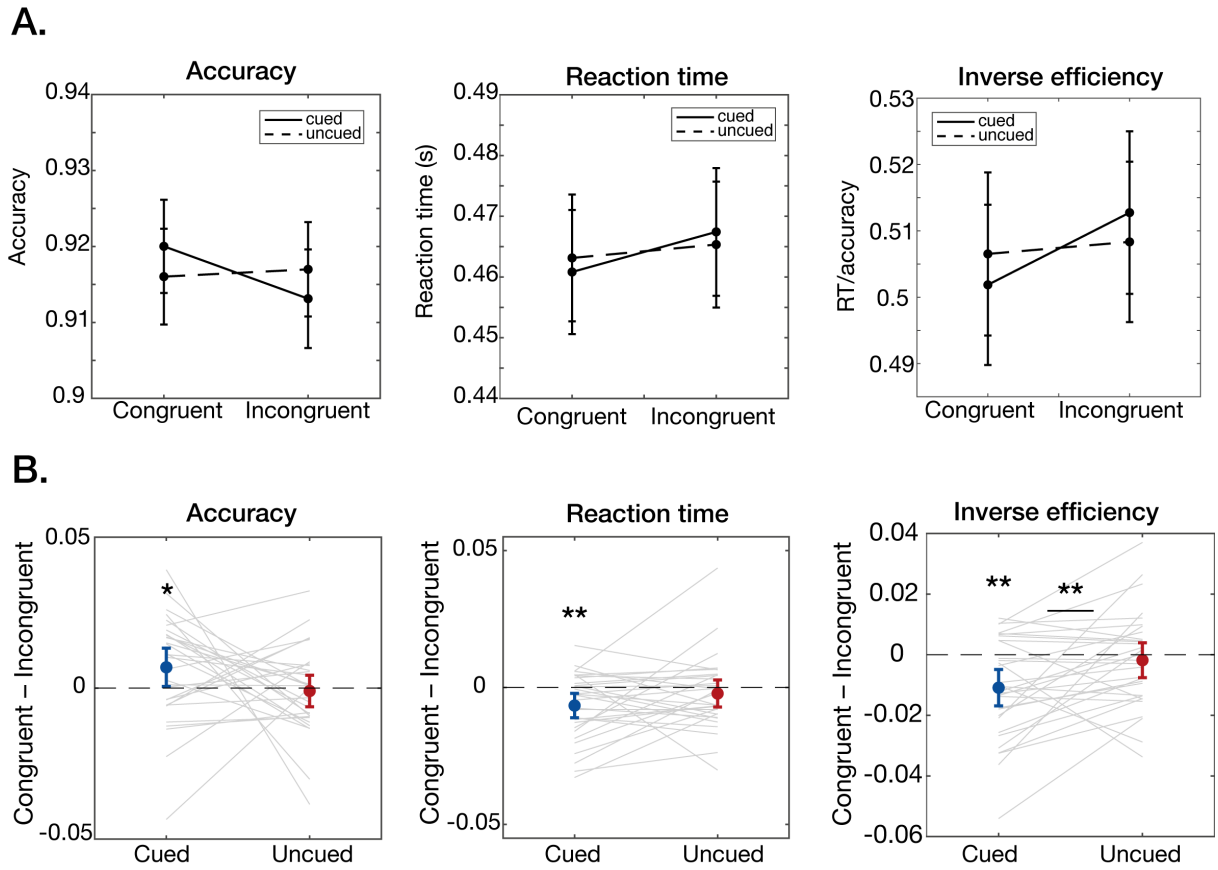

Figure 2. Task-rule congruency effect for diagnostic targets. **A.** Accuracy (left), reaction times (middle) and inverse efficiency (RT/accuracy, right), for congruent and incongruent diagnostic targets when labelled relative to the cued (solid line) and uncued (dashed line) item in memory. Black dots show mean and error bars show standard error of mean. **B.** Congruency benefit (congruent – incongruent) for accuracy (left), reaction times (middle) and inverse efficiency (right), when labelled relative to the cued (blue) and uncued (red) memory items. Dots show mean congruency benefit, error bars show 95% confidence intervals and grey lines show individual participants. *p<.05, **p<.01.

- *Reaction time*

I performed a Shapiro-Wilk test to confirm the RT data were normally distributed (W=.959, p=.279). For RT, there was no significant main effect of cueing condition ($F_{1,30}$=.132, p=.719, BF_{10}=.199). There was a main effect of congruency, suggesting reaction times were faster for congruent than incongruent targets ($F_{1,30}$=5.91, p=.021, n_p^2=.165, BF_{10}=19.05). The cueing condition×congruency interaction was not significant and the inclusion Bayes Factor was inconclusive ($F_{1,30}$=2.40, p=.131, n_p^2=.074, BF_{incl}=.782). As for accuracy, there was a significant congruency benefit relative to the cued item in memory, with faster RTs for congruent (M=.462, SD=.056) than

incongruent (M=.469, SD=.058) diagnostic targets (Figure 2A and 2B, middle panel): t_{30}=-3.04, p=.005, d=-.545, BF_{10}=8.18, but not relative to the uncued WM item (congruent: M=.465, SD=.057, incongruent: M=.467, SD=.057): t_{30}=-.909, p=.370, BF_{10}=.280).

- *Inverse Efficiency*

When combining accuracy and RT into the composite measure of inverse efficiency (RT/accuracy), there was again no main effect of cueing condition ($F_{1,30}$=.120, p=.731, BF_{10}=.188). There was a significant main effect of congruency ($F_{1,30}$=7.10, p=.012, n_p^2=.191, BF_{10}= 63.32) and a significant cueing condition×congruency interaction ($F_{1,30}$=7.91, p=.009, n_p^2=.209, BF_{incl}=5.07). Pairwise comparisons revealed that performance was again significantly better for congruent (M=.504, SD=.067) than incongruent (M=.515, SD=.068) targets relative to the cued item (Figure 2A and 2B, right panel): t_{30}=-3.70, p<.001, d=-.664, BF_{10}=37.12, but not relative to the uncued WM item (congruent: M=.509, SD=.067, incongruent: M=.510, SD=.067; t_{30}=-.645, p=.524, BF_{10}=.232).

No evidence that congruency benefit predicts WM performance

Although the relationship was in the predicted direction, I did not find a significant correlation between the congruency benefit (congruent – incongruent) relative to the cued item in memory and performance on the WM task (Figure 3), for accuracy (r_{29}=.203, p=.273, BF_{10}=.546), RT (r_{29}=-.014, p=.939, BF_{10}=.341), or inverse efficiency (r_{29}=.205, p=.268, BF_{10}=.551).

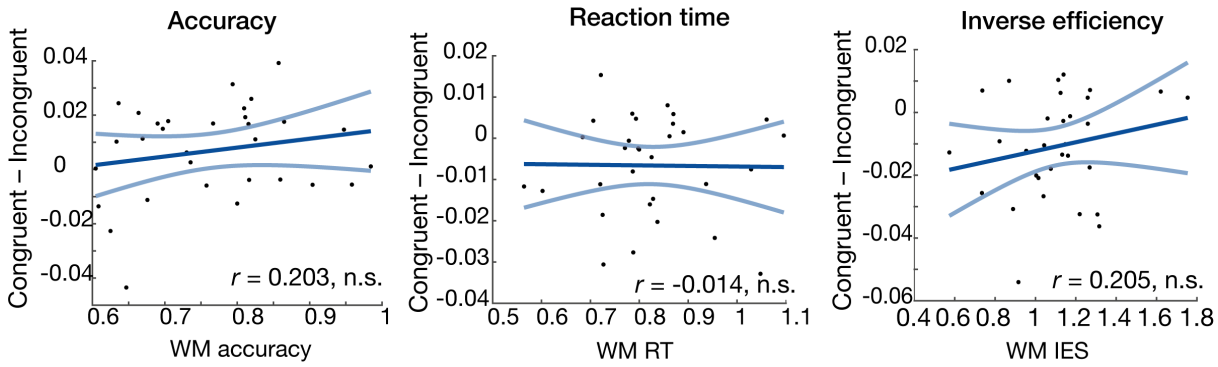

Figure 3. Scatter plots showing correlation between congruency benefit (y-axis) and participants' performance on the WM task (x-axis). Accuracy (left), reaction times (middle) and inverse efficiency (RT/accuracy, right). Black dots show individual participants, dark blue line shows linear regression and light blue lines show 95% confidence intervals.

Interim discussion

In Experiment 1, I investigated whether relevant WM contents are stored as prospective task sets, optimised for quickly computing the relevant decision at recall. In line with my hypothesis, I found that colour judgements on a secondary task during the WM delay were faster and more accurate when S-R mappings were congruent with the cued item held in memory. An exploratory analysis did not provide evidence that the congruency benefit correlated with overall WM task performance, as I would have expected if task-specific coding facilitates WM. However, I note that the direction of the relationship was in the predicted direction for accuracy and inverse efficiency and my power to detect individual differences was small compared to previous studies (Braem et al., 2019).

The results are consistent with a model of WM where prioritised content is reconfigured as task-dependent decision circuits, specific to the current read-out demands. This study serves as a proof-of-principle and it will be necessary to replicate the results with different WM set sizes to test if this effect is specific to a single prioritised item in WM (Myers et al., 2017; Olivers et al., 2011; Olivers & Roelfsema, 2020) and in the context of different WM tasks to test whether the pattern of task-rule congruency effects depends

on the WM read-out demands. In Experiment 2 I aim to test whether task-specific WM coding plays an adaptive role in behaviour.

Experiment 2

If WM is, in part, supported by a re-configuration of maintained contents into prospective task-sets, I would expect WM performance to benefit from having prospective knowledge of the task relative to when the task is uncertain. On the other hand, if WM function is primarily supported by a sensory representation of the stored items, task-foreknowledge should not improve performance, as the full translation from sensory code to motor output should occur following probe onset whether the task is known in advance or not.

To test the hypothesis that advance task knowledge improves WM-guided behaviour, participants were asked to remember the angle of tilted bars for two WM tasks with different read-out demands: a rotation-discrimination task requiring clockwise/counter-clockwise decisions and a change-detection task requiring match/non-match decisions. Task-sets may differ between the two tasks. For example, in the rotation-discrimination task, a task-set may consist of a mental map of the span of target tilts that should trigger a clockwise vs a counter-clockwise response, whereas for the change-detection task, a task-set may consist of a template for acceptance vs rejection of the target. Participants either had advance knowledge of which task they would perform (fixed task blocks) or not (mixed task blocks). If WM content, when possible, is configured into a task-specific code in preparation for the target, I would expect a main effect of task-foreknowledge with more accurate performance on the fixed than on the mixed task blocks.

Previous theoretical and experimental work has suggested only a single WM item, the 'active' or 'attended' memory item, is held in a state that guides attention and drives decision-making (González-García et al., 2020; Olivers et al., 2011; Olivers & Roelfsema, 2020; Peters et al., 2009). If so, a WM representation should be configured in a use-specific format once it is selected as relevant for decision-making, such as following a retro-cue. Myers et al. (2017) proposed a model of retro-cued WM prioritisation that involves two steps: selection and preparation. First, the relevant item in WM is selected by focusing attention to the cued feature. Second, the representation of the relevant item is transformed into an action-oriented code that is optimised to guide behaviour i.e. a stimulus-specific task-set. Preliminary support for this theory of WM prioritisation comes from neuroimaging studies of the retro-cue effect that highlight a role of brain areas often associated with motor planning and task-set preparation, such as the striatum and cingulo-opercular network (Chatham et al., 2014; Myers, Walther, et al., 2015; Schneider et al., 2017; Tamber-Rosenau et al., 2011; Wallis et al., 2015).

If retro-cued prioritisation stems, in part, from the ability to transform a selected item in memory from a stimulus- to an action-oriented representation (Myers et al., 2017), the retro-cue benefit should be larger when the task read-out demands are known in advance. To test this second hypothesis, I included a retro-cue to signal which item would be probed. If task-specific re-configuration supports prioritisation of content in WM, I would expect task-foreknowledge to enhance retro-cued prioritisation, with a stronger retro-cue benefit in fixed task blocks than in mixed task blocks.

In summary, I found that task-foreknowledge improves WM accuracy, suggesting it may be advantageous to configure WM content in a task-oriented code. Retro-cued prioritisation as measured by reaction time, but not accuracy, was enhanced with task-foreknowledge. Prospective task knowledge may allow memory representations of

relevant content to be configured into a task-specific format enabling faster memory-guided decision-making at probe onset, thereby boosting prioritisation.

Methods

Participants

Thirty-five participants took part in the experiment. Five participants performed below chance in one of the experimental conditions and were excluded from subsequent analyses. The remaining thirty participants (18 female) were included in the analyses. They were aged 18-35 (M=26.03, SD=5.01), had normal or corrected-to-normal vision (including normal colour vision), and no history of neurological or neuropsychiatric disorders. 28 were right-handed and two were left-handed. The study was approved by Oxford Central University Research Ethics Committee (CUREC) and all participants provided informed consent prior to participating in the study. Participants were reimbursed for their time.

Procedure

Participants performed a computerised experimental task consisting of two different task types (Figure 4A): a change-detection task requiring match/non-match (M/NM) judgements and a rotation-discrimination task requiring clockwise/counter-clockwise (CW/CCW) judgements. Participants compared the probe to the remembered stimulus according to the current task rule. In the change-detection task, participants responded 'match' if the probe had the same tilt as the memory item and 'non-match' if the probe had a different tilt. In the rotation-discrimination task, participants responded 'clockwise' or 'counter-clockwise' depending on the direction of tilt relative to the memory stimulus. Figure 4B shows the experimental conditions. In half of the blocks, the task context remained fixed for a block of trials (fixed task blocks) and in the other half the

task varied within the block and participants were not informed of the task rule until probe onset (mixed task blocks). Moreover, in half of the blocks, a retro-cue was presented during the delay period to inform participants which of the two items would be probed at the end of the trial.

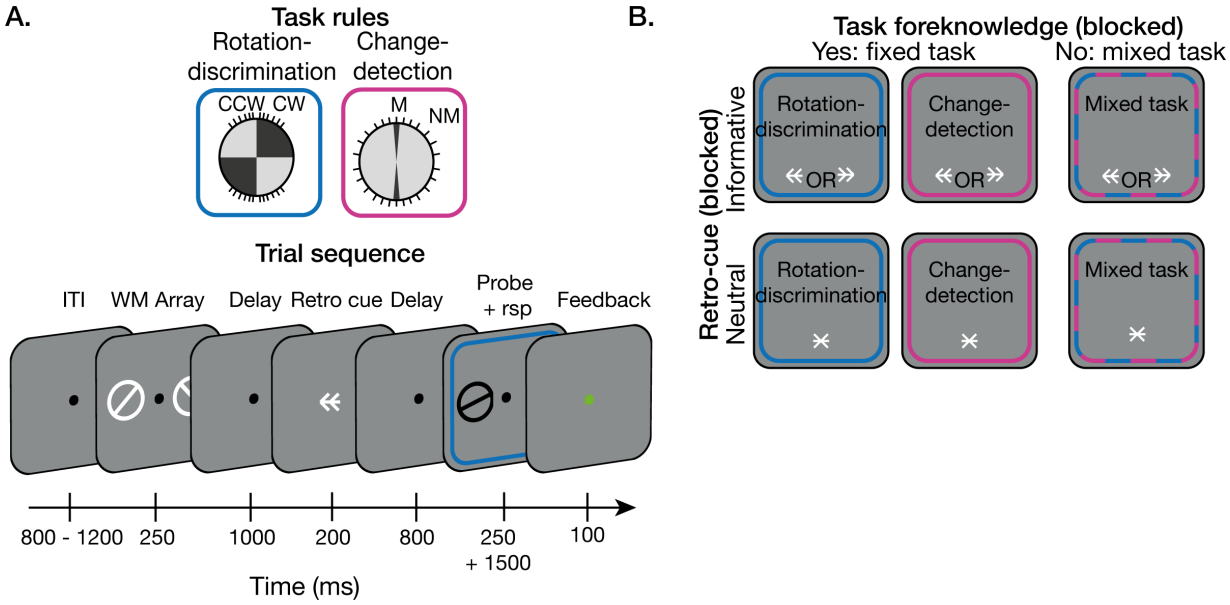

Figure 4. Task design. **A.** Task rules and trial sequence. CW=clockwise, CCW=counter-clockwise, M=match, NM=non-match. Stimuli are not drawn to scale. **B.** Experimental conditions. Task rule (change-detection or rotation discrimination), task-foreknowledge (fixed or mixed task block) and retro-cue (informative or neutral).

Figure 4A shows the trial sequence. Each trial started with a fixation dot (800-1200 ms) followed by presentation of the WM items (250 ms). After an initial delay period (1000 ms), either an informative or a neutral retro-cue appeared (200 ms) depending on the block type, followed by another delay (800 ms). The probe was presented for 250 ms and participants had 1750 ms to respond. The task cue remained on screen throughout the response period. If no response was detected within the response window, the trial would time-out and was recorded as incorrect. Participants responded using the 'F' and 'J' keys on a standard keyboard (QWERTY configuration). For the rotation-discrimination task, participants responded 'F' for 'counter-clockwise' and 'J' for clockwise. For the change-

detection task, the response mappings were counterbalanced across participants. Participants received feedback following the response.

There was a total of 864 trials, divided into 8 block types (36 trials per block): fixed rotation-discrimination with retro-cue, fixed rotation-discrimination without retro-cue, fixed change-detection with retro-cue, fixed change-detection without retro-cue, 2 × mixed task with retro-cue, and 2 × mixed task without retro-cue. The order of the eight block types was pseudorandomised so each block type appeared once every eight blocks, but in random order. Participants practiced the task before starting the experiment.

Apparatus and stimuli

The task was programmed and stimuli presented in Matlab R2017b (The MathWorks Inc., 2017) and Psychtoolbox 3.0.14 (Brainard, 1997b). It was presented on a 24 inch, 1920x1080 pixel monitor running at 100 Hz and participants were seated 60 cm from the screen.

Stimuli were presented on a grey background (RGB = 128, 128, 128). A black fixation dot was presented at the centre of the screen (0.3° diameter) throughout each trial. Memory items were two tilted white bars within circles (6° diameter) and appeared on each side of the screen, at a distance of 6° from the fixation dot. On every trial, the orientation of each stimulus was selected at random between 1 and 180° with replacement. The probe was a black tilted bar within a circle (6°; same size as memory items) and was presented in the position of the to-be-remembered item. The angular offset between the memory item and the probe was uniformly distributed across +/- six angle differences (except for 'match' probes in the change-detection task where the angle difference was always 0°). For the change-detection task, the set of angular offsets were +/-10°, +/-20°, +/-32°, +/-48°, +/-64°, and +/-80°. For the rotation-discrimination task, the angular offsets were +/-5°, +/-10°, +/-16°, +/-24°, +/-32°, and +/-40°. The angular offsets were doubled in the

change-detection task compared to the rotation-discrimination task because angles are compared to 0° (match) rather than their mirror image angle and to match accuracy across tasks. Feedback was provided at the end of each trial. The fixation dot turned green for a correct and red for an incorrect response.

Analyses

I performed a 2×2×2 ANOVA with task-type (change-detection vs rotation-discrimination), task-foreknowledge (present vs absent) and retro-cue (informative vs neutral) as within-subject factors for accuracy (proportion of correct responses for each condition), averaged over the angular offsets between the memorised stimulus and the probe for improved power. See Appendix C, Supplementary Figure S1 for results across angular offsets) and median reaction times (RT) on correct trials. The first question was whether task-foreknowledge improves WM performance. I predicted accuracy would be higher with task-foreknowledge than without. Please note the main effect of task-foreknowledge on RT is confounded by increased stimulus processing demands at probe onset in the mixed than in the fixed task blocks. The second question was whether there was an interaction between the retro-cue benefit and task-foreknowledge. I specifically predicted the retro-cue benefit (accuracy and RT) would be greater with than without task-foreknowledge. I further performed paired-samples t-tests (two-tailed) to test for a significant difference in the size of the retro-cue benefit (informative–neutral) with the presence of foreknowledge or without.

Evidence from task-switching paradigms suggests that performance is better on task-repeat trials relative to task-switch trials (switch-cost)(Monsell, 2003). Next, I tested whether a potential main effect of task-foreknowledge in the above analysis could be accounted for by task repetition. I limited the next analysis to task-repetition trials only and ran a 2×2 repeated measures ANOVA with task-context (change-detection vs

rotation-discrimination) and task-foreknowledge (present vs absent) as within-subject factors.

For completeness, I also tested the role of task-switching on WM performance in general with a 2×2×2 ANOVA on the mixed block trials only with task-context (change-detection vs rotation-discrimination), task-switching (switch vs repeat) and retro-cue (informative vs neutral) as within-subject factors.

As in Experiment 1, I complement the frequentist statistics with Bayes Factor Analysis, performed using JASP (v 0.9.0.1) with default priors (JASPTeam, 2018). Note that there may be discrepancies between frequentist and Bayesian ANOVAs for higher-order interactions (see Chapter 2, Experiment 1 methods section, p. 75).

Results

Performance on the two tasks

Table 1 and 2 contain descriptive statistics for accuracy and RT for all task conditions. Accuracy was generally matched across participants with no main effect of task type (p=.199, BF_{10}=.418). RTs were significantly faster on the change-detection task (M=.702, SD=.105 s) than on the rotation-discrimination task (M=.751, SD=.145 s): $F_{1,29}$=11.64, p=.033, BF_{10}=4.00. This difference was driven by three participants with slow median RTs on the rotation-discrimination task (Appendix C; Supplementary Figure S1A). Thus, the rotation-discrimination and change detection tasks were mostly matched in terms of task performance.

Table 1. Mean and standard deviation (SD) of median RT (s) on correct trials for all conditions.

	Task-foreknowledge with retro-cue		Task-foreknowledge without retro-cue		No task-foreknowledge, with retro-cue		No task-foreknowledge, without retro-cue	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Change-detection	.555	.099	.666	.098	.779	.131	.851	.119
Rotation-discrimination	.617	.163	.737	.138	.796	.164	.889	.141

Table 2. Mean and standard deviation (SD) of accuracy (proportion correct) for all conditions.

	Task-foreknowledge with retro-cue		Task-foreknowledge without retro-cue		No task-foreknowledge, with retro-cue		No task-foreknowledge, without retro-cue	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Change-detection	.839	.067	.768	.063	.801	.065	.744	.071
Rotation-discrimination	.826	.081	.752	.080	.816	.081	.705	.078

Benefit of task-foreknowledge

- *Reaction time*

There was a main effect of task foreknowledge for RT ($F_{1,29}$= 321.82, p<.001, n_p^2=.917, BF_{10}=6.61×10^{30}). RTs were faster with task-foreknowledge (M=.644, SD=.108) than without (M=.829, SD=.123). Please note this main effect is confounded by greater stimulus processing demands at probe onset in the no task-foreknowledge condition and I do not base my conclusions on it.

There was a significant interaction between task-type and task-foreknowledge ($F_{1,29}$=12.18, p=.002, n_p^2=.296, BF_{incl}=1.17). Exploring this interaction further with pairwise comparisons (Bonferroni correction for multiple comparisons, adjusted alpha level=.0125), I found a significant benefit of task foreknowledge in both the change-detection task (t_{29}=-16.81, p<.001) and rotation-discrimination task (t_{29}=-14.74, p<.001). This benefit was bigger for change-detection due to faster RTs on the change-detection

task (M=.604, SD=.099) than on the rotation-discrimination task (M=.671, SD=.150) when participants had foreknowledge about the task (t_{29}=-2.95, p=.006). In the no task-knowledge condition, there was no significant difference in RT for change-detection (M=.812, SD=.119) and rotation-discrimination (M=.841, SD=.150): t_{29}=-1.39, p=.177.

In the task-switching literature, trials following a task-switch are associated with slower RTs and greater errors than trials following task-repetition (Monsell, 2003). As potential switch-costs in the mixed task blocks could also contribute to the main effect of task-foreknowledge, I repeated the previous analysis, but only included task-repeat trials. There was a benefit of task-foreknowledge above and beyond task repetition benefits for RT: $F_{1,29}$=288.54, p<.001, n_p^2=.909, BF_{10}=1.70×10^{15}. Mirroring the effects reported above, there was a trend for a main effect of task type ($F_{1,29}$=4.07, p=.053, n_p^2=.123, BF_{10}=.937) and a significant task-type by task-foreknowledge interaction for task-repeat trials ($F_{1,29}$=13.77, p<.001, n_p^2=.322, BF_{incl}=1.27). Pairwise comparisons showed a significant benefit of task-foreknowledge for both the change-detection task (t_{29}=-15.19, p<.001, d=-2.77, BF_{10}=2.50×10^{12}) and the rotation-discrimination task (t_{29}=-11.78, p<.001, d=-2.15, BF_{10}=5.73×10^{9}) on task-repeat trials (Bonferroni corrected for multiple comparisons, adjusted alpha level=.0125). RTs were faster on the change detection task (M=.604, SD=.099) than on the rotation discrimination task (M=.671, SD= .150) on trials with task-foreknowledge (t_{29}=-2.95, p=.006, d=.539, BF_{10}=6.78), but not on trials without (t_{29}=-0.73, p=.472, d=-.133, BF_{10}=.248).

- *Accuracy*

Figure 5A (left) shows accuracy for trials with or without task foreknowledge. As predicted, accuracy was significantly higher on the fixed task blocks where participants had foreknowledge about the task (M=.796, SD=.059) than on the mixed task blocks where participants did not know the task context until the onset of the probe (M=.767,

SD=.058): $F_{1,29}$=27.34, p<.001, n_p^2=.485, BF_{10}=34.11. There was no significant interaction between task-foreknowledge and task context (p=.812, BF_{incl}=.199).

As for RT, I repeated the previous analysis, but only included task-repeat trials (Figure 5A, right). The main effect of task-foreknowledge remained significant ($F_{1,29}$=11.18, p=.002, n_p^2=.278, BF_{10} = 5.51), showing improved performance with task-foreknowledge on task-repeat trials. As above, there was no significant interaction between task-context and task-foreknowledge for the repeat trials (p=.329, BF_{incl}=.329).

In summary, prospective knowledge of the task-type improved WM accuracy for both the rotation-discrimination task and the change detection task and this benefit was not simply accounted for by task-switching on single trials.

Task-foreknowledge amplifies the retro-cued benefit on reaction time

- *Reaction time*

Replicating previous results, I found that participants performed better with informative retro-cues than with neutral cues across the two tasks. RTs were faster with retro-cues (M=.687, SD=.121) than without (M=.768, SD=.107): $F_{1,29}$=227.25, p<.001, n_p^2=.887, BF_{10}=1.05×10^6. Further, there was a significant interaction between task-foreknowledge and retro-cueing: $F_{1,29}$=11.64, p=.002, n_p^2=.286, BF_{incl}=.653, although the evidence in favour of this interaction compared to matched models without the interaction term was inconclusive as indicated by the inclusion Bayes Factor. There was no interaction between task-type and retro-cueing for RTs ($F_{1,29}$=1.93, p=.176, n_p^2=.062, BF_{incl}=.225) and there was no significant three-way interaction ($F_{1,29}$=.447, p=.509, n_p^2=.015, BF_{incl}=.225).

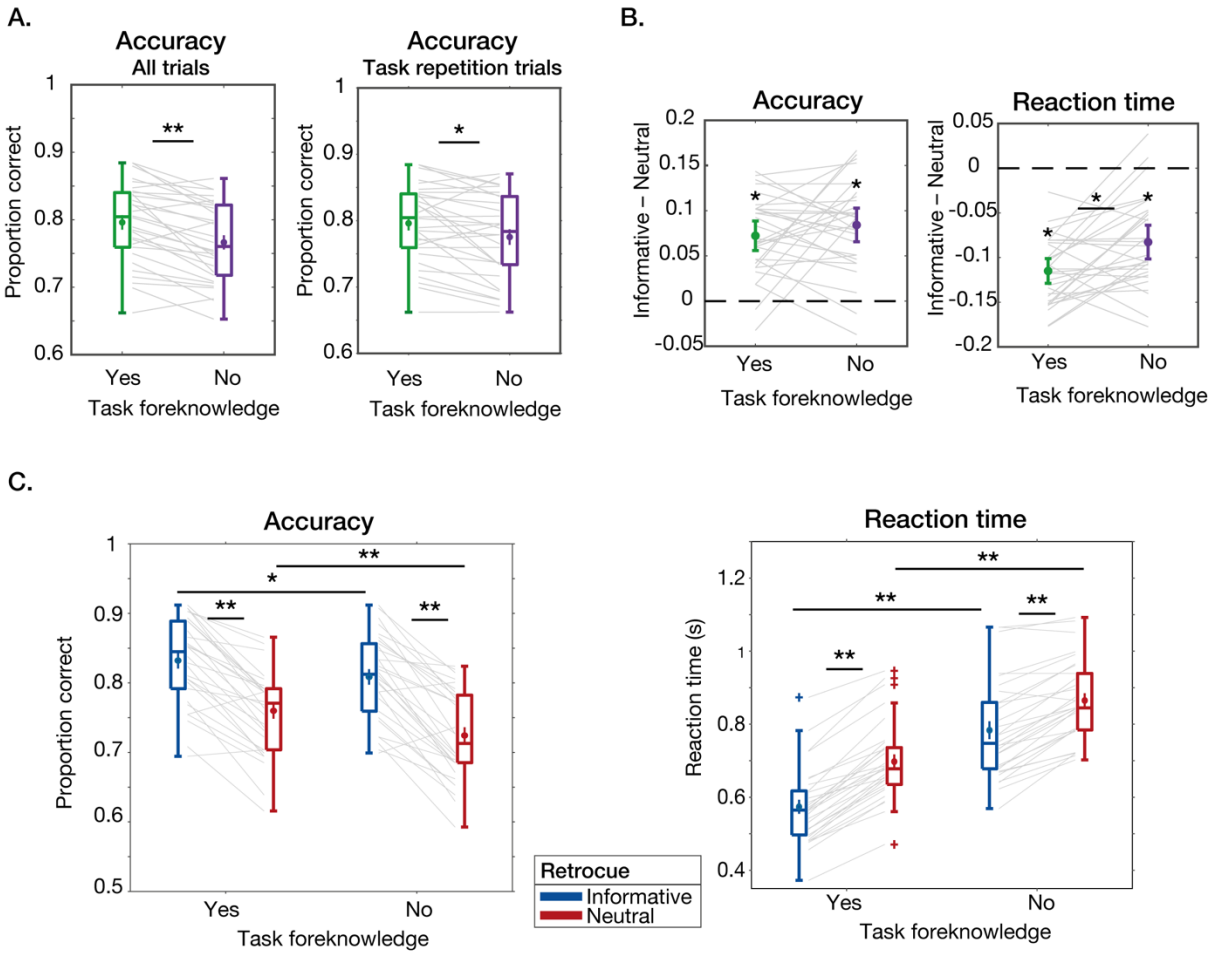

Figure 5. Task-foreknowledge and retro-cueing. **A.** Benefit of task foreknowledge for accuracy for all trials (left) and task-repetition trials only (right). Error bars show standard error of the mean. Grey lines show individual participants. *p<.01, **p<.001. **B.** Retro-cue benefit as a function of task-foreknowledge for accuracy (left) and RT (right). Error bars show 95% confidence intervals. Grey lines show individual participants. **C.** Boxplots of accuracy (left) and RT (right) for trials with and without an informative retro-cue as a function of task-foreknowledge. *p<.05, **p<.001.

To further investigate the significant interaction between task-foreknowledge and retro-cueing, I performed a planned comparison between the RT retro-cue benefit (informative–neutral) for trials with task-foreknowledge and trials without task-foreknowledge (Figure 5B and C). As predicted, the retro-cue benefit was significantly larger with task-foreknowledge (M=-.124, SD=.041) than without (M=-.081, SD=.05): t_{29}=-4.17, p<.001, d=-.762, BF_{10}=114.17, indicating that prospective task knowledge enhances the benefit of retro-cueing on RT. This effect was present for both the change-detection task (t_{29}=-2.90, p=.007, d=-.529, BF_{10}=6.01) and the rotation-discrimination task (t_{29}=-2.06, p=.049, d=-.376, BF_{10}=1.23), though it didn't survive Bonferroni

correction for the rotation-discrimination task (adjusted alpha level=.025) and Bayes Factor was inconclusive.

I further performed four pairwise comparisons (Bonferroni correction for multiple comparisons, adjusted alpha level=.0125) to test for a retro-cue benefit with and without task-foreknowledge, and vice versa. There was a significant retro-cue benefit both when the task was known: t_{29}=-16.69, p<.001, and unknown: t_{29}=-8.98, p<.001. The benefit of task-foreknowledge was present both with (t_{29}=-17.26, p< .001,) and without (t_{29}=-15.61, p< .001) the presence of a retro-cue.

In summary, retro-cues improved RTs and the retro-cue benefit was enhanced in the task-foreknowledge condition relative to no task-foreknowledge condition.

- *Accuracy*

Accuracy was higher with retro-cues (M=.82, SD=.059) than without (M=.742, SD=.059): $F_{1,29}$=140.65, p<.001, n_p^2=.829, BF_{10}=7.70×10^{19}. Figure 5B and C show the benefit of retro-cuing on trials with and without task-foreknowledge. There was no evidence that the retro-cue effect on accuracy benefitted from knowing the task in advance as there was no interaction between retro-cueing and task-foreknowledge ($F_{1,29}$=1.18, p =.286, n_p^2=.039, BF_{incl}=.311). There was also no significant three-way task-type×task-foreknowledge×retro-cue interaction, although there was a small trend for this interaction ($F_{1,29}$=4.05, p=.053, n_p^2=.123, BF_{incl}=1.10).

I performed a planned comparison between the accuracy retro-cue benefit (informative–neutral) for trials with task-foreknowledge and trials without task-foreknowledge (Figure 5B & C). There was no significant difference between the retro-cue benefit with or without task-foreknowledge (t_{29}=-1.09, p=.286, d=-0.198, BF_{10}=.33) and Bayes Factor instead favoured the null. The same was the case when running the comparison separately for the change detection task (t_{29}=0.84, p=.406, d=.154, BF_{10}=.296), while the

rotation-discrimination task showed a small trend towards a reduced retro-cue benefit with task-foreknowledge, though this did not survive Bonferroni correction for multiple comparisons (adjusted alpha level=.025) and Bayes Factor was inconclusive (t_{29}=-2.17, p=.038, d=-0.397, BF_{10}=1.49).

There was a significant task-type×retro-cue interaction for accuracy: $F_{1,29}$=7.68, p<.010, n_p^2= .209, BF_{incl}=1.60, indicating that the retro-cue benefit was larger for the rotation-discrimination task than the change-detection task. Post-hoc pairwise comparisons showed that the retro-cue benefit was significant for both rotation-discrimination (with retro-cue: M=.821, SD=.075, without retro-cue M=.728, SD=.072, t_{29}=4.91, p<.001) and change-detection (with retro-cue: M=.820, SD=.061, without retro-cue: M=.756, SD=.060, t_{29}=3.83, p<.001). There was no significant difference in accuracy for the two tasks either with (p=.551) or without (p=.317) an informative retro-cue.

In summary, retro-cues improved WM performance on both tasks, with some evidence that the retro-cue effect was larger for change-detection than rotation-discrimination judgements. There was no interaction between retro-cued prioritisation and task-foreknowledge for accuracy.

Task switch-cost

There was a main effect of task switching in the mixed task blocks for RTs ($F_{1,29}$=73.69, p<.001, BF_{10}=5.70×10^3) and accuracy ($F_{1,29}$=6.47, p=.017, BF_{10}=.671), suggesting that task-switching has detrimental effects on WM performance in line with switch-costs previously reported in the cognitive control literature. However, it should be noted that the Bayes Factor analysis indicated no evidence for a switching effect for accuracy. There was no switching×task-type interaction, no switching×retro-cue interaction and no three-way task-type×switching×retro-cue interactions for either RT or accuracy (p>.05 and BF_{incl}< 1).

In summary, switching between tasks in the mixed blocks was associated with higher RTs and, to some extent, lower accuracy relative to repeat trials.

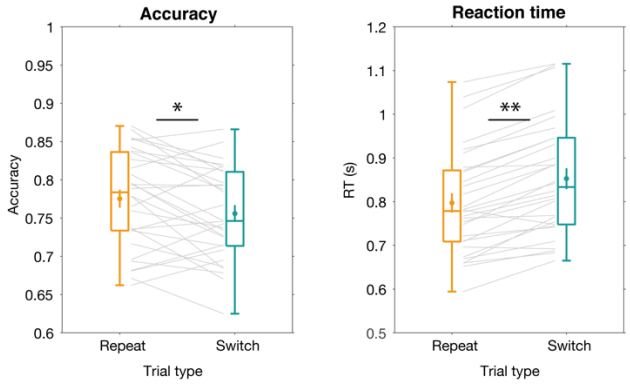

Figure 6. Task switching in mixed task blocks. Accuracy (left) and reaction time (right) for task repeats (orange) and task switch trials (cyan). Error bars show standard error of the mean. Grey lines show individual participants. *p<.05, **p<.001.

Interim discussion

In support of the hypothesis that it is adaptive to store WM content in a prospective task-oriented code, I found that WM accuracy improves with advance knowledge of the task context by which WM content is probed. I further show that accuracy is higher for task-repeats than task-switches in the mixed task blocks, but find evidence that task-foreknowledge benefits performance over and above the effects of task repetition. Retro-cues further improved performance, replicating previous results, and retro-cued prioritisation as measured by RT, but not accuracy, was enhanced with task-foreknowledge, suggesting WM prioritisation may, in part, be supported by re-configuration of WM content from a sensory to a task-oriented code, as predicted by a recent theoretical model (Myers et al., 2017).

While the behavioural data reported here are consistent with the use-specific WM model, they are not definitive. The benefit of prospective task knowledge was assessed relative to a neutral condition without task foreknowledge, making it hard to disentangle benefits due to task-specific WM coding from costs due to additional processing demands

at the time of the probe. Although task-cues were matched on fixed and mixed task blocks, the cue was imperative only in the mixed blocks, incurring additional processing demands. Generally, this would not explain the significant interaction between the retro-cue benefit and task-foreknowledge for RT, as additional probe processing demands should affect the informative and neutral retro-cue conditions similarly. However, because the task used spatial retro-cues that predicted the location of the probe, it remains possible that preparatory attention to the probe-location contributed to the retro-cue benefit. Although the task cue remained on screen for the duration of the response period, perhaps the additional requirements for processing the task cue in mixed task blocks limited the use of anticipatory attentional orienting toward the probe location, thereby reducing the benefit of retro-cues in the mixed task blocks. However, accuracy was still high in the 'no task-foreknowledge, with retro-cue' condition, suggesting participants successfully processed the probe identity on these trials. Future replications preventing the use of preparatory spatial attention as a retro-cueing strategy would strengthen confidence in the results.

Another potential alternative explanation for the results in Experiment 2 is if the mixed task blocks increase the load on WM by requiring two task rules to be maintained in WM simultaneously, rather than a single task rule in the fixed task blocks. This has been proposed as a potential source of the mixing cost in the cognitive control literature (Rogers & Monsell, 1995). If so, fewer resources may be available for remembering orientations with high precision in the mixed task blocks and/or for processing the prioritised item following the retro-cue. This might similarly predict enhanced retro-cued prioritisation with task-foreknowledge as I found for RTs. Whether task rules compete for the same WM resources as those required for WM maintenance is debated (Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008; Oberauer, 2010; Rubin & Meiran, 2005). Rubin and Meiran (2005) suggested that it is the competition between overlapping S-R

mappings and not the number of S-R mappings (i.e. WM load) that is the source of the mixing costs. Nevertheless, future studies varying task-foreknowledge on a trial-wise basis may be better able to distinguish between these possible explanations.

General discussion

Convergent evidence from two behavioural experiments suggest that WM is supported by a transformation of sensory representations into prospective task-sets optimised for context-dependent decision-making.

In Experiment 1, delay representations of parametric stimuli in a classic WM task triggered task-rule congruency effects, a signature of prepared task sets, suggesting WM content is configured as a prospective task-specific decision circuit. I suggest this is analogous to forming a task-set by linking the possible sensory inputs to task-appropriate actions in preparation for the probe (S-R mapping or "prepared reflex")(Hommel, 2000). In Experiment 2, prospective task knowledge improved WM performance, as expected if use-specific WM coding has an adaptive role in guiding behaviour. Additionally, the retro-cue benefit on reaction times was enhanced with task-foreknowledge, consistent with a recent theoretical model of WM prioritisation suggesting that the retro-cue benefit is, in part, supported by re-configuration of WM content into an action-oriented format (Myers et al., 2017). The accuracy retro-cue benefit did not depend on task-foreknowledge, indicating perhaps that task-specific coding of prioritised content enables faster, but not more accurate, translation into decision-signals (and subsequent motor output). However, I note that potential increased WM load in mixed task blocks may also contribute to the pattern of results in Experiment 2 and suggest potential avenues for future studies (see Experiment 2, interim discussion).

The results from Experiment 1 are consistent with previous literature showing task-rule congruency effects in dual-task paradigms (Braem et al., 2019; González-García et al., 2020; Liefooghe et al., 2013, 2012), but also expand on this literature in important ways. Previous studies have shown task-rule congruency effects for explicitly encoded S-R mappings, where a specific probe stimulus should trigger a specific response. Here, it is the relative difference between the orientation of an encoded parametric WM stimulus (orientation) and the random orientations of the diagnostic targets that trigger task-dependent responses. This suggests that the encoded orientation of the WM item is reconfigured from a sensory representation into a format optimised for computing the appropriate decision at recall. I propose that the task-foreknowledge benefit in Experiment 2 may similarly be supported by re-formatting the sensory memoranda in WM into task-sets in preparation for the probe. An important difference between this study and classic task-switching studies from the cognitive control literature (Meiran, 1996; Monsell, 2003; Yeung & Monsell, 2003), is that they tend to instruct explicit S-R mappings that remain stable across many trials, whereas new parametric stimuli are encoded on each trial in the present experiments using classic WM tasks. This would require the formation of new S-R mappings from the parametric WM stimulus on every trial in order to compute the appropriate task-set. Rapid configurations of task sets have previously been reported on the first trial after novel instructed task-sets (Braem et al., 2019; Brass, Liefooghe, Braem, & De Houwer, 2017; Liefooghe et al., 2013, 2012).

Human neuroimaging has shown that 'output gating' in WM (i.e. selecting memorised information to guide future behaviour) engages brain areas traditionally associated with motor planning and task-set preparation (Chatham et al., 2014; Schneider et al., 2017; Tamber-Rosenau et al., 2011; Wallis et al., 2015). Recruitment of these areas may contribute to a task-oriented reconfiguration of the neural state in preparation for upcoming behaviour (Rushworth et al., 2002). Here I showed that task-foreknowledge

boosted the RT retro-cueing benefit, consistent with a model where prioritisation is supported by reconfiguring memorised content into an action-oriented code (Myers et al., 2017; Olivers & Roelfsema, 2020). However, I did not find evidence that task-foreknowledge enhanced retro-cued prioritisation for accuracy, which may suggest task-specific reconfiguration is specifically important for speeded-decision-making, while attentional selection alone supports the accuracy benefit, though further research will be necessary to examine this effect across a wider range of tasks and WM loads.

Studies of WM tend to focus on performance within a single task, for example continuous report tasks (Bays et al., 2009) or change detection tasks (Luck & Vogel, 1997), where subjects always know how the information held in memory will be used. Employing two WM tasks within the same design in Experiment 2 illustrates the importance of considering read-out demands when investigating the nature of WM delay representations. This is novel in the context of WM where the focus has largely been on how the brain maintains detailed representations of past sensory input and less on how WM contents are optimised for guiding future behaviour (Nobre & Stokes, 2019). I do not suggest WM representations are always configured as stimulus-dependent S-R mappings. For example, some free recall tasks may encourage more detailed sensory representations, particularly if they don't require a speeded response. Even so, this would nevertheless be consistent with a model where WM delay representations are optimised for the current task.

In conclusion, relevant WM content triggers task-specific responses, even when irrelevant in a secondary delay task, consistent with a model of WM where prioritised content is reconfigured as task-dependent decision circuits, optimised for the current read-out demands. I further found evidence that advance task knowledge improves WM accuracy, and boosts retro-cued prioritisation, suggesting it may be adaptive to store WM content in a task-oriented code for guiding future behaviour. This highlights the

prospective nature of working memory and challenges the classic notion that WM function is exclusively supported by retrospective representations of past sensory inputs.

Chapter 5. Use-specific working memory: does coding depend on expected future decision-making?[10]

Despite its name, WM is often more concerned with the future than the past, as the core purpose of WM is to guide future behaviour (Nobre & Stokes, 2019; Olivers & Roelfsema, 2020). This future-oriented view of WM predicts that the way information is represented depends on how it will be used. Here, we propose to test the use-dependency of WM with electroencephalography (EEG).

Traditionally, theories of WM have focused on the retrospective mechanisms by which sensory representations are maintained after a stimulus is no longer present in the environment. For example, item-specific persistent activity that bridges the delay between sensory input and motor output in WM tasks has been observed in various areas of the non-human primate brain, including prefrontal, parietal, and sensory areas (Chafee & Goldman-Rakic, 1998; Funahashi et al., 1989; Fuster & Alexander, 1971; Fuster & Jervey, 1982; Gnadt & Andersen, 1988; E. K. Miller et al., 1996; Nakamura & Kubota, 1995). Influential computational models have shown that persistent delay activity can be achieved in networks with recurrent connectivity to sustain mnemonic activity within a stable attractor (Compte et al., 2000; Wimmer et al., 2014). In humans, numerous brain imaging studies have shown that activation patterns in visual and parietal areas encode stimulus-selective representations that bridge the delay period of visual WM tasks (Christophel et al., 2012; Harrison & Tong, 2009; Serences, 2016; Serences et al., 2009).

[10] This chapter is based on a manuscript prepared as a Stage 1 Registered Report not yet accepted for publication: Printzlau, F., Wolff, M., Muhle-Karbe, P., Nobre, A.C., Manohar, S., Myers, N.*, Stokes, M*. Use-specific working memory: does coding depend on expected future decision-making? *Joint contribution.

However, it has long been appreciated that WM is more than a passive store of past experience (Geng et al., 2017; Navalpakkam & Itti, 2007; Scolari et al., 2012; Scolari & Serences, 2009; X. Yu & Geng, 2019). Extensive behavioural and neuroimaging evidence suggests that WM preferentially encodes information that is most likely to be relevant for future behaviour (A. Gazzaley & Nobre, 2012; Adam Gazzaley, 2011; Griffin & Nobre, 2003; Alexandra M. Murray, Nobre, & Stokes, 2011; Vogel & Machizawa, 2004; Vogel et al., 2005; Woodman & Vogel, 2008), thereby improving the odds of achieving a behavioural goal despite WM capacity limits (Luck & Vogel, 1997). Selective encoding of task-relevant objects, as measured by contralateral delay activity in EEG, correlates with WM capacity, pointing to an adaptive role of filtering access to WM (Vogel et al., 2005). Selection based on task demands can also occur at the level of object features. Serences and colleagues (2009) showed that when participants were instructed to memorise either the colour or the orientation of a single stimulus, only the relevant feature dimension could be accurately classified from hemodynamic responses in visual cortex during the delay (Serences et al., 2009). Similarly, Lee and colleagues (2013) showed that the same item was represented in posterior fusiform areas when participants were required to memorise precise visual features, but in lateral prefrontal areas when the task was to memorise more abstract category-level information (Lee et al., 2013). Selective coding of task-relevant information does not necessarily require the presentation of the relevant stimulus. Rainer and colleagues (1999) trained monkeys to perform a delayed paired-associate task, in which each sample image was associated with a different target image (Rainer et al., 1999). Delay activity coded for the expected target stimulus, even though it was not actually presented at encoding. It seems that what is represented is the information relevant for the expected future decision, rather than the presented stimulus itself.

Together, the studies above suggest that expectations about future relevance guide which information is encoded and maintained in WM. Various theoretical models account for selective maintenance by filtering access into WM (Bundesen, 1990; Bundesen et al., 2005, 2015; O'Reilly & Frank, 2006). Recently we have proposed a stronger prediction for future-oriented WM, whereby expectations about future use dictate not only *what* information is encoded, but also *how* it is coded in WM. According to the dynamic coding framework, encoding into WM changes the way neural circuits involved in memory and decision-making respond to incoming information (Stokes, 2015). Empirically, this is supported by a qualitative transition between coding schemes during encoding and delay phases of a WM task (Sreenivasan, Vytlacil, et al., 2014; Wolff et al., 2017). The changed neural pattern in the delay phase could reflect a transition to a specific response profile of the network optimised for behaviour by specifically mediating decision-relevant context-dependent responses to subsequent input. For example, neural states of WM could function as a matched filter, which directly interacts with sensory input to compute the appropriate decision: In a change detection task, a WM item may serve as a template to compute the distance between sample and probe stimuli (e.g. match vs mismatch), whereas in a discrimination task the WM item may serve as a template to compute the rotation direction (e.g. clockwise vs anti-clockwise) (Hayden & Gallant, 2013; Merrikhi et al., 2017; Myers, Rohenkohl, et al., 2015; Sugase-Miyamoto et al., 2008). Within this framework, the neural state of WM is not just a representation of past sensory input, but is a prospective task-dependent decision-circuit that functionally serves as the basis for WM-guided decision-making (González-García et al., 2020; Muhle-Karbe et al., 2017, 2020). From this functional perspective of WM coding, we propose to test the following novel hypothesis: the *same* stimulus feature is represented differently depending on expected use.

Initial evidence consistent with this idea comes from studies that vary the momentary relevance of memory items for guiding behaviour. Items can be prioritized or not depending on how relevant they are to expected decision-making. For example, Warden and Miller (2010) showed that neural representations in monkey prefrontal cortex preferentially coded for one of two images in memory depending on the expected task (recognition vs recall), even though the sequence of events was identical in the two tasks until the onset of the probe. In humans, selecting an item in memory for a forthcoming decision is associated with a boost in the stimulus-specific neural representation, while decoding accuracy drops for other items that might be relevant later (LaRocque et al., 2013; Lewis-Peacock et al., 2012; Watanabe & Funahashi, 2014b; Wolff et al., 2017). At first glance, this appears similar to attentional enhancement of relevant features in WM. However, attentional enhancement in perception comes at the expense of other items, while prioritization in WM sometimes occurs without associated behavioural costs (Myers et al., 2018; Rerko & Oberauer, 2013; Van Moorselaar et al., 2015). Neurally, recent evidence suggests that prioritized and unprioritized items are actually both represented robustly, just in qualitatively different formats. Unprioritized items could be stored in hidden or latent neural states (e.g. temporary changes in synaptic strength) (Mongillo et al., 2008; Rose et al., 2016; Wolff et al., 2015, 2017), in different brain areas (Christophel et al., 2018) or in an opposite (or "reversed") neural code within the same areas of cortex (van Loon et al., 2018; Q. Yu et al., 2020).

This evidence for priority-dependent maintenance of WM items is consistent with a general model of use-dependency. While classic attention-based accounts assume that prioritization reflects amplification of the selected WM representation irrespective of task-context (Larocque et al., 2014; Lepsien et al., 2011), a recent account instead suggests that WM prioritization involves a transformation of the selected representation into a task-dependent decision circuit optimised for the upcoming decision (González-

García et al., 2020; Muhle-Karbe et al., 2020; Myers et al., 2017). If prioritized WM content is stored in a format that directly mediates future decision-making (Olivers et al., 2011), currently unprioritized WM items should be stored in a qualitatively different format that does not interfere with the current decision, as recent studies have found (Christophel et al., 2018; Rose et al., 2016; van Loon et al., 2018; Wolff et al., 2015, 2017; Q. Yu et al., 2020). The model proposed here makes an even stronger prediction: even when the priority-state is held constant, the neural state should differ according to the type of expected decision (e.g., change-detection vs discrimination). We aim to test this strong variant of use-specific WM. Either outcome would be informative in characterising the nature of WM delay representations and advance the ongoing debate in the literature regarding how WM content is prioritized for action.

In this proposed study, we will test whether coding in WM depends on how the information will be used during readout. Storing WM content in a use-specific format could serve an adaptive purpose by preparing the memory item for context-dependent processing, thereby allowing for faster and more accurate translation to the relevant decision signals (and eventually motor output) once a memory probe appears (Monsell, 2003; Sakai, 2008). We previously showed that prospective knowledge of the required type of decision (rotation discrimination vs change detection) improves WM performance, suggesting a benefit of use-specific working memory coding (Chapter 4; Experiment 2)(Printzlau et al., 2019). However, behavioural studies alone face a fundamental challenge when trying to assess the relative contribution of use-oriented WM coding to performance. The benefit of prospective task knowledge is assessed relative to a neutral condition without task foreknowledge, making it difficult to disentangle benefits due to use-specific WM coding from costs due to additional processing demands at recall. While the behavioural data are consistent with the use-specific WM model, using EEG will

provide a more direct test of the hypothesis by allowing us to directly compare delay representations in different task contexts.

To test whether the same WM information is maintained in use-specific configurations, we propose an EEG experiment in which participants remember oriented bars for two different types of WM-based decisions: 1) a *rotation-discrimination* task requiring participants to make clockwise/counter-clockwise judgements, or 2) a *change-detection* task requiring match/mismatch judgements. Both types of WM tasks are widely used in the literature and require high precision representations of the stimulus orientation. Participants will encode two items on every trial and a retro-cue will inform them which item is relevant to their decision with 100% reliability. M/EEG has previously been shown to be sensitive to decode fine-grained mental representations, such as memorised orientations (Carlson, Grootswagers, & Robinson, 2019; Cichy, Ramirez, & Pantazis, 2015; Grootswagers, Wardle, & Carlson, 2017; Stokes, Wolff, & Spaak, 2015; Wolff et al., 2017). If the neural code for WM is use-specific, we would expect orientation decoding to be impaired when training a classifier on data from one task and testing on data from the other task relative to decoding accuracy when training and testing within the same task (cross-generalization cost). Alternatively, if the neural code for WM items does not depend on the expected use there should be no cost of cross-generalization.

Our primary question concerns whether the retro-cued item, prioritized for the upcoming decision, is configured differently depending on the anticipated task. We expect to find a significant cost of cross-generalization for the cued item. If this hypothesis is confirmed, it would suggest that the same relevant information is configured differently depending on anticipated decision-making. Alternatively, a null-effect showing no cost of cross-generalization would require revision of this strong version of the use-specific WM hypothesis and suggest instead that coding of prioritized WM content is task-invariant.

Secondly, if the prioritized item is configured in a use-specific format, the question remains when this configuration occurs. While it might be advantageous to prepare a use-oriented format for a single item about to be probed, what about multiple potentially relevant items? Previous theoretical work has suggested only a single WM item can be held in a decision-oriented state (Olivers et al., 2011; Peters et al., 2009). In our study, this would mean that WM content should only be reformatted into a use-specific code once the retro-cue indicates which item will be probed (Myers et al., 2017). To test whether the WM items are in a use-specific format from encoding, we will also compute the cross-generalization cost following the onset of the WM array. A significant cost of cross-generalization would suggest that the items are configured in a use-specific format from encoding, while a null effect would indicate that items are initially encoded in a format that is not specific to the task. Either outcome is compatible with the primary hypothesis, but distinguishing between these two scenarios will help elucidate whether use-specific coding evolves specifically for WM content that has been selected as currently relevant (Myers et al., 2017; Olivers et al., 2011) or is a more general property of coding in WM.

Table 1. Hypotheses. Summary of primary and secondary research questions, how the hypotheses will be tested and how outcomes will be interpreted.

Primary question: Is prioritized WM content configured in a use-specific format? Test: cost of cross-generalization following retro-cued prioritization		
H_1: cost of cross-generalization Evidence in support of use-specific WM account		**H_0: no cost of cross-generalization** Evidence against use-specific WM account
Secondary question: Is content encoded into WM in a use-specific format? Test: cost of cross-generalization during encoding into WM		
H_1: cost of cross-generalization Content is encoded into WM in a use-specific code	**H_0: no cost of cross-generalization** Use-dependency evolves, such as following retro-cued prioritization	

Methods

Participants

We will use a sequential Bayesian analysis design with a minimum sample size of n=30 and a maximum sample size of n=113 (see below for Bayesian sampling methods). To be eligible, participants must be aged 18-40, have normal or corrected-to-normal vision and have no history of neurological or neuropsychiatric disorders. Participants will first come in for a practice session to ensure they are comfortable performing the task while maintaining fixation. They will then be invited to complete the main experiment in a separate session. The study has been approved by Oxford Central University Research Ethics Committee (CUREC) and all participants will provide informed consent prior to participating in the study. Participants will be reimbursed £15 per hour for their time.

Procedure

Participants perform a computerised experimental task (Figure 1). It consists of two different task contexts: a change-detection task requiring match/mismatch decisions and a rotation-discrimination task requiring clockwise/counter-clockwise decisions. The task context is blocked and participants are informed of the task rule prior to each block. The stimulus parameters during each trial are identical in both tasks until probe onset. Participants compare the probe to the remembered stimulus according to the current task rule. In the change-detection task, participants respond 'match' if the probe has the same tilt and 'mismatch' if the probe has a different tilt. In the rotation-discrimination task, participants respond 'clockwise' or 'counter-clockwise' depending on the direction of the tilt relative to the stimulus. Participants are encouraged to respond as quickly and accurately as possible.

Each trial starts with a fixation marker (300 ms) followed by presentation of the WM items (250 ms). Participants encode both items. After a short initial delay (350 ms), a

retro-cue (100 ms) informs participants which item will be probed. Following a second delay (900 ms), an impulse stimulus is flashed for 100 ms. A strong visual impulse can be used to boost decodability of WM representations (Stokes, 2015; Wolff et al., 2015, 2017). A central dot re-appears inside the fixation marker 300 ms before impulse onset to re-centre fixation. The impulse is followed by a short delay (300 ms). The probe is presented for 250 ms and participants have 1500 ms to respond. Participants receive feedback following the response for 100 ms. The inter-trial interval is jittered between 400-800 ms. In case participants forget the condition during a trial block (32 trials), there is an option to pause the task and review block conditions. At the end of each block, participants are informed of their accuracy and median reaction times on the most recent block and overall.

Participants perform a total of 2240 trials, 1120 trials per task type, divided into 70 blocks. Participants will have the opportunity to take breaks between trial blocks. The experiment can be completed either in one session or divided into two experimental sessions depending on participants' preference. The order of the two block types (change-detection, rotation-discrimination) is pseudorandomised so the block types are equally distributed throughout the session. Participants practice the task in a separate practice session until they are comfortable maintaining fixation and perform above a minimum performance criterion of 80% accuracy on both tasks and on items presented on either side of fixation.

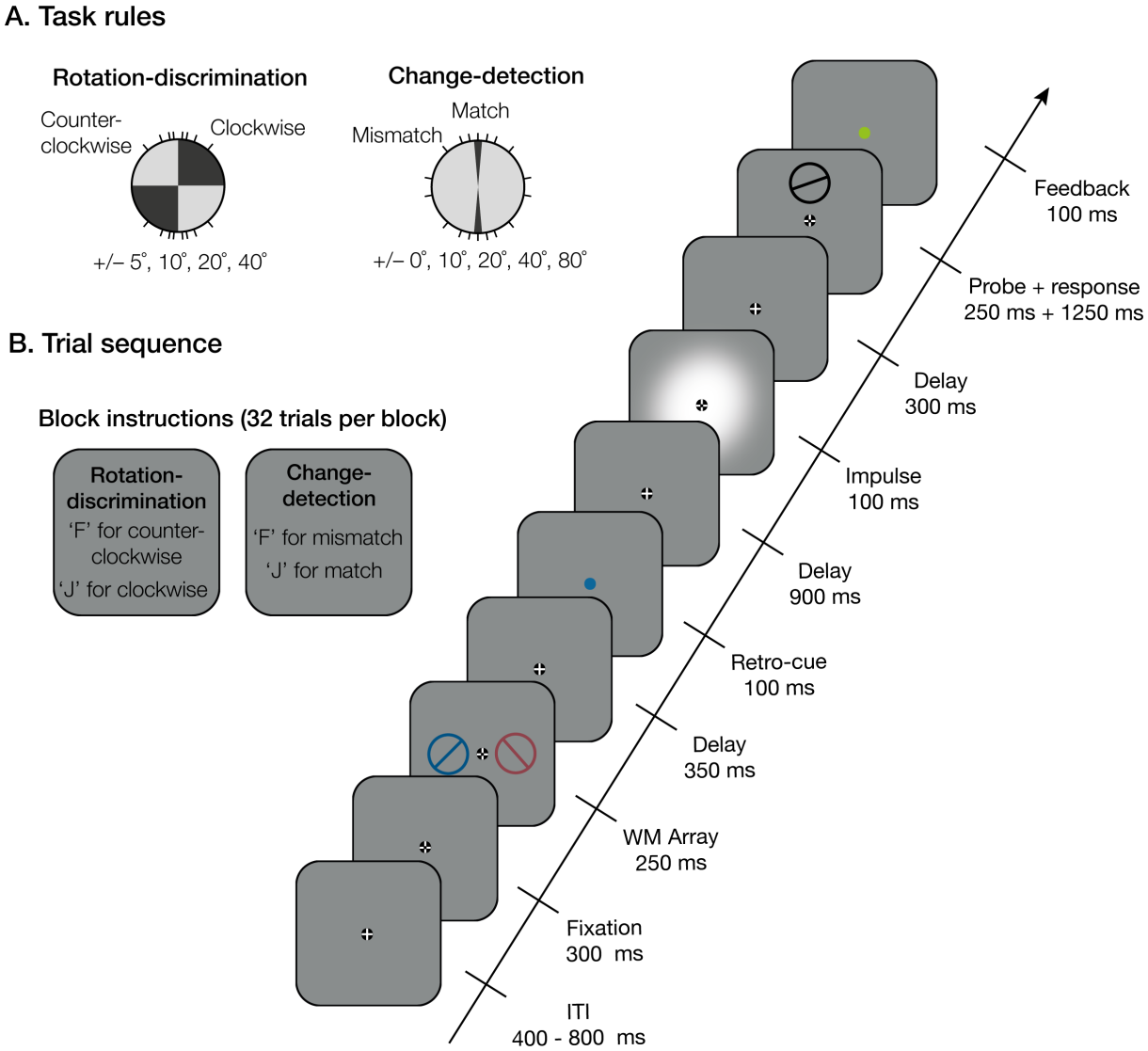

Figure 1. Task design and experimental conditions. **A.** Task rules for rotation-discrimination task and change-detection task. The ticks around the circles represent the possible angular offsets of the probe and shaded regions represent decision boundaries (relative to vertical meridian) for each task. **B.** Trial sequence. Before each block of 32 trials participants receive instructions of the task context and response mappings for the upcoming block.

Apparatus and stimuli

The task is programmed and stimuli presented in Matlab (The MathWorks Inc., 2017) and Psychtoolbox (Brainard, 1997a). The task is presented on a 24-inch, 1920x1080 pixel monitor running at 100 Hz.

Stimuli are presented on a grey background (RGB = 128, 128, 128, CIELAB: 53.6, 0, 0). A fixation marker is presented at the centre of the screen throughout the trial. It

consists of a filled black circle (0.5° diameter) with a white cross inside (width: 0.15°). To signal trial onset and to encourage fixation, a black dot (0.15° diameter) appears inside the fixation marker 300 ms before each event of interest (i.e. the memory array and the impulse). This type of fixation marker has been shown to reduce micro-saccades (Thaler, Schütz, Goodale, & Gegenfurtner, 2013). Memory items are two tilted bars within circles (6° diameter) and appear on each side of the screen 6° from the fixation marker. One item is blue (RGB = 0, 91, 149, CIELAB: 30, -50, -50) and one item is pink (RGB = 125, 37, 80, CIELAB: 30, 42, -5), selected to have similar luminance. The prioritized item is presented left or right of the fixation marker on each trial, selected at random from a uniform distribution within each block. On every trial, the orientation of each stimulus is randomly selected without replacement from a uniform distribution of orientations matching the number of trials in each block (32 angles, 2.81-177.19°, in steps of 5.625°). The stimulus angles are the same for both task contexts. The impulse stimulus is a circular white stimulus (18° diameter), with smooth borders, presented centrally, that covers the area of stimulus and probe presentation, created using the CreateProceduralSmoothDisc function in Psychtoolbox with radius = 9° and edge smoothing value = 1.3°. The probe is a black tilted bar within a circle (6°; same size as memory items) and is presented above or below the fixation marker along the vertical meridian (6° eccentricity) on each trial, with equal probability, pseudorandomised for the full experimental session. The purpose of the specific positioning is twofold. First, the probe is not reproduced at the original location, or in the original colour, making it imperative that participants use the retro cue. The retro cue is the only information that determines which item will be probed, therefore accurate performance on the task implies that participants are directing attention to the relevant, retro-cued, item. Second, we wanted to avoid presenting the probes at fixation, as this could encourage orientation-specific micro-saccades (Mostert et al., 2018). The angular offset between the

memory items and the probes are uniformly distributed across four angular differences (except for 'match' probes in the change-detection task where the angle difference is always 0°). For the change-detection task, the set of angular offsets are +/–10°, +/-20°, +/–40°, +/-80°. For the rotation-discrimination task, the angular offsets are +/–5°, +/–10°, +/–20, +/–40°. In a recent behavioural study, we show that accuracy between the two tasks is matched with a 2:1 relationship of angular offsets between rotation-discrimination and change-detection (Chapter 4; Experiment 2)(Printzlau et al., 2019). Feedback is provided at the end of each trial. The fixation marker turns green for a correct response and red for an incorrect response.

Participants respond using the 'F' and 'J' keys on a standard keyboard. For the change-detection task, participants respond 'F' for 'mismatch' and 'J' for 'match'. For the rotation-discrimination task, participants respond 'F' for 'counter-clockwise' and 'J' for 'clockwise'.

EEG data acquisition

Electroencephalography (EEG) will be recorded at a sampling rate of 1kHz using Synamps amplifiers (Neuroscan, Inc., Albany, CA) and Curry Neuroimaging Suite (v7.0.12 X) software. Sixty-one electrodes will be positioned according to the 10-10 system("American Electroencephalographic Society guidelines for standard electrode position nomenclature.," 1991). Eye movements are recorded from bipolar EOG derivations from electrodes above and below the right eye (vertical) and lateral to each eye (horizontal). A ground electrode will be placed on the left elbow and the left mastoid will be used as the active reference.

In addition, participants' eye movements will be recorded using an eye tracking camera (EyeLink) tracking at least one eye, and both eyes where possible, at a rate of 1kHz.

Planned analyses

Behavioural analyses

Mean accuracy and median response times calculated for correct trials (RT) will be analysed in a Bayesian paired samples t-test (two-tailed, default prior) comparing performance on rotation-discrimination blocks to performance on change-detection blocks (H_1: difference in performance, H_0: no difference in performance). While we do not expect performance to differ between the two tasks, evidence for a difference will not impair our ability to test our hypotheses relating to the underlying neural coding format for stimulus orientation. Potential differences in overall neural decoding between the two tasks is addressed below (see *Cross-generalization methods).*

EEG pre-processing

The EEG signal will be re-referenced offline to the mean of the right and left mastoid and down-sampled to 500 Hz. We will apply a high-pass filter of .1 Hz and a low-pass filter of 45 Hz.

The data will be epoched around the onset of the WM array and around the onset of the impulse stimulus. The epochs will be 0-400 ms following the WM array onset and 0-400 ms following the impulse onset. Artefact rejection and data analysis will be performed separately for each epoch.

We will use the eye-tracking data (Eyelink) to reject epochs with artefacts related to eye-movements. We will reject epochs with eye-blinks (i.e. missing eye-tracking data) and any epochs with fixation deviations > 1° visual angle. We will apply independent components analysis (ICA) to remove any heartbeat artefacts and/or artefacts due to eye-blinks that remain after rejecting eye-blink epochs. For each participant, we will also visually inspect the FieldTrip summary plot displaying the variance over trials to exclude artefacts due to large muscle movements or missing electrodes (Oostenveld et

al., 2011). In addition, any epochs with large voltage steps (> 50 μV) in any electrode will be rejected.

Any participants with fewer than 80% of epochs remaining following artefact rejection will be excluded from further analyses. Based on a previous similar experiment used for computing power analyses for the current study, we expect to reject about five percent of trials on average per participant (Wolff et al., 2017). We will exclude participants that perform below a threshold of 80% correct on either task and on items presented on either side of fixation, left or right, from further analyses. We will not exclude behavioural error trials, as this may result in biases in the distribution of trials for the multivariate decoding analyses. A large proportion of the error rate will likely be driven by the difficulty of the probe and will not necessarily be an indicator that people did not remember the correct stimulus on that trial.

Multivariate decoding methods

To test the extent to which activity patterns in the EEG channels contain information about the orientations of the memory items, we will calculate Mahalanobis distance to estimate the trial-wise multidimensional distances between the activity patterns at each time point for each of the 32 possible orientation angles in memory, grouped into 16 angle bins distributed around 180 degrees (Wolff et al., 2017). As in previous published studies, we will perform the analyses on the 17 posterior channels (P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz and O2) (Muhle-Karbe et al., 2020; Wolff, Jochim, et al., 2020; Wolff et al., 2017). The left and right memory items will be decoded separately and independently for each participant and experimental session.

To take advantage of the dynamic nature of event-related potentials (ERPs), we will use a multivariate decoding approach that pools the relative voltage changes over time and space and thereby leverages information encoded in spatial as well as temporal patterns.

This approach has been shown to boost decoding accuracy (Grootswagers et al., 2017; Nemrodov, Niemeier, Patel, & Nestor, 2018; Wolff, Jochim, et al., 2020; Wolff, Kandemir, Stokes, & Akyürek, 2020a). We will specifically follow the relevant methods in the previous paper (Wolff, Jochim, et al., 2020). We will focus our decoding analysis on a time window from 100-400 ms following onset of the event of interest (impulse and WM array). Another advantage of this approach is that it reduces the multiple-comparisons problem: with a pre-specified time-window we can compare conditions with a single test, rather than with cluster-correction, which improves our sensitivity to detect memory signals. The data will be down-sampled to 100 Hz. We will zero-centre the voltage fluctuations by subtracting the mean activity of the full time window from the data, separately for each trial and channel. This relative baseline is similar to pre-stimulus baseline correction as is common in ERP analyses, but also removes neural activity that is stable within the time-window. As such, this approach is optimised for decoding stimulus-evoked, dynamic activity. The resulting 30 values for each channel will be concatenated and used as the input to the multivariate decoder.

The decoding procedure follows a leave-one-block-out cross-validation approach to calculate the decoding accuracy of the orientation of interest for each trial. As Mahalanobis distances are sensitive to differences in means, decoding will be performed on the residual data after subtracting task-specific mean activity from each trial belonging to the same task. The pattern of activity of the trials in the testing block is compared to the trials in the remaining blocks serving as training data. Each block serves as test data once. The trials of the training data are averaged into 16 orientation bins relative to the orientation of the test trial, each bin containing trials with orientations within a range of 11.25° (variable bins, see below). We will ensure equal trial numbers in each bin by random subsampling. We will compute the pairwise Mahalanobis distances between the test trials and each of the orientation bins using the

noise covariance matrix estimated from trials in training data, using a shrinkage estimator (Ledoit & Wolf, 2004). Noise covariance will be computed on the residual data after removing orientation-specific mean activity from each trial belonging to the same orientation bin. To construct a visual representation of a tuning curve, the 16 distances will be ordered as a function of orientation difference, mean-centred, and sign-reversed, so higher values represent greater similarity between the test trial and the training set. This will be repeated for all train and test fold combinations. To obtain reliable estimates, the above procedure will be repeated 100 times (with random folds and subsamples each time), separately for two orientation spaces (bin centres: 5.625° to 174.015° and 11.25° to 180°, each in steps of 11.25°), and averaged across the resulting 200 samples for each of the 16 Mahalanobis distance values.

To obtain a summary measure of decoding accuracy, we will compute the cosine vector means of the tuning curves (Sprague et al., 2016; Wolff et al., 2017). Higher vector mean values will reflect greater orientation tuning and thereby greater decoding accuracy for that trial. We will take the mean over trials, resulting in a single decoding value for each time window of interest for each participant.

Cross-generalization methods

To test our primary hypothesis relating to use-specific coding in WM, we will conduct the above decoding analysis, but instead of training and testing on the full data set, we will now select the train-and-test sets to be either within or between task context. We will subsample the training data relative to the test data to ensure that the mean distance between test and training data (in number of trials) is equal for within-task and between-task decoding.

Within-task decoding will be computed separately for the two tasks by training and testing on change-detection trials only, and training and testing on rotation-

discrimination trials only. Between-task decoding accuracy will be computed by training on change-detection trials and testing on rotation-discrimination trials, and vice versa, training on rotation-discrimination trials and testing on change-detection trials. We will compute within-task and between-task decoding accuracy separately for the stimulus epoch and the impulse epoch. The entire decoding analysis pipeline has been validated on simulated EEG data (Appendix D; Supplementary Figure S1), where various scenarios show we are consistently able to identify both true effects and true null effects, ensuring that our analysis approach will lead to interpretable results.

To quantify the cost of cross-generalization, we will compare within-task decoding and cross-task decoding separately for each task on the basis of the test set. The cross-generalization cost for the change-detection task is computed by taking the difference between the decoding accuracy for train: change-detection, test: change-detection and train: rotation-discrimination, test: change-detection. Similarly the cross-generalization cost for rotation-discrimination is computed by taking the difference between decoding accuracy for train: rotation-discrimination, test: rotation-discrimination and train: change-detection, test: rotation-discrimination.

To control for any differences in overall strength of within-task decoding, we will test whether there is a significant difference in within-task decoding strength between change-detection and rotation-discrimination in each epoch (paired sample t-test, two tailed, $p<.05$). If, as expected, the conditions are well matched and there is no difference in overall strength of within-task decoding, the mean cross-generalization costs for the two tasks will serve as our dependent variable in subsequent analyses. In case of a significant difference in strength of overall within-task decoding, we will instead use the cross-generalization cost for the task with the poorest overall within-task decoding strength, defined per subject, as our dependent variable. This conservative approach

avoids any potential false cross-generalization costs due to differences in signal-to-noise ratio between the two tasks (Appendix D; Supplementary Figure S2B).

However, we identified one scenario where this approach could miss a subtle true effect: specifically if a task-specific code exists for one task only and is supplemented by task-invariant coding while there is a large difference in noise levels between the two tasks (Appendix D; Supplementary material, Scenario D). We suggest that this particular combination of multiple unexpected outcomes (i.e. large systematic differences in noise levels between tasks combined with a unique task-specific code for one task only) would be very unlikely. Nevertheless, two clear diagnostic measures would indicate if this was the case: asymmetries in cross-decoding costs (i.e. within – cross A ≠ within – cross B) and significant differences in noise levels between tasks. We will test for differences in noise levels between tasks for each participant by computing the first eigenvalue of the noise covariance matrix separately for each task and compare this to noise differences estimated from data with shuffled task labels, using a permutation test. We will calculate z-values of noise differences for each participant from the permutation test p-values and test whether these are significantly different from zero using a one-sample t-test (one-tailed, >0). This would warrant a follow-up analysis to recover a potential true effect by excluding subjects with significant differences in noise levels between tasks until they are matched in the remaining sample.

To test our primary and secondary hypothesis, we will evaluate whether the cross-generalization cost is significantly greater than zero. A positive difference (H_1) will reflect greater within- than between task decoding. No difference (H_0) would suggest a task-independent WM code for stimulus orientations (see next section for details on statistical comparisons).

The presence of a cross-generalization cost (H_1) in the data might reflect an underlying neural code where stimulus orientations are coded in a purely task-specific format or alternatively that task-specific coding is supplemented by a common code for stimulus orientation shared between the two tasks. Either scenario would be consistent with our hypothesis of use-dependent coding in WM. We will distinguish between these two possibilities by testing for the presence of significant overall cross-decoding for each task. Zero cross-decoding would suggest a purely task-specific coding format, while positive cross-decoding would suggest that task-specific coding is supplemented by a common stimulus code shared between the two tasks.

While we expect there to be a strong neural signal for task context, this does not trivially imply a task-specific code for stimulus orientations. We have recently shown that strong decodable context signal (e.g. temporal context) still allows for near-perfect cross-context decoding in EEG (Wolff, Jochim, et al., 2020). Additionally, simulated neural data specific to this study design confirms that it is in principle possible to have zero cost of cross-generalization even when task context is strongly decodable (Appendix D; Supplementary Figure 1B). As such, the existence of task-general neural patterns does not necessarily imply a cost of cross-generalization between tasks.

Test of primary question: Cross-generalization cost during impulse epoch

Orientation decoding of the cued item following the impulse must be greater than zero for us to test our primary hypothesis. Consequently, we will first test whether the cued item is decodable during the impulse epoch (100-400 ms from impulse onset) using a one sample t-test (one-tailed, >0, $p < .05$). Significant decoding of the cued item will act as a positive control for our primary hypothesis. For completeness, we will also test for decoding of the uncued item and for any difference between the cued and uncued item using one sample (one-tailed, >0) and paired sample t-tests (one-tailed, cued > uncued).

The outcome of these additional analyses will not affect our ability to test the primary hypothesis.

To test our primary hypothesis of a use-specific code for prioritized content in WM, our main analysis will focus on whether there is a cross-generalization cost for the prioritized item following the impulse. A significant cost of cross-generalization would suggest that relevant information in WM is coded in a use-specific format, as predicted (H_1). Alternatively, a non-significant cost of cross-generalization would suggest that relevant information in WM is coded in a use-invariant format (H_0). We will perform a Bayesian one sample t-test to test for evidence of a cross generalization cost (one-tailed, >0, default prior distribution). We will stop sampling when the data is 10 times more likely under either H_1 or H_0 than under the alternative hypothesis ($BF_{10} = 10$ or $BF_{10} = 1/10$) or when we reach the maximum feasible number of participants ($n_{max} = 113$, see Bayes factor design analysis below*).* BF will be computed after a sample of $n_{min} = 30$ participants has been collected and subsequently after every participant.

Additionally, to test whether a potential use-specific pattern is supplemented by a common stimulus pattern shared between the two tasks during the impulse epoch, we will perform Bayesian one-sample t-tests on the cross-decoding values separately for each task (one-tailed, >0, default prior distribution). Either outcome is consistent with our hypothesis of use-specific coding in WM. Evidence for or against this hypothesis will not be used as an evidence threshold for sampling purposes.

Test of secondary question: Cross-generalization cost during stimulus epoch

We will test the extent to which the orientations of the memory items are decodable during stimulus presentation (100-400 ms from WM array onset). As the two items are equally likely to be probed, we will use the mean decoding accuracy for the two items as the index of orientation decoding during stimulus presentation. We will test whether

orientation decoding is greater than zero, using a one sample t-test (one-tailed, >0, $p <$ .05). Significant decoding will act as a positive control for our secondary hypothesis.

In case we find evidence in support of our primary hypothesis of a cross-generalization cost for the prioritized item, we will run a second analysis to test whether items are encoded into WM in a task-specific format. If both items are encoded in a task-specific format, there should be a cost of cross-generalization during the stimulus epoch (H_1). Alternatively, if task-specific re-configuration occurs later during WM maintenance, such as following retro-cued prioritization, we would not expect to also see a significant cost of cross-generalization at this stage (H_0). We will perform a Bayesian paired samples t-test (one-tailed, within > between-task decoding of orientation, default prior distribution) to distinguish between these two alternatives for task-specific coding during encoding into WM.

Additionally, to test whether a potential use-specific code is supplemented by a common stimulus code shared between the two tasks during the stimulus epoch against the alternative, we will perform Bayesian one-sample t-tests on the cross-decoding values separately for each task (one-tailed, >0, default prior distribution).

Evidence for or against the secondary hypothesis will not be used as an evidence threshold for sampling purposes.

Exploratory analyses

In addition to the planned analyses specified in the previous section, the data collected as part of this study may benefit from further exploratory analyses that could shed light on the way the brain maintains information in the short-term in the service of a future goal. Specifically, we have planned two exploratory analyses to examine (1) whether

higher levels of use-dependent coding in WM are associated with faster and more accurate WM performance, and (2) which neural areas show evidence of use-dependent coding, using a searchlight analysis on the sensor-level EEG data. We describe these planned exploratory analyses in more detail next.

Relationship between cross-generalization costs and WM performance

To explore whether a potential use-dependent code promotes WM-guided behaviour, perhaps by optimizing WM content for a faster and more accurate translation from sensory input to the relevant decision, we will test whether the magnitude of the cross-generalization cost is predictive of behavioural performance (median RT and accuracy), separately for the change-detection task and the rotation-discrimination task. For RTs, evidence of a negative correlation would be consistent with our hypothesis that use-dependent coding formats facilitates a faster translation from sensory input to context-dependent motor output. For accuracy, a positive correlation would be consistent with our hypothesis that use-dependent coding promotes a more accurate translation from sensory input to motor output.

First, we will compute Pearson correlations across participants to test the hypothesis that participants with larger cross-generalization costs show better task performance. For each task, we will run separate correlations between cross-generalization cost magnitude and median RTs and between cross-generalization cost magnitude and mean accuracy. We expect to find a negative correlation between cross-generalization cost and median RT and a positive correlation between cross-generalization cost and mean accuracy.

Next, we will test whether participants' trial wise behavioural performance relates to the use-specificity of the neural code for stimulus orientations on individual trials. For RTs, we will run a linear regression for each participant to test whether trial wise cross-

generalization costs predict trial wise RTs (log-transformed). For accuracy, we will run a logistic regression for each participant to test whether trial wise cross-generalization costs predict trial wise accuracy. We will obtain betas from each regression for each participant and test whether these are significantly different from zero using Bayesian one-sample t-tests. We expect trial wise cross-generalization costs to negatively predict reaction times (one-tailed, <0), and to positively predict accuracy (one-tailed, >0).

Full topography and searchlight analysis

While previous studies report orientation decoding to be strongest in posterior EEG channels (Wolff, Jochim, et al., 2020; Wolff et al., 2017), task-related differences in WM coding schemes may recruit a wider network of neural areas involved in goal-directed behaviour and motor planning. Consequently, we will repeat the same decoding analyses as described in the previous section using the full EEG topography of 61 electrodes as input to the decoder, instead of posterior channels only. Next, to explore the topography of use-dependent WM coding, we will repeat the orientation decoding iteratively for a subset of electrodes in a searchlight analysis across all 61 EEG electrodes to obtain topographies for within- and between-task decoding separately for each task, as in (Wolff, Jochim, et al., 2020). In each iteration, the searchlight will be centered on a different electrode, which will be included in the analysis along with the two closest neighbouring electrodes. As above, we will take the difference over channels between within- and between-task decoding separately for each task before computing the mean cross-generalization cost. The resulting decoding topography will be visualized using the Matlab extension Fieldtrip (Oostenveld et al., 2011).

Bayes factor design analysis

We will use a sequential Bayesian sampling to increase the informativeness and efficiency of the study and avoid relying on published effect sizes. Using the BFDA toolbox in R (R Core Team, 2019; Schönbrodt & Wagenmakers, 2018) we computed estimates of expected sample size for a Bayesian sequential paired sample t-test (one-tailed) using Monte-Carlo simulations (10,000). We used a default prior Cauchy distribution (scaling parameter of $\sqrt{2}/2$) on a conservative estimated effect size of 0.4 for H_1 and a point effect size of 0 for H_0. An effect size of interest of 0.4 for the H_1 prior was chosen based on separate effect size values for orientation decoding of a prioritized item in WM from three similar previous studies (0.5437, 0.4975, 0.7594)(Wolff, Jochim, et al., 2020; Wolff et al., 2017). We took the mean of these previously reported effect sizes ($M = 0.6001$) and scaled this by 66%, assuming effect size estimates based on published studies might be inflated ($0.6001 \times 0.66 \approx 0.4$). For the simulations, we used a minimum sample size of $n_{min} = 30$ and an initial maximum sample size of $n_{max} = 300$, a step-size of 1 (i.e. computing BF after every sample), and an evidence threshold of $BF_{10} = 10$. Table 2 shows the outcome of these simulations ($N_{max} = 300$). When H_1 is the ground truth, the median sample size at the stopping point is 41 and 95% of all studies reach the evidence threshold with fewer than 113 participants. When H_0 is the ground truth, the median sample size is 69 and the 95% quantile is 300 (n_{max} in the simulations), reflecting the fact that 10.4% of studies haven't reached the evidence boundary with fewer than 300 samples.

Since 95% of studies under H_1 reach the evidence threshold with fewer than 113 participants, we re-analysed the simulations using a maximum sample size of $n_{max} = 113$. The outcome of these analyses can be seen in Table 2 ($N_{max} = 113$) and Figure 2. When H_1 is the ground truth, 95% of studies stop at the H_1 evidence boundary (true positives), 0.2% stop at the H_0 evidence boundary (false negatives) and 4.8% terminate at

the maximum sample size. Of these, 2.4% show sub-threshold, but interpretable evidence for the H_1, 2.4% show inconclusive evidence and 0.1% show sub-threshold evidence for the null hypothesis. When H_0 is the ground truth, the lower n_{max} results in fewer studies terminate at the H_0 boundary (66.1%) and there is a corresponding increase in studies stopping at n_{max} (31.9%). However, of these studies, the large majority (21.1%) show interpretable sub-threshold evidence for the null hypothesis. 10.4% of studies show inconclusive evidence (3 > BF > 1/3) and 0.3% show sub-threshold evidence for the H1. The proportion of studies stopping at the H_1 boundary (false positives) is still small at 2.1%.

Table 2. Outcome of simulations of sequential Bayesian paired-samples t-tests (one-tailed, >0) in population with a Cauchy distribution (scaling parameter=$\sqrt{2}/2$) over a true effect size 0.4 for H_1 and a point effect size of 0 for H_0. The simulations were analyzed separately assuming a maximum sample size of nmax = 300 (left) and n_{max} = 113 (right). In both cases, the minimum sample size was n_{min}=30, the step-size was 1 (computing BF after every sample) and the evidence boundary was BF_{10}=10 and BF_{10}=1/10.

	N_{max} = 300		**N_{max} = 113**	
	H_1	H_0	H_1	H_0
Studies terminating at n_{max} (%)	0	10.4	4.8	31.9
Studies terminating at H_1 boundary (%)	99.8	2.7	95	2.1
Studies terminating at H_0 boundary (%)	0.2	86.9	0.2	66.1
50% quantile (sample size)	41	69	41	69
95% quantile (sample size)	113	n_{max}=300	n_{max}=113	n_{max}=113
Of studies terminating at n_{max}:				
Evidence for H_1, BF > 3 (%)	0	0.2	2.4	0.3
Inconclusive, 3 > BF > 1/3 (%)	0	3.9	2.4	10.4
Evidence for H_0, BF < 1/3 (%)	0	6.3	0.1	21.1

Due to resource limitations and steeply diminishing returns at larger sample sizes, we believe a maximum sample size of $n_{max} = 113$ achieves the right balance between feasibility and interpretability of results. We believe this maximum sample size is sufficiently high that an inconclusive result would still be an important message to the field.

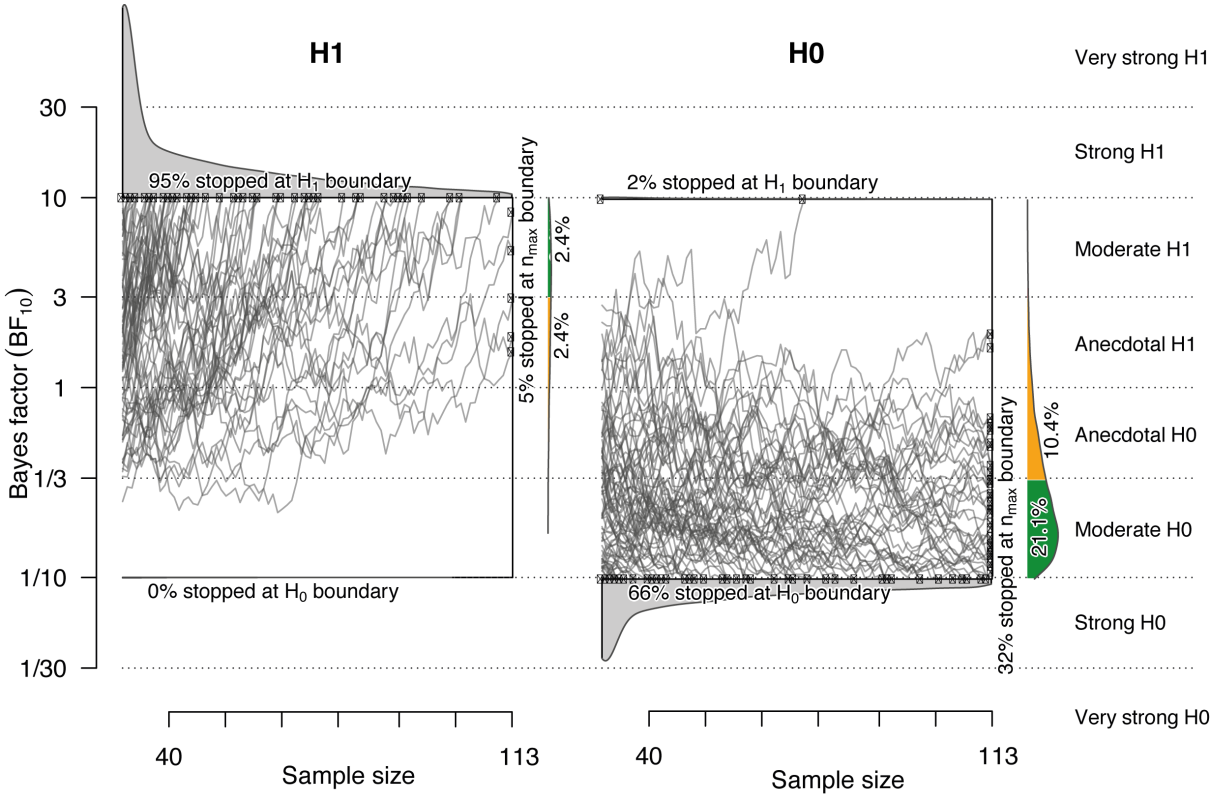

Figure 2. Bayes Factor Design Analysis. Depicts 100 simulations of studies in a population with a default prior Cauchy distribution (scaling parameter of $\sqrt{2}/2$) on an effect size of 0.4 for H_1 (left) and a point effect size of 0 for H_0 (right). Evidence boundary is BF=10 and BF=1/10, minimum sample size is 30 and maximal sample size is 113.

Timeline

If Stage 1 review is successful, we expect it will take a total of up to 9 months to complete this study before we would submit a final manuscript for Stage 2 review. This would include up to four months for data collection. Following data collection, we expect it will take 4-5 months to pre-process and analyse the data and write up the results for Stage 2 submission.

Data sharing statement

We are committed to sharing all study materials. Study materials, including anonymized raw and preprocessed data, the task script, EEG preprocessing and analysis code, will be deposited on the Open Science Framework (OSF) indefinitely upon acceptance of the final manuscript.

Chapter 6. General Discussion

The aim of this thesis was to investigate how information in WM is selected and prioritized for future action. I developed two lines of research that each address different aspects of the main research question. In a first set of experiments, I examined attentional selection in WM (Chapter 2-3). I specifically hypothesized that attention automatically spreads between object features in WM. The second line of research concerned how WM content is prioritized for action (Chapter 4-5). I predicted that task-relevant WM contents are configured as prospective task-sets in preparation for behaviour. I will present the main findings of each project, discuss the implications and how the results help shed light on the ways in which WM supports goal-directed behaviour.

Project 1: Mechanisms for attentional selection in WM

In Chapter 2-3, I presented behavioural and neural evidence in support of the hypothesis that attentional enhancement spreads between an object's features in WM. Using an incidental cueing paradigm inspired by Zokaei and colleagues (Zokaei, Manohar, et al., 2014; Zokaei, Ning, et al., 2014), I tested the consequences of selecting a single feature of an object in WM during the delay (incidental cue) for the representation of other features belonging to the same object.

The 'incidental' delay task asked participants to respond only if the cued feature did not match any of the features in memory and thus did not require participants to recall any other features of the cued object. Even so, behavioural data reported in Chapter 2 (Experiment 1) revealed that the precision of orientation recall was higher and responses faster when a feature of the same object was cued during the memory delay (congruent) compared to when a different object was cued (incongruent). Thus, orienting attention to

the cued feature privileged other features belonging to the same object in WM. The behavioural congruency effect was reproduced across two subsequent experiments in Chapter 2 (Experiment 2) and Chapter 3. Auto-associative enhancement of object-features is predicted by recent models of attention in WM, where retrieval occurs through obligatory pattern completion (Manohar et al., 2019), but contrasts with research suggesting attention can be allocated to selectively enhance task-relevant features in memory (Bocincova & Johnson, 2019; Hajonides et al., 2019; A. Heuer & Schubö, 2016; Serences et al., 2009). Overlap and discrepancies with previous literature will be discussed in more detail next.

The behavioural observations are consistent with previous studies of attention in WM demonstrating relative performance benefits following non-predictive cues during the maintenance period, such as 'refreshing cues' (M. R. Johnson et al., 2007; Souza et al., 2015; Yi et al., 2008) or 'incidental cues' as in this task (Zokaei, Manohar, et al., 2014; Zokaei, Ning, et al., 2014). These established that attending to a cued object during the WM delay privileges the object's status in memory. In the experiments presented in this thesis, I was interested in whether attending to a *single feature* of an object similarly changes the status of the *entire* object in memory. In contrast to previous incidental cueing studies, the present task did not require participants to report any other features associated with the incidental cue. Thus, the behavioural findings in Chapter 2-3 ruled out *active* recall of binding information during the delay as a source of the incidental cueing benefit and showed that simply attending to a single feature of an item privileged other features belonging to that item in memory.

Moreover, I varied cue and probe feature characteristics independently to examine the extent to which the congruency benefit was dependent on the feature dimension used to cue recall (colour vs location). There was a robust congruency benefit on reaction times regardless of cue and probe characteristics. For error, colour cues were associated with a

congruency benefit whether orientation recall was probed by colour or location, but location cues improved memory precision on congruent relative to incongruent trials only when recall was also probed by location.

These results are novel in two ways. First, the presence of a congruency effect even when orientation recall was probed by a different feature dimension than the cue, in most cases, indicates that the benefit was not merely explained by heightened accessibility of the cued feature or enhanced connectivity between the cued feature and its orientation. Instead, it suggests that the entire object was brought into a privileged state, consistent with obligatory pattern completion and object-based selection in WM (Manohar et al., 2019). Second, the asymmetry for location cues may point to a special status of location in WM, supported by extensive evidence (Jiang et al., 2000; Pertzov & Husain, 2014; Schneegans & Bays, 2017a; Treisman & Zhang, 2006; B. Wang et al., 2016; Wood, 2011). The asymmetry was initially surprising, as I had suspected location primacy in WM might make location a stronger cue for automatic retrieval. Instead, the opposite was the case. This pattern of results may be expected if location information can be accessed independently of colour information, while colour information cannot be accessed independently of location. This would be in accordance with recent computational models positing retinotopic mapping of non-spatial features in WM, which predicts an important role for location in selecting non-spatial features from memory (Bocincova et al., 2020; Schneegans & Bays, 2017a).

A special role of location in colour-selection in WM was corroborated by two combined EEG and eye-tracking experiments presented in Chapter 2 (Experiment 2) and Chapter 3. Participants performed a modified version of the experimental task where the incidental task always cued a colour. This improved statistical power to investigate the role of spatial attention in colour-selection from WM. I specifically looked at alpha lateralization (Worden et al., 2000) and gaze bias (van Ede, Chekroud, & Nobre, 2019) as

indices of spatial attention following the centrally presented, non-predictive colour cue. Consistent with the hypothesis that spatial attention is involved in selecting the cued colour from WM, oscillatory power in the alpha frequency band was suppressed over posterior EEG channels contra- relative to ipsilateral to the hemifield where the cued item was originally presented. Additionally, participants' gaze position was biased toward the cued item's memorised location. In Chapter 2, orientation recall was probed by either location or colour, leaving open the possibility that spatial attention was engaged in selecting the cued colour from memory because spatial position remained a relevant feature for retrieval at the probe. However, this explanation was ruled out in Chapter 3 where recall was never probed by spatial position making location a fully redundant feature.

Previous studies have shown contralateral alpha suppression relative to attended positions in WM cued by predictive, spatial retro-cues (Myers, Walther, et al., 2015; Wallis et al., 2015). Recent evidence suggests alpha lateralization may similarly be modulated by predictive, *non-spatial* retro-cues, such as colour, suggesting a role for spatial attentional mechanisms in selecting objects from WM based on non-spatial features (Poch et al., 2017). The results presented in this thesis support a role of spatial attention in colour-selection from WM even when the cue does not motivate retrieval of any other object-features.

While the precise mechanisms underpinning alpha oscillations are unresolved, a prominent theory suggests alpha power is important for gating neural information flow by increasing excitability in relevant brain areas when desynchronised, and/or inhibiting activity in neural populations coding for distractors when synchronized (Bollimunta et al., 2011; Bonnefond & Jensen, 2012; Foxe et al., 1998; Jensen & Mazaheri, 2010), perhaps controlled top-down through low-frequency oscillatory activity in frontal areas (de Vries et al., 2018; Lara & Wallis, 2014). In the present study, alpha lateralization

may contribute to attentional selection of task-relevant content in retinotopically organized portions of the posterior cortex by increasing excitability and facilitating information flow between relevant areas of sensory cortex and frontal areas representing current goals.

Small changes in gaze position are thought to track the focus of spatial attention within WM (Van Ede et al., 2020; van Ede, Chekroud, & Nobre, 2019). There is substantial overlap in the brain areas engaged during spatial attention and eye-movements (Kustov & Robinson, 1996; Nobre et al., 1997), leading to proposals that attention may even be an emergent property of motor planning (Deubel & Schneider, 1996; Olivers & Roelfsema, 2020; Rizzolatti et al., 1987). Indeed, a crucial function of spatial attention is to prepare for where to look next. However, once information is encoded into WM the need to prepare for where to look dissipates, particularly when the memory probe is presented centrally as was the case here. Gaze bias observed during internal orienting in WM may instead be a consequence of oculomotor system engagement in spatial attention and as a result "spilling over" into eye-movements (van Ede, Chekroud, & Nobre, 2019). Thus the shift in gaze bias following the cue may reflect orienting of spatial attention to select the cued colour in WM, perhaps stored retinotopically (Schneegans & Bays, 2017a).

Spatial attention has been proposed as a mechanisms for "rehearsing" or "refreshing" contents in visual WM periodically to prevent information loss (Awh & Jonides, 2001; Awh & Vogel, 2006). Attentional refreshing is thought to bring WM content into a temporary focus of attention and increase its accessibility for later retrieval (Souza et al., 2015; Vergauwe & Langerock, 2017). Traditionally, retro-cue benefits have been explained in terms of sustained internal attention toward the cued item in memory (Chun, 2011; Griffin & Nobre, 2003; Larocque et al., 2014; Ma, Husain, & Bays, 2014; Schneegans & Bays, 2017b). However, it is worth noting that contralateral alpha

suppression and shifts in gaze bias following the cue were temporary in both experiments and realigned approximately one second following cue onset. This is consistent with previous studies showing transient alpha lateralization following predictive retro-cues (Myers, Walther, et al., 2015; Wallis et al., 2015) and transient gaze bias following non-predictive cues as in the current experiments (Van Ede et al., 2020). This implies that once selected, the cued item remained in a privileged state even if it is no longer actively refreshed via spatial attention.

Instead, measures of spatial attention in Chapter 2 (Experiment 2) suggested participants may actually have oriented spatial attention toward the uncued memory items after the incidental task component was completed. Alpha lateralization reversed ~1s after the cue in a way that predicted orientation recall on incongruent trials, supporting a role in re-orienting spatial attention and alleviating potential costs for uncued items. Further, the positions of the two uncued items remained decodable from eye-position later in the second delay, while decoding for the cued item was at chance, suggesting spatial attention may refresh the uncued items before probe onset. However, as these effects were not significant in Chapter 3, they are interpreted with some caution. Nevertheless, both studies corroborate the finding that sustained spatial attention was not required to maintain the cued item in a privileged state. It is possible that the selected item remains in a focus of attention that is independent of spatial attention, but this seems unlikely given that other studies have reported robust retro-cue benefits even if attention is occupied by a dual task or directed towards a different item following the cue (Myers et al., 2017; Rerko & Oberauer, 2013; Rerko et al., 2014; Souza et al., 2015; Van Moorselaar et al., 2015). Instead, the results suggest sustained attention is not required to maintain content in a privileged state in WM following selection. I will discuss potential alternative mechanisms for the behavioural cueing effects below (p. 240-241).

The results discussed so far suggest that selecting a single feature of a multi-feature object temporarily engages spatial attention toward the memorized location of the cued object and privileges memory for associated features. A wealth of evidence shows that attended, task-relevant objects and features in WM are preferentially represented in brain activity, as decoded from EEG and BOLD activity, particularly in sensory cortex (Christophel et al., 2018; LaRocque et al., 2013; Larocque et al., 2014; Lewis-Peacock et al., 2012; Rose et al., 2016; Sprague et al., 2016; Wolff et al., 2017). Using multivariate pattern analyses of EEG, I tested whether focusing attention on the cued colour by extension strengthens the representational quality of the cued object's memorized location and orientation, relative to uncued object-features. Confirming the hypothesis, there was a selective improvement in decoding quality for the location (Chapter 2) and orientation (Chapter 3) of the cued memory item following the cue. The results suggest colour-selection in WM improves the neural representation of other features originally encoded on the same object as the cued colour, even when these were redundant at the time of the incidental task. Importantly, the observations suggest objects, rather than individual features, are the unit of selection in WM.

A recent computational model posits that features in WM are bound and retrieved through strengthened synaptic traces between neurons coding for sensory features and flexibly coding conjunctive neurons (Manohar et al., 2019). Encoding occurs through serial attention and the last attended item remains the focus of attention through sustained activity in the units coding for the item's features. Unattended items are represented in patterns of synaptic traces established during encoding. An item may be brought into the focus of attention through cued-recall, when one of the sensory neurons is stimulated, activating the relevant conjunctive neuron and associated object-features by auto-associative pattern completion. The model predicts that neural coding for all features of an object in WM should be re-activated when orienting attention to a single

feature. Thus, re-emergence of location and orientation decoding quality following the cue is consistent with the predictions of the model.

Would models without synaptic, "activity-silent" mechanisms similarly predict re-emergence of decoding quality for associated object-features following the cue? In general, models where WM contents are stored in persistent activity in selective neurons (Compte et al., 2000; X. J. Wang, 2001; Wimmer et al., 2014), struggle to account for different representational states in WM. However, a recent model proposes that differences in decoding quality for cued and uncued WM content may nevertheless be explained without appealing to latent memory states (Schneegans & Bays, 2017b). Instead, differences in representational quality are explained in terms of biased competition (Desimone & Duncan, 1995). When multiple items are encoded into WM, they compete for representational resources in a noisy neural population. Decoding accuracy should decrease over time and with increasing set size due to due to noise and lateral inhibition between items. Thus, chance-level decoding is explained in terms of lower signal-to-noise rather than activity-silent coding mechanisms. When a memory item is cued as relevant, resources are re-allocated toward the cued item, at expense of the uncued item, resulting in re-emergence of decoding quality for the cued item.

Schneegans & Bays' (2017b) model makes use of fixed conjunctive coding such that non-spatial features are represented in neural fields with retinotopic selectivity, as was the case in a model of feature-binding in WM by the same authors (Schneegans & Bays, 2017a). Thus, selecting the colour of a memory item should strengthen decoding quality for the cued item's spatial location, as found in Chapter 2. However, it is not clear that the model would similarly predict obligatory re-emergence of decoding quality for the cued item's orientation (Chapter 3). Moreover, the model does not easily account for neural and behavioural cueing effects when the cue is non-predictive, as was the case in the present experiments. This is because retro-cue benefits in the model are achieved by

dropping irrelevant items, thereby allowing more resources to be allocated to the cued item. The cue in the current experiments did not allow uncued memory items to be dropped from memory. Yet, without a neutral cue condition, I could not rule out that cueing effects came at the expense of uncued items, as in a previous similar study (Zokaei, Ning, et al., 2014). However, others have shown that attention can be flexibly shifted between items in WM (LaRocque et al., 2013; Lewis-Peacock et al., 2012; van Ede et al., 2017; Zokaei, Ning, et al., 2014) and retro-cue benefits are sometimes observed without associated costs to other memory items (Li & Saiki, 2014; Myers et al., 2018). Consequently, the findings of the experiments presented in Chapter 2-3 are better explained by a model in which features are retrieved through auto-associative mechanisms and where attention allows cued memory contents to be strengthened without dropping uncued items (Manohar et al., 2019).

Obligatory restoration of associated object-features following the non-predictive cue initially seems to contrast with previous work showing selective coding for task-relevant features in WM (Bocincova & Johnson, 2019; Serences et al., 2009). Particularly, the current results seem at odds with recent work showing successful prioritization of features in WM (Hajonides et al., 2019; Niklaus et al., 2017). Hajonides et al. (2019) showed retro-cue benefits were of a similar magnitude for feature- and object-based retro-cues indicating that it is possible to allocate feature-based attention to prioritise a single feature-dimension of multi-feature objects. If attention spreads automatically between features belonging to the same object in WM, there should be no benefit of feature-based retro-cueing. However, a crucial difference between the experiments presented in this thesis and the abovementioned predictive retro-cueing paradigms is that the latter allow irrelevant features to be *dropped* from memory entirely (Souza & Oberauer, 2016), potentially freeing up resources for processing the cued features. In the present experiments, all features remained relevant to orientation recall (except location

in Chapter 3) and may constitute a relevant bundle of connected features that are refreshed whenever a single feature of that bundle is selected.

Two ensuing questions include whether a predictive feature-based retro-cue initially retrieves all features belonging to the same object after which the irrelevant features are either strategically dropped through top-down control and/or through lack of refreshing. This would square with one hypothesis of object-based encoding proposing that all features are initially encoded into WM after which irrelevant features are dropped (Bocincova & Johnson, 2019; Jaswal, 2012; Soto & Humphreys, 2009). A second question is whether a non-predictive incidental cue only strengthens task-relevant object features as determined by a predictive retro-cue. I tested a version of this hypothesis in Chapter 3 where location constituted an irrelevant feature, but spatial attention was nevertheless directed toward the memorized position of the cued item. However, location may be a special case and this prediction should be tested with non-spatial features in memory.

Although neural decodability for the cued item's memorised location and orientation re-emerged following the cue, neither was sustained until the onset of the probe. If decoding strength is indicative of attentional states within WM, corroborated by a wealth of literature (LaRocque et al., 2013; Larocque et al., 2014; Lewis-Peacock et al., 2012; Rose et al., 2016; Wolff et al., 2017), it raises an important question of how the behavioural benefit for the cued item arises if not through sustained focused attention. Both models of attention in WM, discussed above, predict that retro-cueing benefits are a consequence of the cued item being maintained in a heightened state of neural activation corresponding to the focus of attention, while uncued items are represented in lower levels of activation (Schneegans & Bays, 2017b) or in activity-silent synaptic traces (Manohar et al., 2019). In one model, a heightened state of activation for the cued item supports memory precision by means of greater signal-to-noise (Schneegans & Bays,

2017b), while the active state in the other model is more accessible and does not depend on retrieval through an error-prone pattern completion process (Manohar et al., 2019).

If cueing benefits are supported by the neural activity state, why was improved decoding for cued item features transient following the cue in the present experiments? It is possible that decoding from raw EEG voltages was not well-suited to detect sustained differences in activation levels of neural populations coding for cued and uncued item-features, particularly in this case where all items remained relevant to orientation recall. Alternatively, neural coding for selected memory items is perhaps transferred to frontal areas of cortex and therefore not detected in the current experiment focusing on decoding from posterior areas. However, this would contrast with previous research showing strong classification of task-relevant WM content specifically from sensory regions, while decoding of detailed WM representations from PFC is generally poor (Christophel et al., 2018; Lara & Wallis, 2015; Lee & Baker, 2016; Serences et al., 2009).

Finally, it is possible that the nature of the relationship between the cognitive state in WM (prioritized vs unprioritized) and the neural activity state (active vs latent) is not 1:1 (Stokes et al., 2020). Instead attention may be required during the initial selection of content from WM, engaging spatial attention to activate the relevant neural representation. During this stage, attended WM content is proposed to be represented in heightened levels of activity. That would be consistent with recent theoretical and experimental work showing that active neural states are required for manipulation of WM content (Masse, Yang, Song, Wang, & Freedman, 2019; Trübutschek et al., 2017). Once manipulation is complete, sustained attention (and perhaps active neural coding) is no longer required, but the selected item remains in a privileged state in WM. In the present experiments, participants "manipulated" the colour in WM by making a match/non-match judgement. Although they were not required to manipulate any other features in memory such as location or orientation, selection and manipulation may be

performed at the object-level. Convergent evidence from behavioural and neuroimaging experiments support privileged WM states in the absence of focused attention (Myers et al., 2017; Rerko & Oberauer, 2013; Rerko et al., 2014; Souza et al., 2015). Consequently, attention may only briefly be engaged to select relevant WM content but nevertheless support a transition of the selected item into a privileged state enhancing WM-guided behaviour. A major outstanding question concerns the nature of that privileged WM state if it is not a matter of the attentional state or level of neural activation.

One account predicting retro-cue benefits without the need for sustained attention proposes that cued selection strengthens WM traces and the connectivity between object-features in memory (Rerko & Oberauer, 2013; Rerko et al., 2014; Souza et al., 2015; Tabi, Bocincova, Husain, & Manohar, submitted). The reinforced trace lingers even after attention is removed leaving the cued item in a more accessible state for retrieval. Evidence for a strengthening hypothesis of the retro-cue benefit comes from a study asking participants to "think of" or refresh items in WM in turn (Souza et al., 2015). Memory error decreased linearly with the number of times an item was refreshed during the delay, without costs to other items. The benefit of refreshing was present in the absence of sustained attention, supporting previous results (Rerko et al., 2014).

In a WM task with dual sequential retro-cues, Rerko et al. (2013) further showed that previously cued colours in memory were better rejected than uncued colours, suggesting that these previously selected items had stronger location-colour bindings than baseline, consistent with a 'strengthened traces' account of privileged states in WM in the absence of focused attention. Enhanced binding following refreshing cues is also predicted by recent theoretical models where attention strengthens the binding between co-activated features (Manohar et al., 2019; Oberauer & Lin, 2017). While these models generally predict sustained activation for the last attended item, it is possible that attending an item in WM strengthens synaptic connectivity between units coding that item such that

benefits remain as a function of the recency with which an item was attended rather than sustained attention per se (Tabi et al., submitted).

An alternative, though not mutually exclusive, theory for potential retro-cue benefits without sustained attention is one where the cued item is re-configured into a *functionally* active state that interacts with ongoing processing and drives decision-making (A. Heuer et al., 2020; Myers et al., 2017; Olivers et al., 2011; Olivers & Roelfsema, 2020; Stokes et al., 2020). The crucial difference between currently relevant vs prospectively relevant memory content is the extent to which the information should be allowed to interact with new sensory input and drive behaviour (Stokes et al., 2020).

Analyses of neural processing at the probe in Chaper 3 provided support for an account where there cued item is placed in a privileged state for interacting with the probe at decision-making, though remains ambivalent to whether this state is activity-silent or not. Facilitated comparison between the cued memory item and the probe is predicted by recent theoretical models of attention in WM positing that selected memory content may be configured in an action-oriented state (Myers et al., 2017; Olivers et al., 2011; Olivers & Roelfsema, 2020). For example, filtering new sensory input through a neural pattern coding for the prioritized memory item (i.e. matched-filter) may serve as an efficient mechanism for computing the relevant decision signal—in this case, the relative angular distance of the mnemonic template relative to the probe (Hayden & Gallant, 2013; Muhle-Karbe et al., 2020; Myers, Rohenkohl, et al., 2015; Sugase-Miyamoto et al., 2008). The cued item may be maintained in a functionally active state, that influences ongoing processing by virtue of its recent use in decision-making, thereby facilitating performance on congruent trials, while interfering with performance on incongruent trials (A. Heuer et al., 2020; Myers et al., 2017; Olivers & Roelfsema, 2020; Stokes et al., 2020).

Previous theoretical and experimental work has suggested different mechanisms for WM retrieval depending on their priority-state in memory (Nee & Jonides, 2008, 2013a, 2013b; Oberauer, 2002; Oztekin et al., 2010). For example, Nee and Jonides (2008) demonstrated that the most recently attended WM content, thought to remain in the focus of attention, could be accessed directly, while retrieval of other items engaged the medial temporal lobe, commonly involved in retrieval from long-term memory through hippocampus-mediated pattern completion (Cohen et al., 1999; Eichenbaum, Yonelinas, & Ranganath, 2007; Nee & Jonides, 2011, 2013a). As discussed previously, there are reasons to believe the cued item in this study did not remain in the focus of attention in the standard use of the term. Nevertheless, attentional selection following the cue may establish a functionally active state through strengthened functional connectivity in the network coding for the cued item (Oberauer & Lin, 2017; Rerko & Oberauer, 2013; Tabi et al., submitted), perhaps combined with enhanced links to neural areas responsible for configuration of task-sets, motor planning and decision-making (Myers et al., 2017; Olivers & Roelfsema, 2020). This may facilitate faster and more efficient comparison between the probe and the cued item compared to other items in memory that may depend on slowed and error-prone retrieval mechanisms. This action-oriented view of WM prioritization motivated the experiments described in Chapter 4-5, discussed next.

Project 2: How is WM content prioritized for action?

In recent years, several authors have proposed that prioritization of relevant WM content arises, in part, through coupling selected WM content to its anticipated use in decision-making and behaviour (A. Heuer et al., 2020; Myers et al., 2017; Nobre & Stokes, 2019; Olivers & Roelfsema, 2020). In a second line of research, I investigated how WM content is prioritized for action. In two behavioural studies, I showed that relevant WM content is configured as prospective templates for decision-making, and

that successful prioritization depends, in part, on such reconfiguration (Chapter 4). In Chapter 5, I presented a detailed plan for a future EEG study, in the form of a Stage 1 Registered Report, to test whether neural coding for relevant WM content depends on its anticipated use.

In Chapter 4 (Experiment 1), I developed a modified version of the task-rule congruency paradigm, previously used to study configuration of instruction-based task sets in the cognitive control literature (Braem et al., 2019; Liefooghe et al., 2013, 2012). Crucially, task-rule congruency effects are found when stimulus-response mappings are actively being prepared for a context-dependent response, and not when they are merely memorized (Liefooghe et al., 2013). Thus, I expected to find similar task-rule congruency effects only if items in WM were configured as prospective task-sets, prepared for context-dependent decision-making.

As predicted, orientations encoded into WM for a delayed rotation-discrimination task automatically drove responses for an orthogonal delay task in a task-specific manner. Colour-responses on a nested delay task were faster and more precise when targets were *congruent* and the S-R mapping for responding to the target's colour matched the S-R mapping for responding to the target's (irrelevant) tilt direction, compared to when targets were *incongruent* and the S-R mappings for the target's colour and relative orientation were conflicting (task-rule congruency effect). Interestingly, task-rule congruency effects were specific to WM items that were retro-cued as relevant for subsequent recall and were not present for uncued items. These results are consistent with a model where prioritized WM content is configured as a prospective task-set, such that a probe stimulus quickly triggers the appropriate response in a reflex-like manner (Hommel, 2000; Myers et al., 2017). Configuring memory content in a task-specific manner may be advantageous for task-relevant content, thus allowing information in

WM to drive behaviour, but not for other memory items that might interfere with the current task (Olivers et al., 2011; Peters et al., 2009).

However, the present task utilized a 100% valid retro-cue and a single probe meaning that the uncued item could be dropped from memory before the nested task. In other words, the experiment does not rule out the possibility that all content in WM is maintained in a prepared, action-oriented state. This could be tested in a dual serial retro-cue design, where both memory items are relevant to behaviour but at different points in the trial (LaRocque et al., 2013; Lepsien & Nobre, 2007; Lewis-Peacock et al., 2012). Additionally, to confirm that the effects reported in Chapter 4 (Experiment 1) are indeed specific to the nature of the current task read-out demands, future studies should test for task-rule congruency effects in the context of different WM tasks (e.g. change-detection). In the context of a change-detection task, I would expect responses on the delay task to be modulated by the *absolute* angular distance between the memory item and the target rather than the relative tilt direction.

Two recent behavioural studies further support the hypothesis that prioritized information in WM is maintained in a state that influences ongoing motor behaviour. First, prioritized verbal content in WM was shown to influence hand movements (J. A. Miller et al., 2020). In a dual-task experiment, the speed and trajectory of participants' hand movements was biased by a verbal direction maintained in WM, even when the direction in WM was antithetical to the motor task demands. Interestingly, motor bias was greater for retro-cued WM content compared to uncued content, suggesting that prioritized information in memory may be in a special position to influence motor behaviour.

In a task-rule congruency paradigm similar to the one presented here, González-García et al. (2020) showed that pairs of explicit stimulus-response (S-R) mappings encoded into

WM influenced responses on a nested delay task only when retro-cued as relevant for execution (González-García et al., 2020). Crucially, uncued S-R pairs did not affect responses even though the entire memory display was probed on a subset of trials, adding to the results of this thesis by showing that WM may hold uncued information in a state that does not influence ongoing processing. This is consistent with previous studies showing that only task sets prepared for execution, not ones that are merely memorized, drive behaviour (Liefooghe et al., 2013) and suggest retro-cueing may have a similar function of preparing a remembered task-set for implementation. The important novel aspect of Chapter 4 (Experiment 1) is showing that retro-cueing may further transform *detailed sensory representations* in memory (e.g. orientation angles) into task-sets in preparation for context-dependent action.

If action-oriented states serve to optimize WM-guided behaviour, I would have expected the task-rule congruency benefit in Chapter 4 (Experiment 1) to correlate with performance on the WM task, which it did not. However, the sample size of the present experiment was small compared to a previous, similar study showing individual differences in prepared task sets (Braem et al., 2019), suggesting the power to detect individual differences may have been insufficient. Future studies specifically designed to look at individual variability will be better suited to test this prediction.

Another way to assess the extent to which WM-guided behaviour is supported by prospective task-specific states in memory is by observing behaviour under conditions that would allow for task-specific coding compared to conditions that would not. Myers et al. (2017) proposed a two-step model of retro-cued prioritization in WM. First, top-down attention is engaged to select relevant content in WM. Next, the selected item is reconfigured into an action-oriented state for guiding behaviour. If prioritization is supported by a task-specific reconfiguration of WM content in preparation for behaviour, foreknowledge about the expected task should enhance the retro-cue benefit. In Chapter

4 (Experiment 2), I compared the size of the retro-cue benefit when participants had prospective knowledge of the WM task rule to the size of the retro-cue benefit when the task was unknown. As expected, the retro-cue benefit on reaction times was enhanced with prospective task knowledge. This was not the case for accuracy, indicating that task-specific coding of prioritized content may enable faster, but not more accurate, translation into decision-signals.

Retro-cueing in WM may be considered as 'selection-for-action' (Allport, 1987; A. Heuer et al., 2020). When multiple, potentially task-relevant objects are stored in WM, a predictive retro-cue may function to select *one* object to act upon next. Allport (1987) argued that selection is intimately tied to the desired action – the purpose of attentional selection is generally to act upon the selected object. While it is possible to perceive multiple objects, it is generally only possible to complete one action at a time. Selection for action is also known as 'output gating'–that is, gating which information in memory is allowed to drive behaviour. In line with a 'selection-for-action' hypothesis of retro-cueing, retro-cues compared to pre-cues uniquely recruit neural areas commonly engaged during motor planning and task-set preparation, including areas of the pre-motor cortex and pre-supplementary motor area, basal ganglia and the frontal operculum (Chatham et al., 2014; Higo et al., 2011; O'Reilly & Frank, 2006; Rushworth et al., 2002; Schneider et al., 2017; Tamber-Rosenau et al., 2011; Wallis et al., 2015). Retro-cues facilitate internal attention in order to select an existing mental representation. In theory, this allows a motor plan to be configured right away for how to use the selected information. By comparison, pre-cues motivate anticipatory attention toward relevant information that has not yet been presented, making it impossible to plan for how to use the information in advance. Thus, recruitment of regions typically involved in action planning and task-set configuration is consistent with an output gating hypothesis

where retro-cues facilitate the configuration of a task-oriented neural state in preparation for behaviour.

While the behavioural data presented in Chapter 4 (Experiment 2) support a use-specific model of WM prioritization, there are potential alternative explanations that need to be considered. First, the benefit of prospective task knowledge was assessed relative to a neutral condition, making it difficult to tease apart contributions of prospective WM coding from costs arising due to additional resources required to process the task cue at probe onset. Overall, this should affect informative and neutral retro-cue conditions similarly and thus does not predict the interaction between task-foreknowledge and retro-cueing found here. However, if participants employed a strategy of anticipatory spatial attention to facilitate probe processing on informative retro-cue trials, probe processing may have been impaired on trials where they were also required to process the task cue. While this source of the interaction effect cannot be ruled out, two factors make it less likely. First, accuracy was still high in conditions with no task foreknowledge but informative retro-cues, suggesting probe processing was still good on these trials. Second, the task cue remained on screen until a response was made, so participants did not risk missing the task rule even if directing spatial attention to the anticipated location of the probe on informative retro-cue trials.

In the task-switching literature, performance costs in mixed task blocks compared to fixed task blocks are referred to as 'mixing costs' (Monsell, 2003). Mixing costs have been proposed to arise, in part, due to greater load on WM in mixed compared to single task blocks as two task rules need to be maintained in WM simultaneously (Rogers & Monsell, 1995). If so, this could similarly affect WM performance in the current study if fewer WM resources are available for maintaining orientations with high precision in the mixed task blocks. However, it is not clear that task rules compete for the same WM resources as required for more standard WM tasks (Liefooghe et al., 2008; Rubin &

Meiran, 2005). Rubin and Meiran (2005) found that the number of S-R mappings needed for a task did not affect the size of the mixing cost, suggesting instead mixing costs arise due to competition between overlapping S-R mappings. Overlapping S-R mappings were also used in the present experiment. Even so, varying prospective task-knowledge on a trial-wise basis may help distinguish between these scenarios in the future.

Traditional models of WM emphasised mechanisms by which the brain sustains sensory activity in selective neural populations after the external source of the input is withdrawn. For example, recurrent neural networks with stable attractor states have successfully modelled persistent activity observed at the single neuron level in the monkey prefrontal cortex during WM tasks (Compte et al., 2000; Wimmer et al., 2014). In humans, an influential view, known as the sensory recruitment hypothesis, suggests that information in WM is maintained in the same populations of neurons that code for the information during perception (D'Esposito & Postle, 2015; Scimeca et al., 2018; Serences, 2016). The sensory recruitment hypothesis is supported by observations from multivariate analyses of neuroimaging data showing that fine-grained visual WM representations can be decoded from visual cortex during WM delays and that representational quality predicts subsequent memory precision (Emrich et al., 2013; Ester et al., 2013; Sligte et al., 2009). However, it is not clear how models focusing exclusively on mechanisms for maintaining past sensory input can account for the present behavioural results showing that fine-grained WM representations may be configured as prospective, task-specific templates for decision-making.

It seems likely that retrospective and prospective mechanisms interact to support WM functions. In support of complementary functions of sensory and motor-related representations in WM is evidence suggesting the brain will, when possible, link a sensory representation to the appropriate motor plan in WM early in the memory delay and re-activate these in parallel for decision-making (Boettcher et al., 2020; van Ede,

Chekroud, Stokes, et al., 2019). Perhaps prospective, action-oriented mechanisms of WM are particularly important for speeded decision-making, while retrospective sensory-like representations support memory for precision. This could help explain why task-foreknowledge enhanced the retro-cue effect on reaction times, but not on accuracy in Chapter 4 (Experiment 2). Knowledge of task read-out demands may facilitate a reconfiguration from a pure sensory representation into a prospective task-set, speeding up decision-making, but not necessarily affecting the precision of recall.

According to one theoretical perspective, selecting a WM item for action recruits attention to establish links between the sensory- and action-oriented representations in WM, temporarily enhancing the relevant representation in sensory cortex through feedback from motor areas (Olivers & Roelfsema, 2020). The model predicts that attentional selection further engages reward-based plasticity mechanisms, assigning credit to the selected representation in anticipation of reward by strengthening synaptic connectivity between neurons coding for the item (Roelfsema & Holtmaat, 2018; Roelfsema & van Ooyen, 2005). Selection-for-action may thus engage attention to simultaneously establish sensory-motor links and consolidate memory traces, even in the absence of sustained attention. It represents a possible common framework for how attention could support WM through a combination of enhanced memory traces in sensory areas, as discussed for Chapter 2-3 (Manohar et al., 2019; Rerko & Oberauer, 2013; Souza et al., 2015), as well as action-oriented states optimized for producing the appropriate decision at recall (Myers et al., 2017; Stokes et al., 2020).

Together, the two behavioural experiments presented in Chapter 4 show that retro-cued information in WM triggers task-dependent responses even when irrelevant in the context of a secondary task, indicating prioritized memory items may be configured as prospective task-sets, such that a probe quickly triggers the appropriate response. Prospective task knowledge improves WM-guided behaviour and enhances retro-cued

prioritization, suggesting an adaptive function to storing WM content in a prospective format optimized for the current read-out demands. Until recently, most WM experiments have studied WM within the context of a single task such as change-detection of continuous recall (Bays et al., 2009; Luck & Vogel, 2013; Wilken & Ma, 2004). The current results emphasize the importance of considering task read-out demands when studying the nature of delay representations in WM. This challenges the classic notion that WM is predominantly supported by retrospective representations of sensory history and instead represents a new direction for WM research by shifting focus toward the mechanisms by which WM guides behaviour (A. Heuer et al., 2020; Myers et al., 2017; Olivers & Roelfsema, 2020; Postle, 2006; Sreenivasan, Vytlacil, et al., 2014; Stokes et al., 2020).

Yet, behavioural studies face a fundamental challenge in addressing the nature of delay representations in WM as they rely on indirect measures that may introduce additional demands on cognitive processing or WM load. Consequently, I planned an EEG experiment in collaboration with my co-authors to investigate the format of neural representations under different anticipated task read-out demands, which was the focus of Chapter 5.

Beyond *quantitative* differences in the representational strength of certain features as a function of task-relevance, reported in the literature (LaRocque et al., 2013; Lee et al., 2013; Lewis-Peacock et al., 2012; Mack et al., 2020; Serences et al., 2009), the results of the behavioural experiments presented in Chapter 4 motivate a stronger prediction: the same task-relevant feature in WM is represented in a different *qualitative* format depending on its expected use. If so, a classifier decoding task-relevant memory content from patterns of EEG activity should perform better when it is trained and tested on data where the anticipated read-out is the same (within-task decoding) compared to when it is trained and tested on data from tasks with different read-outs (cross-task

decoding). That is, there should be a cost of cross-generalisation (Lee et al., 2013; Muhle-Karbe et al., 2020; van Loon et al., 2018; Wolff, Jochim, et al., 2020; Q. Yu et al., 2020).

Prior to commencing data collection, I prepared a Stage 1 Registered Report specifying the hypotheses, experimental design and analysis plans (Chapter 5). I specifically expect to find evidence for use-specific coding for the retro-cued item in WM, motivated by previous experimental and theoretical work suggesting only a single item in WM is held in a decision-oriented state (González-García et al., 2020; Olivers et al., 2011; Peters et al., 2009). However, an ensuing question is at which processing stage a potential use-specific configuration occurs: at encoding into WM or once a single item has been selected as relevant, such as following retro-cueing? Either outcome is compatible with the primary hypothesis but teasing apart these scenarios will help elucidate whether use-dependent coding is specific for WM content that has been selected as currently relevant (González-García et al., 2020; J. A. Miller et al., 2020; Myers et al., 2017; Olivers et al., 2011) or is a general property of coding in WM.

While the proposed hypotheses and analyses will help shed light on whether coding in WM depends on anticipated use, they are not designed to reveal exactly *how* the anticipated task alters the underlying neural representation or the precise coding format under different task conditions. Despite active theoretical discussion concerning prospective functional states in WM (A. Heuer et al., 2020; Myers et al., 2017; Nobre & Stokes, 2019; Olivers & Roelfsema, 2020), I am not aware of any empirical studies showing evidence of use-specific coding formats of the *same* task-relevant feature (e.g. orientation in this case). Previous examples of task-dependent coding in WM could be explained in terms of attentional prioritization of different object characteristics depending on the current goal, e.g. colour vs orientation (Serences et al., 2009) or level of abstraction (Lee et al., 2013). Thus, a first empirical test in humans would provide an important piece of evidence either in favour or against a strong version of the use-

specific coding hypothesis, potentially challenging the standard WM model. However, the question of how anticipated use alters the coding format in WM is of great interest and may be followed up in future studies.

In Chapter 5, I proposed one way in which expected use may change the neural state coding for WM content through dynamic coding (Stokes, 2015; Stokes et al., 2013). According to the dynamic coding framework, encoding content into WM alters the way neural circuits involved in memory and decision-making respond to new sensory input (Stokes, 2015). As shown in Chapter 2, neural patterns coding for a stimulus during encoding into WM tend to be qualitatively different from the pattern coding for the same stimulus during a WM delay (Sreenivasan, Curtis, et al., 2014; Stokes, 2015; Wolff et al., 2017), which could reflect a task-specific response profile of the system optimized for context-dependent responses to incoming information. Neural states may function as a matched filter directly interacting with new input to compute the relevant decision, such as the absolute distance between the memory sample and the probe (change-detection) or the direction of rotation (rotation-discrimination)(Hayden & Gallant, 2013; Myers, Rohenkohl, et al., 2015). Consequently, according to a dynamic coding framework, the neural state coding for WM content is not merely a representation of past sensory input, but a prospective template for decision-making, perhaps established through adaptive allocation of attentional gain to neurons tuned to the portion of feature-space most informative to decision-making in the current task context (Navalpakkam & Itti, 2007; Scolari et al., 2012; Scolari & Serences, 2009; X. Yu & Geng, 2019).

An important consideration for Registered Reports is the interpretability of a potential null result. In Chapter 5, factors to consider included (1) the suitability of EEG to answer the proposed hypotheses, (2) statistical power to detect effects of interest, and (3) the suitability of the analysis pipeline to uncover true effects and true null results. I will address these in turn.

First, M/EEG has been shown to be sensitive to detect fine-grained representations in WM (Carlson et al., 2019; Cichy et al., 2015; Grootswagers et al., 2016; Stokes et al., 2015). The high temporal resolution of EEG compared to fMRI makes it ideal for detecting rapid attentional shifts in WM, such as following retro-cueing. I specifically adopted a Mahalanobis distance decoding approach based on previous similar studies detecting neural patterns coding for orientations in WM using EEG (Muhle-Karbe et al., 2020; Wolff, Jochim, et al., 2020; Wolff et al., 2017). Thus, I believe EEG will be suitable to detect differences in the underlying coding format. Combined with a sequential Bayesian analysis approach ensuring high statistical power, a null result would serve as strong evidence against the proposed hypothesis.

A sequential Bayesian analysis framework with an evidence threshold of BF_{10}=10 was adopted to ensure that outcomes of the primary analyses are interpretable, whether they support the null or the alternative hypothesis. Given the anticipated effect size and desired strength of evidence, a Bayes Factor Design Analysis was computed using Monte Carlo simulations. The results were used to estimate the expected sample size and determine a maximum sample size that achieves the desired balance between feasibility and interpretability of the results (Schönbrodt & Wagenmakers, 2018).

Finally, I included simulations of neural data to illustrate the feasibility of the analysis pipeline and ensure that the proposed measure of use-specific coding (cross-generalization cost) was able to detect true effects and true null effects under different scenarios. The simulations raised potential confounds as well as possible ways of controlling for these that would be important to consider for anybody conducting cross-decoding analyses of neuroimaging data. First, there was a concern that a strong univariate and/or multivariate signal coding for the task condition would necessarily produce task-related differences in the decodable neural signal for stimulus orientation. Using simulated neural data, I demonstrated that a task-specific code for orientation

was not trivial, even in the presence of a strong signal coding for the current task-condition.

A second concern regarded potential differences in the level of noise between the two task-conditions. Even if the two task-conditions are well matched, it is possible the EEG signal is noisier in one condition—if participants are less motivated, for example. Reduced signal-to-noise in one condition may result in greater within-condition stimulus decoding in one task condition compared to the other. If so, taking the mean across the two conditions may result in false cross-generalization costs. That is because training and testing within the better task condition is likely to produce superior decoding performance, as compared to training on the worse task condition and testing on the better task condition. The simulations suggested that this false cross-generalization cost is not necessarily balanced out by averaging across the two conditions. Therefore, checking for any differences in within-condition decoding is an important step before computing cross-decoding costs to avoid false positives.

In case of a significant difference in within-task decoding, I adopted a conservative approach and proposed to use the cross-generalization cost for the task with the poorest overall within-task decoding strength, defined per participant. The simulations showed this eliminated the risk of false positives due to differences in signal-to-noise ratio between the two conditions. An alternative approach may be to exclude participants with significant differences in noise-levels between conditions.

Simulations of neural data proved a useful tool for evaluating analysis procedures prior to data collection and data analysis, thereby reducing the risks of confounds and increasing the chances of detecting true effects/null effects. The simulations suggested the cross-decoding procedure proposed in Chapter 5 should be well-suited to test the hypothesis that coding for prioritized content in WM is use-specific.

Conclusion

WM is a highly flexible system that serves to bridge the gap between sensation and action. The aim of this thesis was to investigate how WM contents are selected and prioritized for future action. In a first set of experiments, I examined mechanisms for attentional selection in WM and showed behavioural and neural evidence that attentional enhancement spreads between object features in WM. Selection in WM engages spatial attention, suggesting location information may support retrieval of non-spatial features. Interestingly, despite behavioural benefits of non-predictive cueing, signatures of spatial attention (alpha lateralization and gaze bias) and enhanced representational strength of associated features (location and orientation) were transient rather than sustained following the cue. This is consistent with previous research suggesting sustained internal attention is not needed for retro-cue benefits (Myers et al., 2017; Myers, Walther, et al., 2015; Rerko et al., 2014; Souza & Oberauer, 2016; Wallis et al., 2015). Nevertheless, the cued item was automatically compared with the probe at recall, whether it corresponded to the probed item (congruent) or not (incongruent), suggesting the cued item was in a state of privileged access. Rather than sustained attention, behavioural benefits from incidental cueing may be driven by enhanced connectivity between the features belonging to the cued object even after attention is withdrawn (Manohar et al., 2019; Oberauer & Lin, 2017; Souza & Oberauer, 2016; Souza et al., 2015). Alternatively, cue-induced selection may encourage a re-configuration of the selected content into an action-oriented state optimized for driving decision-making and motor behaviour (A. Heuer et al., 2020; Myers et al., 2017; Olivers & Roelfsema, 2020). Finally, a recent theoretical framework highlighted that these two mechanisms are not mutually exclusive and may both emerge as a result of attentional selection in WM to support prioritization (Olivers & Roelfsema, 2020).

In a second set of experiments, I specifically investigated how WM content is prioritized for action. Behaviourally, a prioritized memory item showed signatures of prepared task-sets. That is, retro-cued items in WM drove responses in a task-specific manner even when irrelevant in the context of a nested delay task. This indicates prioritized WM contents are formatted as prospective templates for decision-making, such that new sensory inputs trigger memory-dependent responses in a reflex-like manner. Configuring selected memory items in such a format may be adaptive as information is generally selected for action (output-gating). Myers et al. (2017) predicted that retro-cue benefits may be supported by a re-configuration of selected items into a prospective action-oriented state. Consistently, the benefit afforded by retro-cues on reaction times was enhanced with prospective knowledge of the task read-out demands compared to when the task-rule was uncertain. Together, the results provide early evidence that retro-cued prioritization in WM is supported by configuration of selected content into a prospective task-specific state that facilitates speeded decision-making. A high-powered EEG study, directly comparing the format of brain representations of the same task-relevant WM content under different anticipated uses, as proposed in Chapter 5, will help elucidate whether neural delay representations in WM are indeed task-specific.

In conclusion, attentional selection of a single feature in WM recruits spatial attention and improves the representation of other features belonging to the same item. However, it is not clear that behavioural cueing benefits in WM are always supported by *sustained* internal attention. Instead, attentional selection may consolidate an item's trace in memory through enhanced connectivity, perhaps placing the item in a privileged position for comparison with the probe (Manohar et al., 2019; Olivers & Roelfsema, 2020; Rerko & Oberauer, 2013; Souza et al., 2015). Additionally, selecting an item for a specific task could enhance sensory-motor links and support the configuration of a prepared task

set, optimized for producing the appropriate action at recall (A. Heuer et al., 2020; Myers et al., 2017; Olivers & Roelfsema, 2020).

Studying attentional states in WM is crucial for understanding how WM bridges the gap between perception and behaviour. This was recognised in early models such as the Multi-Store Model of Memory that considered attention to play a key role in transferring relevant information into short-term storage and preventing information loss (Atkinson & Shiffrin, 1968). The Multicomponent Model of WM highlighted the importance of attention for coordinating information storage and processing in the service of goal-directed behaviour (Baddeley, 2000; Baddeley & Hitch, 1974). On these points the results of this thesis agree.

An area of debate between the Multicomponent model of WM and State-based models of WM is whether WM is supported by dedicated stores or whether WM is best-considered the 'activated portion of long-term memory' (Baddeley, 2012; Cowan, 2008; McElree & Dosher, 1989; Oberauer, 2002). While the experiments presented here were not designed to disambiguate this question, the results tended to agree with theories that highlight the distributed nature of WM and shared processes with perception, long-term memory and motor planning (Fiebig & Lansner, 2017; González-García et al., 2020; Manohar et al., 2019; Myers et al., 2017; Oberauer & Lin, 2017; Olivers & Roelfsema, 2020; Schneegans & Bays, 2017a). Consequently, I favour models that consider WM to operate within the representational structures involved in perception, LTM and action control (Cowan, 1998; Jonides et al., 2008; McElree & Dosher, 1989; Oberauer, 2002; Postle, 2006).

Importantly, attentional selection seems to play a key role in coordinating activity between sensory and motor regions in order to enhance task-relevant representations and configure processing for decision-making and action. Models of WM may benefit

from incorporating mechanisms for attention and action planning in a way that offers a more integrated account of human cognition. Considering the functional purpose of WM will be crucial to future research attempting to understand the mechanisms by which WM supports goal-directed behaviour.

References

Allport, A. (1987). Selection for action: Some beahavioural and neurophysiological considerations of attention and action. In H. Heuer & A. Sanders (Eds.), *Perspectives of Perception and Action* (pp. 395–419). Hilsdale, New Jersey: Erlbaum.

Allport, A., Styles, E. A., & Hsieh, S. (1994). *Shifting intentional set: Exploring the dynamic control of tasks A Family-Centered Intervention Program for Preterm Infants: Effects and Their Biosocial Pathways View project*.

Alvarez, G. A., & 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

American Electroencephalographic Society guidelines for standard electrode position nomenclature. (1991). *Journal of Clinical Neurophysiology : Official Publication of the American Electroencephalographic Society*, *8*(2), 200–202.

Astle, D. E., Summerfield, J., Griffin, I. C., & Nobre, A. C. (2012). Orienting attention to locations in mental representations. *Attention, Perception, and Psychophysics*, *74*(1), 146–162. https://doi.org/10.3758/s13414-011-0218-3

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. *Psychology of Learning and Motivation - Advances in Research and Theory*, *2*(C), 89–195. https://doi.org/10.1016/S0079-7421(08)60422-3

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

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *TRENDS in Cognitive Sciences*, *5*(3).

Awh, E., Jonides, J., & Reuter-Lorenz, P. A. (1998). Rehearsal in spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 780–790. https://doi.org/10.1037//0096-1523.24.3.780

Awh, E., & Vogel, E. K. (2006). Interactions between attention and working memory. *Neuroscience*, *139*, 201–208. https://doi.org/10.1016/j.neuroscience.2005.08.023

Baddeley, A. D. (1992). Working Memory, *255*(5044), 556–559.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*. Elsevier Current Trends. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *Psychology of Learning and Motivation - Advances in Research and Theory*, *8*(C), 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Baddeley, A. D., Jarrold, C., & Vargha-Khadem, F. (2011). Working memory and the hippocampus. *Journal of Cognitive Neuroscience*, *23*(12), 3855–3861. https://doi.org/10.1162/jocn_a_00066

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

Bae, G.-Y., & Luck, S. J. (2017). Dissociable Decoding of Spatial Attention and Working Memory from EEG Oscillations and Sustained Potentials. *The Journal of Neuroscience*, *38*(2), 409–422. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

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), 7.1-11. https://doi.org/10.1167/9.10.7

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

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954-967.e21. https://doi.org/https://doi.org/10.1016/j.cell.2020.09.031

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

Bhandari, A., Gagne, C., & Badre, D. (2018). Just above Chance: Is It Harder to Decode Information from Prefrontal Cortex Hemodynamic Activity Patterns? *Journal of Cognitive Neuroscience*, *30*(10), 1473–1498. https://doi.org/10.1162/jocn_a_01291

Bichot, N. P., Heard, M. T., DeGennaro, E. M., & Desimone, R. (2015). A Source for Feature-Based Attention in the Prefrontal Cortex. *Neuron*, *88*(4), 832–844. https://doi.org/10.1016/j.neuron.2015.10.001

Bichot, N. P., Xu, R., Ghadooshahy, A., Williams, M. L., & Desimone, R. (2019). The role of prefrontal cortex in the control of feature attention in area V4. *Nature Communications*, *10*(1), 5727. https://doi.org/10.1038/s41467-019-13761-7

Bocincova, A., & Johnson, J. S. (2019). The time course of encoding and maintenance of task-relevant versus irrelevant object features in working memory. *Cortex*, *111*, 196–209. https://doi.org/10.1016/j.cortex.2018.10.013

Bocincova, A., Olivers, C. N. L., Stokes, M. G., & Manohar, S. G. (2020). A common neural network architecture for visual search and working memory. *Visual Cognition*, *28*(5–8), 356–371. https://doi.org/10.1080/13506285.2020.1825142

Bocincova, A., van Lamsweerde, A. E., & Johnson, J. S. (2017). The role of top-down suppression in mitigating the disruptive effects of task-irrelevant feature changes in visual working memory. *Memory and Cognition*, *45*(8), 1411–1422. https://doi.org/10.3758/s13421-017-0738-8

Boettcher, S. E. P., Gresch, D., Nobre, A. C., & van Ede, F. (2020). Output Planning at the Input Stage: Action Imprinting for Future Memory-Guided Behaviour. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3550240

Bollimunta, A., Mo, J., Schroeder, C. E., & Ding, M. (2011). Neuronal mechanisms and attentional modulation of corticothalamic alpha oscillations. *Journal of Neuroscience*, *31*(13), 4935–4943. https://doi.org/10.1523/JNEUROSCI.5580-10.2011

Bonnefond, M., & Jensen, O. (2012). Alpha oscillations serve to protect working memory maintenance against anticipated distracters. *Current Biology*, *22*(20), 1969–1974. https://doi.org/10.1016/j.cub.2012.08.029

Bouchacourt, F., & Buschman, T. J. (2019). A Flexible Model of Working Memory. *Neuron*, *0*(0). 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

Braem, S., Deltomme, B., & Liefooghe, B. (2019). The instruction-based congruency effect predicts task execution efficiency: Evidence from inter- and intra-individual differences. *Memory and Cognition*, *47*(8), 1582–1591. https://doi.org/10.3758/s13421-019-00951-3

Brainard, D. H. (1997a). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Brainard, D. H. (1997b). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2017.02.012

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547. https://doi.org/10.1037/0033-295X.97.4.523

Bundesen, C., Habekost, T., & Kyllingsbæk, S. (2005). A neural theory of visual attention: Bridging cognition and neurophysiology. *Psychological Review*, *112*(2), 291–328. https://doi.org/10.1037/0033-295X.112.2.291

Bundesen, C., Habekost, T., & Kyllingsbæk, S. (2011). A neural theory of visual attention and short-term memory (NTVA). *Neuropsychologia*, *49*(6), 1446–1457. https://doi.org/10.1016/j.neuropsychologia.2010.12.006

Bundesen, C., Vangkilde, S., & Petersen, A. (2015). Recent developments in a computational theory of visual attention (TVA). *Vision Research*, *116*, 210–218. https://doi.org/10.1016/j.visres.2014.11.005

Butters, N., & Pandya, D. (1969). Retention of delayed-alternation: Effect of selective lesions of sulcus principalis. *Science*, *165*(3899), 1271–1273. https://doi.org/10.1126/science.165.3899.1271

Cai, Y., Sheldon, A. D., Yu, Q., & Postle, B. R. (2019). Overlapping and distinct contributions of stimulus location and of spatial context to nonspatial visual short-term memory. *J Neurophysiol*, *121*, 1222–1231. https://doi.org/10.1152/jn.00062.2019.-Stim

Carlson, T. A., Grootswagers, T., & Robinson, A. K. (2019). An introduction to time-resolved decoding analysis for M/EEG. *ArXiv*.

Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T., & Kennerley, S. W. (2018). Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nature Communications*, *9*(1), 3498. https://doi.org/10.1038/s41467-018-05873-3

Chafee, M. V., & Goldman-Rakic, P. S. (1998). Matching Patterns of Activity in Primate Prefrontal Area 8a and Parietal Area 7ip Neurons During a Spatial Working MemoryTask. *Journal of Neurophysiology*, *79*(6), 2919–2940. https://doi.org/10.1152/jn.1998.79.6.2919

Chafee, M. V, & Goldman-Rakic, P. S. (2000). Inactivation of parietal and prefrontal cortex reveals interdependence of neural activity during memory-guided saccades. *Journal of Neurophysiology*, *83*(3), 1550–1566. https://doi.org/10.1152/jn.2000.83.3.1550

Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal output gating during selection from working memory. *Neuron*, *81*(4), 930–942. https://doi.org/10.1016/j.neuron.2014.01.002

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*. 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

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, *49*(6), 1407–1409. https://doi.org/10.1016/j.neuropsychologia.2011.01.029

Cichy, R. M., Ramirez, F. M., & Pantazis, D. (2015). Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *NeuroImage*, *121*, 193–204. https://doi.org/10.1016/j.neuroimage.2015.07.011

Cohen, N. J., Ryan, J., Hunt, C., Romine, L., Wszalek, T., & Nash, C. (1999). Hippocampal system and declarative (relational) memory: summarizing the data from functional neuroimaging studies. *Hippocampus*, *9*(1), 83–98. https://doi.org/10.1002/(SICI)1098-1063(1999)9:1<83::AID-HIPO9>3.0.CO;2-7

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (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

Corbetta, M., Kincade, J. M., & Shulman, G. L. (2002). Neural systems for visual orienting and their relationships to spatial working memory. *Journal of Cognitive Neuroscience*, *14*(3), 508–523. https://doi.org/10.1162/089892902317362029

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., & Petersen, S. E. (1991). Selective and divided attention during visual discriminations of shape, color, and speed: Functional anatomy by positron emission tomography. *Journal of Neuroscience*, *11*(8), 2383–2402.

https://doi.org/10.1523/jneurosci.11-08-02383.1991

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

Corkin, S. (2002). What's new with the amnesic patient H.M.? *Nature Reviews Neuroscience*, *3*(2), 153–160. https://doi.org/10.1038/nrn726

Corkin, S., Amaral, D. G., Gilberto González, R., Johnson, K. A., & Hyman, B. T. (1997). H. M.'s Medial temporal lobe lesion: Findings from magnetic resonance imaging. *Journal of Neuroscience*, *17*(10), 3964–3979. https://doi.org/10.1523/jneurosci.17-10-03964.1997

Corneil, B. D., & Munoz, D. P. (2014). Overt responses during covert orienting. *Neuron*, *82*(6), 1230–1243. https://doi.org/10.1016/j.neuron.2014.05.040

Cowan, N. (1998). *Attention and Memory: An Integrated Framework*. *Attention and Memory: An Integrated Framework*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195119107.001.0001

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114-85.

Cowan, N. (2005). *Working Memory Capacity*. *Working Memory Capacity*. Psychology Press. https://doi.org/10.4324/9780203342398

Cowan, N. (2008). *Attention and Memory: An Integrated Framework*. *Attention and Memory: An Integrated Framework*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195119107.001.0001

Cowan, N. (2011). The focus of attention as observed in visual working memory tasks: Making sense of competing claims. *Neuropsychologia*, *49*(6), 1401–1406. https://doi.org/10.1016/j.neuropsychologia.2011.01.035

Curtis, C. E. (2006). Prefrontal and parietal contributions to spatial working memory. *Neuroscience*. https://doi.org/10.1016/j.neuroscience.2005.04.070

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423. https://doi.org/10.1016/S1364-6613(03)00197-9

D'Esposito, M., & Postle, B. R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia*, *37*(11), 1303–1315. https://doi.org/10.1016/S0028-3932(99)00021-4

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

de Vries, I. E. J., van Driel, J., Karacaoglu, M., & Olivers, C. N. L. (2018). Priority Switches in Visual Working Memory are Supported by Frontal Delta and Posterior Alpha Interactions. *Cerebral Cortex*, *28*(11), 4090–4104. https://doi.org/10.1093/cercor/bhy223

de Vries, I. E. J., van Driel, J., & Olivers, C. N. L. (2017). Posterior α EEG Dynamics Dissociate Current from Future Goals in Working Memory-Guided Visual Search. *The Journal of Neuroscience*, *37*(6), 1591–1603. https://doi.org/10.1523/JNEUROSCI.2945-16.2016

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837. https://doi.org/https://doi.org/10.1016/0042-6989(95)00294-4

Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychological Science*, *11*(6), 467–473. https://doi.org/10.1111/1467-9280.00290

Downing, P. E., & Dodds, C. M. (2004). Competition in visual working memory for control of search. *Visual Cognition*, *11*(6), 689–703. https://doi.org/10.1080/13506280344000446

Duncan, J., & Owen, A. M. (2000, October 1). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*. Elsevier.

https://doi.org/10.1016/S0166-2236(00)01633-7

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*. Annu Rev Neurosci. https://doi.org/10.1146/annurev.neuro.30.051606.094328

Emrich, S. M., Riggall, A. C., La Rocque, J. J., & Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *Journal of Neuroscience*, *33*(15), 6516–6523. https://doi.org/10.1523/JNEUROSCI.5732-12.2013

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1

Engbert, R., & Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(18), 7192–7197. https://doi.org/10.1073/pnas.0509557103

Ester, E. F., Anderson, D. E., Serences, J. T., & Awh, E. (2013). A neural measure of precision in visual working memory. *Journal of Cognitive Neuroscience*, *25*(5), 754–761. https://doi.org/10.1162/jocn_a_00357

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

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.1989-16.2017

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12), 3. https://doi.org/10.1167/11.12.3

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12), 1–11. https://doi.org/10.1167/10.12.27

Foxe, J. J., Ca, G. V. S., & Ahlfors, S. P. (1998). Anticipatory State of Visual Attention Mechanisms, *9*(17), 3929–3933.

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: the relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, *17*(5), 673–679. https://doi.org/10.3758/17.5.673

Funahashi, S. (2015). Functions of delay-period activity in the prefrontal cortex and mnemonic scotomas revisited. *Frontiers in Systems Neuroscience*, *9*(February), 2. https://doi.org/10.3389/fnsys.2015.00002

Funahashi, S. (2017). Working memory in the prefrontal cortex. *Brain Sciences*, *7*(5). https://doi.org/10.3390/brainsci7050049

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*, *61*(2), 331–349.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*. https://doi.org/10.1016/j.conb.2016.01.010

Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *Journal of Neurophysiology*, *36*(1), 61–78. https://doi.org/10.1152/jn.1973.36.1.61

Fuster, J. M. (1980). *The Prefrontal Cortex. Anatomy, Physiology and Neuropsychology of the Frontal Lobe*. New York: Raven Press.

Fuster, J. M. (2001). *The Prefrontal Cortex - An Update: Time Is of the Essence*. *Neuron* (Vol. 30).

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

Fuster, J. M., & Jervey, J. P. (1982). Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. *The Journal of Neuroscience*, *2*(3), 361–375.

Gallivan, J. P., Bowman, N. A. R., Chapman, C. S., Wolpert, D. M., & Flanagan, J. R. (2016). The sequential encoding of competing action goals involves dynamic restructuring of motor plans in working memory. *Journal of Neurophysiology*, *115*(6), 3113–3122.

https://doi.org/10.1152/jn.00951.2015

Gallivan, J. P., Logan, L., Wolpert, D. M., & Flanagan, J. R. (2016). Parallel specification of competing sensorimotor control policies for alternative action options. *Nature Neuroscience*, *19*(2), 320–326. https://doi.org/10.1038/nn.4214

Gao, Z., Yu, S., Zhu, C., Shui, R., Weng, X., Li, P., & Shen, M. (2016). Object-based Encoding in Visual Working Memory: Evidence from Memory-driven Attentional Capture. *Scientific Reports*, *6*(1), 22822. https://doi.org/10.1038/srep22822

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2011.11.014

Gazzaley, Adam. (2011). Influence of early attentional modulation on working memory. *Neuropsychologia*, *49*(6), 1410–1424. https://doi.org/10.1016/j.neuropsychologia.2010.12.022

Geng, J. J., Di Quattro, N. E., & Helm, J. (2017). Distractor probability changes the shape of the attentional template. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1993–2007. https://doi.org/10.1037/xhp0000430

Gnadt, J. W., & Andersen, R. A. (1988). Memory related motor planning activity in posterior parietal cortex of macaque. *Experimental Brain Research*, *70*(1), 216–220. https://doi.org/10.1007/BF00271862

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485. https://doi.org/10.1016/0896-6273(95)90304-6

González-García, C., Formica, S., Liefooghe, B., & Brass, M. (2020). Attentional prioritization reconfigures novel instructions into action-oriented task sets. *Cognition*, *194*, 104059. https://doi.org/10.1016/j.cognition.2019.104059

Gosseries, O., Yu, Q., LaRocque, J. J., Starrett, M. J., Rose, N. S., Cowan, N., & Postle, B. R. (2018). Parietal-occipital interactions underlying control- and representation-related processes in working memory for nonspatial visual features. *The Journal of Neuroscience*, 2747–17. https://doi.org/10.1523/JNEUROSCI.2747-17.2018

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, *391*(6666), 481–484. https://doi.org/10.1038/35135

Graham, K. S., Barense, M. D., & Lee, A. C. H. (2010). Going beyond LTM in the MTL: A synthesis of neuropsychological and neuroimaging findings on the role of the medial temporal lobe in memory and perception. *Neuropsychologia*, *48*(4), 831–853. https://doi.org/10.1016/j.neuropsychologia.2010.01.001

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/10.3758/s13414-011-0218-3

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2016). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Cognitive Neuroscience, Journal Of*, *57*, 1–21. https://doi.org/10.1162/jocn

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Haegens, S., Nácher, V., Luna, R., Romo, R., & Jensen, O. (2011). α-Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(48), 19377–19382. https://doi.org/10.1073/pnas.1117190108

Hafed, Z. M., & Clark, J. J. (2002). Microsaccades as an overt measure of covert attention shifts. *Vision Research*, *42*(22), 2533–2545. https://doi.org/10.1016/S0042-6989(02)00263-8

Hajonides, J. E., Nobre, A. C., Van Ede, F., & Stokes, M. G. (2020). Decoding visual colour from scalp electroencephalography measurements. *BioRxiv*, 2020.07.30.228437. https://doi.org/10.1101/2020.07.30.228437

Hajonides, J. E., van Ede, F., Stokes, M. G., & Nobre, A. C. (2019). Comparing the prioritisation of items and feature-dimensions in visual working memory. *BioRxiv*, 863191. https://doi.org/10.1101/863191

Hannula, D. E., Tranel, D., & Cohen, N. J. (2006). The Long and the Short of It: Relational Memory Impairments in Amnesia, Even at Short Lags. *Journal of Neuroscience*, *26*(32), 8352–8359. https://doi.org/10.1523/JNEUROSCI.5222-05.2006

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

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Hayden, B. Y., & Gallant, J. L. (2013). Working memory and decision processes in visual area v4. *Frontiers in Neuroscience*, *7*, 18. https://doi.org/10.3389/fnins.2013.00018

Heinen, K., Feredoes, E., Weiskopf, N., Ruff, C. C., & Driver, J. (2014). Direct Evidence for Attention-Dependent Influences of the Frontal Eye-Fields on Feature-Responsive Visual Cortex. *Cerebral Cortex*, *24*(11), 2815–2821. https://doi.org/10.1093/cercor/bht157

Heuer, A., Ohl, S., & Rolfs, M. (2020). Memory for action: a functional view of selection in visual working memory. *Visual Cognition*, *28*(5–8), 388–400. https://doi.org/10.1080/13506285.2020.1764156

Heuer, A., & Schubö, A. (2016). Feature-based and spatial attentional selection in visual working memory. *Memory and Cognition*, *44*(4), 621–632. https://doi.org/10.3758/s13421-015-0584-5

Higo, T., Mars, R. B., Boorman, E. D., Buch, E. R., & Rushworth, M. F. S. (2011). Distributed and causal influence of frontal operculum in task control. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(10), 4230–4235. https://doi.org/10.1073/pnas.1013361108

Hollingworth, A., & Maxcey-Richard, A. M. (2013). Selective maintenance in visual working memory does not require sustained visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 1047–1058. https://doi.org/10.1037/a0030238

Hommel, B. (2000). The prepared reflex: Automaticity and control in stimulus-response translation. *Attention and Performance*, *18*, 246–273.

IBM Corp. (2016). IBM SPSS Statistics for Macintosh. Armonk, NY: Corp., IBM.

Ikkai, A., & Curtis, C. E. (2011). Common neural mechanisms supporting spatial working memory, attention and motor intention. *Neuropsychologia*, *49*(6), 1428–1434. https://doi.org/10.1016/j.neuropsychologia.2010.12.020

Jacobsen, C. F. (1936). Studies of cerebral function in primates. I. The functions of the frontal association areas in monkeys. *Comparative Psychology Monographs*, *13*, *3*, 1–60.

James, W. (1890). *The Principles of Psychology*. New York: Henry Holt and Company.

JASPTeam. (2018). JASP (Version 0.9.0.1)[Computer software].

Jaswal, S. (2012). The importance of being relevant. *Frontiers in Psychology*, *3*(AUG), 309. https://doi.org/10.3389/fpsyg.2012.00309

Jaswal, S., & Logie, R. H. (2011). Configural encoding in visual feature binding. *Journal of Cognitive Psychology*, *23*(5), 586–603. https://doi.org/10.1080/20445911.2011.570256

Jensen, O., & Mazaheri, A. (2010). Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition. *Frontiers in Human Neuroscience*, *4*(November), 1–8. https://doi.org/10.3389/fnhum.2010.00186

Jeong, S. K., & Xu, Y. (2016). Behaviorally relevant abstract object identity representation in the human parietal cortex. *Journal of Neuroscience*, *36*(5), 1607–1619. https://doi.org/10.1523/JNEUROSCI.1016-15.2016

Jerde, T. A., Merriam, E. P., Riggall, A. C., Hedges, J. H., & Curtis, C. E. (2012). Prioritized Maps of Space in Human Frontoparietal Cortex. *The Journal of Neuroscience*, *32*(48), 17382 LP – 17390. https://doi.org/10.1523/JNEUROSCI.3810-12.2012

Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(3), 683–702. https://doi.org/10.1037/0278-7393.26.3.683

Jin, W., Nobre, A. C., & van Ede, F. (2020). Temporal Expectations Prepare Visual Working Memory for Behavior. *Journal of Cognitive Neuroscience*, 1–12. https://doi.org/10.1162/jocn_a_01626

Johnson, J. S., Hollingworth, A., & Luck, S. J. (2008). The Role of Attention in the Maintenance of Feature Bindings in Visual Short-Term Memory. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 41–55. https://doi.org/10.1037/0096-1523.34.1.41

Johnson, M. R., Mitchell, K. J., Raye, C. L., D'Esposito, M., & Johnson, M. K. (2007). A brief thought can modulate activity in extrastriate visual areas: Top-down effects of refreshing just-seen visual stimuli. *NeuroImage*, *37*(1), 290–299. https://doi.org/10.1016/j.neuroimage.2007.05.017

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*(1), 193–224. https://doi.org/10.1146/annurev.psych.59.103006.093615

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The Reviewing of Object File: Object-Specific Integration of Information. *Cognitive Psychology*, *24*, 175–219.

Kamiński, J., & Rutishauser, U. (2020). Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. *Annals of the New York Academy of Sciences*, *1464*(1), 64–75. https://doi.org/10.1111/nyas.14213

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

Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, *282*(5386), 108–111. https://doi.org/10.1126/science.282.5386.108

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761. https://doi.org/10.1016/S0896-6273(00)80734-5

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*. Annu Rev Neurosci. https://doi.org/10.1146/annurev.neuro.23.1.315

Kayser, J., & Tenke, C. E. (2015). On the benefits of using surface Laplacian (current source density) methodology in electrophysiology. *International Journal of Psychophysiology*, *97*(3), 171–173. https://doi.org/10.1016/j.ijpsycho.2015.06.001

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching-a review. *Psychological Bulletin*, *136*(5), 849–874. https://doi.org/10.1037/a0019842

King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2014.01.002

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic Bulletin and Review*, *20*(2), 228–242. https://doi.org/10.3758/s13423-012-0359-y

Kiyonaga, A., Egner, T., & Soto, D. (2012). Cognitive control over working memory biases of selection. *Psychonomic Bulletin & Review 2012 19:4*, *19*(4), 639–646. https://doi.org/10.3758/S13423-012-0253-7

Klauer, K. C., & Zhao, Z. (2004). Double Dissociations in Visual and Spatial Short-Term Memory. *Journal of Experimental Psychology: General*. Klauer, Karl Christoph: Institut fur Psychologie, Albert-Ludwigs-Universitat Freiburg, Freiburg, Germany, D-79085, christoph.klauer@psychologie.uni-freiburg.de: American Psychological Association. https://doi.org/10.1037/0096-3445.133.3.355

Kuo, B. C., Stokes, M. G., Murray, A. M., & Nobre, A. C. (2014). Attention biases visual activity in visual short-term memory. *Journal of Cognitive Neuroscience*, *26*(7), 1377–1389. https://doi.org/10.1162/jocn_a_00577

Kuo, B. C., Stokes, M. G., & Nobre, A. C. (2012). Attention modulates maintenance of representations in visual short-term memory. *Journal of Cognitive Neuroscience*, *24*(1), 51–60. https://doi.org/10.1162/jocn_a_00087

Kuo, B. C., Yeh, Y. Y., Chen, A. J. W., & D'Esposito, M. (2011). Functional connectivity during top-down modulation of visual short-term memory representations. *Neuropsychologia*, *49*(6),

1589–1596. https://doi.org/10.1016/j.neuropsychologia.2010.12.043

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature*, *384*(6604), 74–77. https://doi.org/10.1038/384074a0

Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, *43*(2), 149–164. https://doi.org/10.1016/S0042-6989(02)00402-9

Lara, A. H., & Wallis, J. D. (2014). Executive control processes underlying multi-item working memory. *Nature Neuroscience*, *17*(6), 876–883. https://doi.org/10.1038/nn.3702

Lara, A. H., & Wallis, J. D. (2015). The Role of Prefrontal Cortex in Working Memory: A Mini Review. *Frontiers in Systems Neuroscience*, *9*, 173. https://doi.org/10.3389/fnsys.2015.00173

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding Attended Information in Short-term Memory: An EEG Study. *Journal of Cognitive Neuroscience*, *25*(1), 127–142. https://doi.org/10.1162/jocn_a_00305

Larocque, J. J., Lewis-Peacock, J. A., & Postle, B. R. (2014). Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience*, *8*, 1–14. https://doi.org/10.3389/fnhum.2014.00005

Larocque, J. J., Riggall, A. C., Emrich, S. M., & Postle, B. R. (2017). Within-Category Decoding of Information in Different Attentional States in Short-Term Memory. *Cerebral Cortex*, *27*, 4881–4890. https://doi.org/10.1093/cercor/bhw283

Lawrence, S. J. D., van Mourik, T., Kok, P., Koopmans, P. J., Norris, D. G., & de Lange, F. P. (2018). Laminar Organization of Working Memory Signals in Human Visual Cortex. *Current Biology*, *28*(21), 3435-3440.e4. https://doi.org/10.1016/j.cub.2018.08.043

Lebedev, M. A., Messinger, A., Kralik, J. D., & Wise, S. P. (2004). Representation of attended versus remembered locations in prefrontal cortex. *PLoS Biology*, *2*(11). https://doi.org/10.1371/journal.pbio.0020365

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*, 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

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*(February), 2. 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

Lepsien, J., & Nobre, A. C. (2007). Attentional modulation of object representations in working memory. *Cerebral Cortex*, *17*(9), 2072–2083. https://doi.org/10.1093/cercor/bhl116

Lepsien, J., Thornton, I., & Nobre, A. C. (2011). Modulation of working-memory maintenance by directed attention. *Neuropsychologia*, *49*(6), 1569–1577. https://doi.org/10.1016/j.neuropsychologia.2011.03.011

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

Li, Q., & Saiki, J. (2014). Different effects of color-based and location-based selection on visual working memory. https://doi.org/10.3758/s13414-014-0775-3

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working Memory Costs of Task Switching. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(3), 478–494. https://doi.org/10.1037/0278-7393.34.3.478

Liefooghe, B., De Houwer, J., & Wenke, D. (2013). Instruction-based response activation depends on task preparation. *Psychonomic Bulletin and Review*, *20*(3), 481–487. https://doi.org/10.3758/s13423-013-0374-7

Liefooghe, B., Wenke, D., & De Houwer, J. (2012). Instruction-based task-rule congruency effects. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(5), 1325–1335. https://doi.org/10.1037/a0028148

Lopes da Silva, F. (1991). Neural mechanisms underlying brain waves: from neural membranes to

networks. *Electroencephalography and Clinical Neurophysiology*. Electroencephalogr Clin Neurophysiol. https://doi.org/10.1016/0013-4694(91)90044-5

Luck, S. J., & Vogel, E. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2013.06.006

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L. L., Buschman, T. J. 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, A. R. (1966). *Higher Cortical Functions in Man*. New York: Basic Books. https://doi.org/10.1007/978-1-4684-7741-2

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience and 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*. https://doi.org/10.1038/nn.3655

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-019-13930-8

Magen, H., Emmanouil, T. A., McMains, S. A., Kastner, S., & Treisman, A. (2009). Attentional demands predict short-term memory load response in posterior parietal cortex. *Neuropsychologia*, *47*(8–9), 1790–1798. https://doi.org/10.1016/j.neuropsychologia.2009.02.015

Makovski, T., & Jiang, Y. V. (2007). Distributing versus focusing attention in visual short-term memory. *Psychonomic Bulletin and Review*, *14*(6), 1072–1078. https://doi.org/10.3758/BF03193093

Makovski, T., Sussman, R., & Jiang, Y. V. (2008). Orienting Attention in Visual Working Memory Reduces Interference From Memory Probes. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(2), 369–380. https://doi.org/10.1037/0278-7393.34.2.369

Makovski, T., Watson, L. M., Koutstaal, W., & Jiang, Y. V. (2010). Method Matters: Systematic Effects of Testing Procedure on Visual Working Memory Sensitivity. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(6), 1466–1479. https://doi.org/10.1037/a0020851

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P., & Husain, M. (2019, June). Neural mechanisms of attending to items in working memory. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2019.03.017

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*, 1159–1167. https://doi.org/10.1038/s41593-019-0414-3

McElree, B., & Dosher, B. A. (1989). Serial Position and Set Size in Short-Term Memory: The Time Course of Recognition. *Journal of Experimental Psychology: General*, *118*(4), 346–373. https://doi.org/10.1037/0096-3445.118.4.346

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*(6), 1423–1442. https://doi.org/10.1037/0278-7393.22.6.1423

Merrikhi, Y., Clark, K., Albarran, E., Parsa, M., Zirnsak, M., Moore, T., & Noudoost, B. (2017). Spatial working memory alters the efficacy of input to visual cortex. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms15041

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, *16*(16), 5154–5167. https://doi.org/10.1523/jneurosci.16-16-05154.1996

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., Kiyonaga, A., Ivry, R. B., & D'Esposito, M. (2020). Prioritized verbal working memory content biases ongoing action. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(12), 1443–1457. https://doi.org/10.1037/xhp0000868

Milner, B. (1966). Amnesia following operation on the temporal lobes. In C. Whitty & O. Zangwill (Eds.), *Amnesia* (pp. 109–133). London: Butterworths.

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*. https://doi.org/10.1016/S1364-6613(03)00028-7

Monsell, S., Sumner, P., & Waters, H. (2003). Task-set reconfiguration with predictable and unpredictable task switches. *Memory and Cognition*, *31*(3), 327–342. https://doi.org/10.3758/BF03194391

Moore, T., & Fallah, M. (2004). Microstimulation of the Frontal Eye Field and Its Effects on Covert Spatial Attention. *Journal of Neurophysiology*, *91*(1), 152–162. https://doi.org/10.1152/jn.00741.2002

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784. https://doi.org/10.1126/science.4023713

Mostert, P., Albers, A. M., Brinkman, L., Todorova, L., Kok, P., & de Lange, F. P. (2018). Eye Movement-Related Confounds in Neural Decoding of Visual Working Memory Representations. *Eneuro*, *5*(4), ENEURO.0401-17.2018. https://doi.org/10.1523/ENEURO.0401-17.2018

Muhle-Karbe, P. S., Duncan, J., De Baene, W., Mitchell, D. J., & Brass, M. (2017). Neural Coding for Instruction-Based Task Sets in Human Frontoparietal and Visual Cortex. *Cerebral Cortex (New York, N.Y. : 1991)*, *27*(3), 1891–1905. https://doi.org/10.1093/cercor/bhw032

Muhle-Karbe, P. S., Myers, N. E., & Stokes, M. G. (2020). A hierarchy of functional states in working memory. *BioRxiv*, 2020.04.16.044511. https://doi.org/10.1101/2020.04.16.044511

Murray, A. M., Nobre, A. C., Clark, I. A., Cravo, A. M., & Stokes, M. G. (2013). Attention Restores Discrete Items to Visual Short-Term Memory. *Psychological Science*, *24*(4), 550–556. https://doi.org/10.1177/0956797612457782

Murray, Alexandra M., Nobre, A. C., & Stokes, M. G. (2011). Markers of preparatory attention predict visual short-term memory performance. *Neuropsychologia*, *49*(6), 1458–1465. https://doi.org/10.1016/j.neuropsychologia.2011.02.016

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), 398–411. https://doi.org/10.1037/xhp0000449

Myers, N. E., Rohenkohl, G., Wyart, V., Woolrich, M. W., Nobre, A. C., & Stokes, M. G. (2015). Testing sensory evidence against mnemonic templates. *ELife*, *4*(DECEMBER2015), 1–25. https://doi.org/10.7554/eLife.09000

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2017.03.010

Myers, N. E., Walther, L., Wallis, G., Stokes, M. G., & Nobre, A. C. (2015). Temporal Dynamics of Attention during Encoding versus Maintenance of Working Memory: Complementary Views from Event-related Potentials and Alpha-band Oscillations. *Journal of Cognitive Neuroscience*, *27*(3), 492–508. https://doi.org/10.1162/jocn_a_00727

Nairne, J. S. (2002). Remembering over the short-term: The Case Against the Standard Model. *Annual Review of Psychology*, *53*, 53–81. https://doi.org/10.1146/annurev.psych.53.100901.135131

Nakamura, K., & Kubota, K. (1995). Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. *Journal of Neurophysiology*, *74*(1), 162–178. https://doi.org/10.1152/jn.1995.74.1.162

Navalpakkam, V., & Itti, L. (2007). Search Goal Tunes Visual Features Optimally. *Neuron*, *53*(4),

605–617. https://doi.org/10.1016/j.neuron.2007.01.018

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A Meta-analysis of Executive Components of Working Memory. *Cerebral Cortex*, *23*(2), 264–282. https://doi.org/10.1093/cercor/bhs007

Nee, D. E., & Jonides, J. (2008). Neural correlates of access to short-term memory. *Proceedings of the National Academy of Sciences*, *105*(37), 14228–14233. https://doi.org/10.1073/pnas.0802081105

Nee, D. E., & Jonides, J. (2011). Dissociable contributions of prefrontal cortex and the hippocampus to short-term memory: Evidence for a 3-state model of memory. *NeuroImage*, *54*(2), 1540–1548. https://doi.org/10.1016/j.neuroimage.2010.09.002

Nee, D. E., & Jonides, J. (2013a). Neural evidence for a 3-state model of visual short-term memory. *NeuroImage*, *74*, 1–11. https://doi.org/10.1016/j.neuroimage.2013.02.019

Nee, D. E., & Jonides, J. (2013b). Trisecting representational states in short-term memory. *Frontiers in Human Neuroscience*, *7*(NOV). https://doi.org/10.3389/fnhum.2013.00796

Nemrodov, D., Niemeier, M., Patel, A., & Nestor, A. (2018). The neural dynamics of facial identity processing: Insights from EEG-based pattern analysis and image reconstruction. *ENeuro*, *5*(1), ENEURO.0358-17.2018. https://doi.org/10.1523/ENEURO.0358-17.2018

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. https://doi.org/10.1038/srep42384

Nobre, A. C., Coull, J. T., Maquet, P., Frith, C. D., Vandenberghe, R., & Mesulam, M. M. (2004). Orienting Attention to Locations in Perceptual Versus Mental Representations. *Journal of Cognitive Neuroscience*, *16*(3), 363–373. https://doi.org/10.1162/089892904322926700

Nobre, A. C., Sebestyen, G. N., Gitelman, D. R., Mesulam, M. M., Frackowiak, R. S. J., & Frith, C. D. (1997). Functional localization of the system for visuospatial attention using positron emission tomography. *Brain*, *120*(3), 515–533. https://doi.org/10.1093/brain/120.3.515

Nobre, A. C., & Stokes, M. G. (2019). Premembering Experience: A Hierarchy of Time-Scales for Proactive Attention. *Neuron*, *104*(1), 132–146. https://doi.org/10.1016/j.neuron.2019.08.030

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

Oberauer, K. (2001). Removing Irrelevant Information from Working Memory: A Cognitive Aging Study with the Modified Sternberg Task. *Journal of Experimental Psychology: Learning Memory and Cognition*, *27*(4), 948–957. https://doi.org/10.1037/0278-7393.27.4.948

Oberauer, K. (2002). Access to Information in Working Memory: Exploring the Focus of Attention. *Journal of Experimental Psychology: Learning Memory and Cognition*, *28*(3), 411–421. https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K. (2010). Declarative and Procedural Working Memory: Common Principles, Common Capacity Limits? *Psychologica Belgica*, *50*(3–4), 277. https://doi.org/10.5334/pb-50-3-4-277

Oberauer, K., & Eichenberger, S. (2013). Visual working memory declines when more features must be remembered for each object. *Memory & Cognition*, *41*(8), 1212–1227. https://doi.org/10.3758/s13421-013-0333-6

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. (2008). Forgetting in immediate serial recall: decay, temporal distinctiveness, or interference? *Psychological Review*, *115*(3), 544–576. https://doi.org/10.1037/0033-295X.115.3.544

Oberauer, K., & Lin, H.-Y. (2017). An Interference Model of Visual Working Memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Olivers, C. N. L., Meijer, F., & Theeuwes, J. (2006). Feature-based memory-driven attentional capture: Visual working memory content affects visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1243–1265. https://doi.org/10.1037/0096-1523.32.5.1243

Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual

working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*(7), 327–334. https://doi.org/10.1016/j.tics.2011.05.004

Olivers, C. N. L., & Roelfsema, P. R. (2020). Attention for action in visual working memory. *Cortex*. https://doi.org/10.1016/j.cortex.2020.07.011

Olsen, R. K., Moses, S. N., Riggs, L., & Ryan, J. D. (2012). The hippocampus supports multiple cognitive processes through relational binding and comparison. *Frontiers in Human Neuroscience*, *6*, 146. https://doi.org/10.3389/fnhum.2012.00146

Olson, I. R., & Jiang, Y. (2002). Is visual short-term memory object based? Rejection of the "strong-object" hypothesis. *Perception & Psychophysics*, *64*(7), 1055–1067. https://doi.org/10.3758/bf03194756

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 1–9. https://doi.org/10.1155/2011/156869

Oztekin, I., Davachi, L., & McElree, B. (2010). Are representations in working memory distinct from representations in long-term memory? Neural evidence in support of a single store. *Psychological Science*, *21*(8), 1123–1133. https://doi.org/10.1177/0956797610376651

Panichello, M., & Buschman, T. (2020). Selective control of working memory in prefrontal, parietal, and visual cortex. *BioRxiv*, 2020.04.07.030718. https://doi.org/10.1101/2020.04.07.030718

Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S. C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. *Nature Neuroscience*, *20*(12), 1770–1779. https://doi.org/10.1038/s41593-017-0003-2

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

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*(2), 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1224–1231. https://doi.org/10.1037/a0030947

Pertzov, Y., & Husain, M. (2014). The privileged role of location in visual working memory. *Attention, Perception, & Psychophysics*, *76*(7), 1914–1924. https://doi.org/10.3758/s13414-013-0541-y

Pertzov, Y., Miller, T. D., Gorgoraptis, N., Caine, D., Schott, J. M., Butler, C., & Husain, M. (2013). Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain*, *136*(8), 2474–2485. https://doi.org/10.1093/brain/awt129

Peters, J. C., Goebel, R., & Roelfsema, P. R. (2009). Remembered but Unused: The Accessory Items in Working Memory that Do Not Guide Attention. *Journal of Cognitive Neuroscience*, *21*(6), 1081–1091. https://doi.org/10.1162/jocn.2009.21083

Poch, C., Campo, P., & Barnes, G. R. (2014). Modulation of alpha and gamma oscillations related to retrospectively orienting attention within working memory. *European Journal of Neuroscience*, *40*(2), 2399–2405. https://doi.org/10.1111/ejn.12589

Poch, C., Capilla, A., Hinojosa, J. A., & Campo, P. (2017). Selection within working memory based on a color retro-cue modulates alpha oscillations. *Neuropsychologia*, *106*, 133–137. https://doi.org/10.1016/j.neuropsychologia.2017.09.027

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38. https://doi.org/10.1016/j.neuroscience.2005.06.005

Printzlau, F., Myers, N. E., Muhle-Karbe, P. S., Manohar, S. G., & Stokes, M. G. (2019). Prospective task knowledge improves working memory-guided behaviour. *PsyArxiv*. https://doi.org/10.31234/OSF.IO/QHKU6

Quax, S. C., Dijkstra, N., van Staveren, M. J., Bosch, S. E., & van Gerven, M. A. J. (2019). Eye movements explain decodability during perception and cued attention in MEG. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2019.03.069

R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna: R

Foundation for Statistical Computing.

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

Rainer, G., Rao, S. C., & Miller, E. K. (1999). Prospective coding for objects in primate prefrontal cortex. *J Neurosci*, *19*(13), 5493–5505.

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

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481. https://doi.org/10.3758/bf03196302

Rerko, L., & Oberauer, K. (2013). Focused, unfocused, and defocused information in working memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(4), 1075–1096. https://doi.org/10.1037/a0031172

Rerko, L., Souza, A. S., & Oberauer, K. (2014). Retro-cue benefits in working memory without sustained focal attention. *Memory and Cognition*, *42*(5), 712–728. https://doi.org/10.3758/s13421-013-0392-8

Richter, F. R., & Yeung, N. (2015). Neuroimaging Studies of Task Switching. In *Task Switching and Cognitive Control* (pp. 237–271). https://doi.org/10.1093/acprof:osobl/9780199921959.003.0010

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), 1–6. https://doi.org/10.1038/nature12160

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1 PART 1), 31–40. https://doi.org/10.1016/0028-3932(87)90041-8

Roelfsema, P. R., & Holtmaat, A. (2018). Control of synaptic plasticity in deep cortical networks. *Nature Reviews. Neuroscience*, *19*(3), 166–180. https://doi.org/10.1038/nrn.2018.6

Roelfsema, P. R., & van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural Computation*, *17*(10), 2176–2214. https://doi.org/10.1162/0899766054615699

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139. https://doi.org/10.1126/science.aah7011

Rossi, A. F., Bichot, N. P., Desimone, R., & Ungerleider, L. G. (2007). Top–Down Attentional Deficits in Macaques with Lesions of Lateral Prefrontal Cortex. *The Journal of Neuroscience*, *27*(42), 11306 LP – 11314. https://doi.org/10.1523/JNEUROSCI.2939-07.2007

Rubin, O., & Meiran, N. (2005). On the Origins of the Task Mixing Cost in the Cuing Task-Switching Paradigm. *Journal of Experimental Psychology*, *31*(6), 1477–1491. https://doi.org/10.1037/0278-7393.31.6.1477

Rushworth, M. F. S., Hadland, K. A., Paus, T., & Sipila, P. K. (2002). Role of the human medial frontal cortex in task switching: a combined fMRI and TMS study. *Journal of Neurophysiology*, *87*(5), 2577–2592. https://doi.org/10.1152/jn.2002.87.5.2577

Sakai, K. (2008). Task Set and Prefrontal Cortex. *Annual Review of Neuroscience*, *31*(1), 219–245. https://doi.org/10.1146/annurev.neuro.31.060407.125642

Schneegans, S., & Bays, P. M. (2017a). 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. (2017b). Restoration of fMRI Decodability Does Not Imply Latent Working Memory States. *Journal of Cognitive Neuroscience*, *26*(3), 1–18. https://doi.org/10.1162/jocn_a_01180

Schneider, D., Barth, A., & Wascher, E. (2017). On the contribution of motor planning to the retroactive cuing benefit in working memory: Evidence by mu and beta oscillatory activity in the EEG. *NeuroImage*, *162*, 73–85. https://doi.org/10.1016/j.neuroimage.2017.08.057

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Scimeca, J. M., Kiyonaga, A., & D'Esposito, M. (2018). Reaffirming the Sensory Recruitment Account of Working Memory. *Trends in Cognitive Sciences*, *22*(3), 190–192. https://doi.org/10.1016/j.tics.2017.12.007

Scolari, M., Byers, A., & Serences, J. T. (2012). Optimal deployment of attentional gain during fine discriminations. *Journal of Neuroscience*, *32*(22), 7723–7733. https://doi.org/10.1523/JNEUROSCI.5558-11.2012

Scolari, M., & Serences, J. T. (2009). Adaptive allocation of attentional gain. *Journal of Neuroscience*, *29*(38), 11933–11942. https://doi.org/10.1523/JNEUROSCI.5642-08.2009

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*(1), 11–21. https://doi.org/10.1136/jnnp.20.1.11

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

Shafi, M. M., Zhou, Y., Quintana, J., Chow, C. C., Fuster, J. M., & Bodner, M. (2007). Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience*, *146*(3), 1082–1108. https://doi.org/10.1016/j.neuroscience.2006.12.072

Shallice, T., & Warrington, E. K. (1974). The dissociation between short term retention of meaningful sounds and verbal material. *Neuropsychologia*, *12*(4), 553–555. https://doi.org/10.1016/0028-3932(74)90087-6

Shen, M., Tang, N., Wu, F., Shui, R., & Gao, Z. (2013). Robust object-based encoding in visual working memory. *Journal of Vision*, *13*(2), 1–1. https://doi.org/10.1167/13.2.1

Shomstein, S., & Gottlieb, J. (2016). Spatial and non-spatial aspects of visual attention: Interactive cognitive mechanisms and neural underpinnings. *Neuropsychologia*, *92*, 9–19. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2016.05.021

Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2009). V4 activity predicts the strength of visual short-term memory representations. *Journal of Neuroscience*, *29*(23), 7432–7438. https://doi.org/10.1523/JNEUROSCI.0784-09.2009

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*. Netherlands: Elsevier Science. https://doi.org/10.1006/cogp.1997.0658

Smyth, M. M., & Scholey, K. A. (1994). Interference in immediate spatial memory. *Memory & Cognition*, *22*(1), 1–13. https://doi.org/10.3758/BF03202756

Soto, D., Heinke, D., Humphreys, G. W., & Blanco, M. J. (2005). Early, involuntary top-down guidance of attention from working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(2), 248–261. https://doi.org/10.1037/0096-1523.31.2.248

Soto, D., Hodsoll, J., Rotshtein, P., & Humphreys, G. W. (2008). Automatic guidance of attention from working memory. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2008.05.007

Soto, D., & Humphreys, G. W. (2009). Automatic selection of irrelevant object features through working memory evidence for top-down attentional capture. *Experimental Psychology*, *56*(3), 165–172. https://doi.org/10.1027/1618-3169.56.3.165

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, *78*(7), 1–22. https://doi.org/10.3758/s13414-016-1108-5

Souza, A. S., Rerko, L., & Oberauer, K. (2014). Unloading and reloading working memory: Attending to one item frees capacity. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1237–1256. https://doi.org/10.1037/a0036331

Souza, A. S., Rerko, L., & Oberauer, K. (2015). Refreshing memory traces: Thinking of an item improves retrieval from visual working memory. *Annals of the New York Academy of Sciences*, *1339*(1), 20–31. https://doi.org/10.1111/nyas.12603

Souza, A. S., Rerko, L., & Oberauer, K. (2016). Getting More From Visual Working Memory: Retro-Cues Enhance Retrieval and Protect From Visual Interference. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(6), 890–910. https://doi.org/10.1037/xhp0000192

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29. https://doi.org/10.1037/h0093759

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

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *The Journal of Neuroscience*, *37*(27), 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, *253*(5026), 1380 LP – 1386. https://doi.org/10.1126/science.1896849

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Sreenivasan, K. K., Vytlacil, J., & D'Esposito, M. (2014). Distributed and Dynamic Storage of Working Memory Stimulus Information in Extrastriate Cortex. *Journal of Cognitive Neuroscience*, *26*(5), 1141–1153. https://doi.org/10.1162/jocn

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

Stokes, M. G., Muhle-Karbe, P. S., & Myers, N. E. (2020). Theoretical distinction between functional states in working memory and their corresponding neural states. *Visual Cognition*, *28*(5–8), 420–432. https://doi.org/10.1080/13506285.2020.1825141

Stokes, M. G., & Spaak, E. (2016). The Importance of Single-Trial Analyses in Cognitive Neuroscience. *Trends in Cognitive Sciences*, *20*(7), 483–486. https://doi.org/10.1016/j.tics.2016.05.008

Stokes, M. G., Wolff, M. J., & Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution. *Trends in Cognitive Sciences*, *19*(11), 636–638. https://doi.org/10.1016/j.tics.2015.08.016

Sugase-Miyamoto, Y., Liu, Z., Wiener, M. C., Optican, L. M., & Richmond, B. J. (2008). Short-Term Memory Trace in Rapidly Adapting Synapses of Inferior Temporal Cortex. *PLoS Comput Biol*, *4*(5). https://doi.org/10.1371/journal.pcbi.1000073

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

Tabi, Y. A., Husain, M., & Manohar, S. G. (2019). Recall cues interfere with retrieval from visuospatial working memory. In *British Journal of Psychology* (Vol. 110, pp. 288–305). John Wiley and Sons Ltd. https://doi.org/10.1111/bjop.12374

Tabi, Y., Bocincova, A., Husain, M., & Manohar, S. G. (n.d.). Recency but not retro-cued shifts of attention protects working memory from interference.

Takeda, K., & Funahashi, S. (2002). Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks. *Journal of Neurophysiology*, *87*(1), 567–588. https://doi.org/10.1152/jn.00249.2001

Tamber-Rosenau, B. J., Esterman, M., Chiu, Y. C., & Yantis, S. (2011). Cortical mechanisms of cognitive control for shifting attention in vision and working memory. *Journal of Cognitive Neuroscience*, *23*(10), 2905–2919. https://doi.org/10.1162/jocn.2011.21608

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. https://doi.org/10.1016/j.visres.2012.10.012

The MathWorks Inc. (2015). MATLAB and Statistics Tookbox Release 2015b. Natick, Massachusetts, United States.

The MathWorks Inc. (2017). MATLAB and Statistics Toolbox Release 2017b. Natick, Massachusetts, United States.

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

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective and Behavioral Neuroscience*, *5*(2), 144–155. https://doi.org/10.3758/CABN.5.2.144

Treisman, A. M. (1998). The perception of features and objects. In R. D. Wright (Ed.), *Vancouver studies in cognitive science, Vol. 8. Visual attention* (pp. 26–54). Oxford University Press.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Treisman, A. M., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*(1), 107–141. https://doi.org/10.1016/0010-0285(82)90006-8

Treisman, A. M., & Zhang, W. (2006). Location and binding in visual working memory. *Memory & Cognition*, *34*(8), 1704–1719. https://doi.org/10.3758/BF03195932

Trübutschek, D., Marti, S., Ojeda, A., King, J. R., Mi, Y., Tsodyks, M., & Dehaene, S. (2017). A theory of working memory without consciousness or sustained activity. *ELife*, *6*, e23871. https://doi.org/10.7554/eLife.23871

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26. https://doi.org/10.1016/j.cogpsych.2014.01.003

Van Ede, F., Board, A. G., & Nobre, A. C. (2020). Goal-directed and stimulus-driven selection of internal representations. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(39), 24590–24598. https://doi.org/10.1073/pnas.2013432117

van Ede, F., Chekroud, S. R., & Nobre, A. C. (2019, May 1). Human gaze tracks attentional focusing in memorized visual space. *Nature Human Behaviour*. Nature Publishing Group. https://doi.org/10.1038/s41562-019-0549-y

van Ede, F., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nature Neuroscience*. https://doi.org/10.1038/s41593-018-0335-6

van Ede, F., Niklaus, M., & Nobre, A. C. (2017). Temporal Expectations Guide Dynamic Prioritization in Visual Working Memory through Attenuated α Oscillations. *The Journal of Neuroscience*, *37*(2), 437–445. https://doi.org/10.1523/JNEUROSCI.2272-16.2016

van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J., & Olivers, C. N. (2018). Current and future goals are represented in opposite patterns in object-selective cortex. *ELife*, *7*. https://doi.org/10.7554/elife.38677

Van Moorselaar, D., Olivers, C. N. L., Theeuwes, J., Lamme, V. A. F., & Sligte, I. G. (2015). Forgotten but not gone: Retro-cue costs and benefits in a double-cueing paradigm suggest multiple states in visual short-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(6), 1755–1763. https://doi.org/10.1037/xlm0000124

Vanni, S., Revonsuo, A., & Hari, R. (1997). Modulation of the parieto-occipital alpha rhythm during object detection. *Journal of Neuroscience*, *17*(18), 7141–7147. https://doi.org/10.1523/jneurosci.17-18-07141.1997

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

Vergauwe, E., & Langerock, N. (2017). Attentional refreshing of information in working memory: Increased immediate accessibility of just-refreshed representations. *Journal of Memory and Language*, *96*, 23–35. https://doi.org/10.1016/j.jml.2017.05.001

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., Mccollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*(7067), 500–503. https://doi.org/Doi 10.1038/Nature04171

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 LP – 18172. https://doi.org/10.1073/pnas.1007277107

Wager, T., & Smith, E. (2004). Neuroimaging studies of working memory: A meta-analysis. *Cognitive, Affective & Behavioral Neuroscience*, *3*, 255–274. https://doi.org/10.3758/CABN.3.4.255

Wallis, G., Stokes, M. G., Cousijn, H., Woolrich, M. G., & Nobre, A. C. (2015). Frontoparietal and cingulo-opercular networks play dissociable roles in control of working memory. *Journal of Cognitive Neuroscience*, *27*(10), 2019–2034. https://doi.org/10.1162/jocn_a_00838

Wang, B., Cao, X., Theeuwes, J., & Olivers, C. N. L. (2016). Location-based effects underlie feature conjunction benefits in visual working memory. *Journal of Vision*, *16*(11), 1–13. https://doi.org/10.1167/16.11.12

Wang, X. J. (2001). Synaptic reverberations underlying mnemonic persistent activity. *Trends in Neurosciences*, *24*(8), 455–463. https://doi.org/10.1016/S0166-2236(00)01868-3

Warden, M. R., & Miller, E. K. (2010). Task-Dependent Changes in Short-Term Memory in the Prefrontal Cortex. *The Journal of Neuroscience*, *30*(47), 15801–15810. https://doi.org/10.1523/JNEUROSCI.1569-10.2010

Warrington, E. K., & Shallice, T. (1969). The selective impairment of auditory verbal short-term memory. *Brain*, *92*(4), 885–896. https://doi.org/10.1093/brain/92.4.885

Watanabe, K., & Funahashi, S. (2014a). Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature Neuroscience*, *17*(4), 601–611. https://doi.org/10.1038/nn.3667

Watanabe, K., & Funahashi, S. (2014b). Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature NeurOSCIenCe*, *17*(6). https://doi.org/10.1038/nn.3667

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

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135. https://doi.org/10.1167/4.12.11

Williams, M., Hong, S. W., Kang, M. S., Carlisle, N. B., & Woodman, G. F. (2013). The benefit of forgetting. *Psychonomic Bulletin and Review*, *20*(2), 348–355. https://doi.org/10.3758/s13423-012-0354-3

Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, *17*(3), 431–439. https://doi.org/10.1038/nn.3645

Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience*, *9*(5), 1412–1427. https://doi.org/10.3389/fnsys.2015.00123

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.

Wolff, M. J., Kandemir, G., Stokes, M. G., & Akyürek, E. G. (2020a). Unimodal and Bimodal Access to Sensory Working Memories by Auditory and Visual Impulses. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *40*(3), 671–681. https://doi.org/10.1523/JNEUROSCI.1194-19.2019

Wolff, M. J., Kandemir, G., Stokes, M. G., & Akyürek, E. G. (2020b). Unimodal and bimodal access to sensory working memories by auditory and visual impulses. *Journal of Neuroscience*, *40*(3), 671–681. https://doi.org/10.1523/JNEUROSCI.1194-19.2019

Wood, J. N. (2011). When do spatial and visual working memory interact? *Attention, Perception, and Psychophysics*, *73*(2), 420–439. https://doi.org/10.3758/s13414-010-0048-8

Woodman, G. F., & Luck, S. J. (2007). Do the Contents of Visual Working Memory Automatically Influence Attentional Selection During Visual Search? *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 363–377. https://doi.org/10.1037/0096-1523.33.2.363

Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin and Review*, *15*(1), 223–229. https://doi.org/10.3758/PBR.15.1.223

Worden, M. S., Foxe, J. J., Wang, N., & Simpson, G. V. (2000). Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *The Journal of Neuroscience*, *20*(6), RC63. https://doi.org/Rc63

Xu, Y. (2010). The neural fate of task-irrelevant features in object-based processing. *Journal of Neuroscience*, *30*(42), 14020–14028. https://doi.org/10.1523/JNEUROSCI.3011-10.2010

Xu, Y. (2020). Revisit once more the sensory storage account of visual working memory. *Visual Cognition*, *28*(5–8), 433–446. https://doi.org/10.1080/13506285.2020.1818659

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

Yeung, N., & Monsell, S. (2003). Switching Between Tasks of Unequal Familiarity: The Role of Stimulus-Attribute and Response-Set Selection. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 455–469. https://doi.org/10.1037/0096-1523.29.2.455

Yi, D. J., Turk-Browne, N. B., Chun, M. M., & Johnson, M. K. (2008). When a thought equals a look: Refreshing enhances perceptual memory. *Journal of Cognitive Neuroscience*, *20*(8), 1371–1380. https://doi.org/10.1162/jocn.2008.20094

Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLoS Biology*, *18*(6), e3000769. https://doi.org/10.1371/journal.pbio.3000769

Yu, X., & Geng, J. J. (2019). The attentional template is shifted and asymmetrically sharpened by distractor context. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(3), 336–353. https://doi.org/10.1037/xhp0000609

Zanto, T. P., & Gazzaley, A. (2009). Neural Suppression of Irrelevant Information Underlies Optimal Working Memory Performance. *Journal of Neuroscience*, *29*(10), 3059–3066. https://doi.org/10.1523/JNEUROSCI.4621-08.2009

Zanto, T. P., Rubens, M. T., Bollinger, J., & Gazzaley, A. (2010). Top-down modulation of visual feature processing: The role of the inferior frontal junction. *NeuroImage*, *53*(2), 736–745. https://doi.org/10.1016/j.neuroimage.2010.06.012

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

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*(November), 1–12. https://doi.org/10.3389/fnhum.2014.00853

Zola-Morgan, S., Squire, L. R., & Amaral, D. G. (1986). Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field CA1 of the hippocampus. *The Journal of Neuroscience*, *6*(10), 2950 LP – 2967. https://doi.org/10.1523/JNEUROSCI.06-10-02950.1986

Appendix A. Supplementary material for Chapter 2

Regressing out decoding from eye-position

As the memorised location of the WM items were decodable from the eye-position data, small eye-movements may have confounded the EEG decoding analyses, even when limiting analyses to posterior channels (Mostert et al., 2018; Quax et al., 2019).

To test the extent to which location information from the eye-position predicted the location information decodable from raw EEG, I ran a general linear model for each time-point with each subject's trial-wise location decoding from eye-position as the predictor variable and trial-wise location decoding from EEG channels as the outcome variable. I obtained the regression coefficient (β weights) and the residual EEG decoding values after regressing out decoding from eye-position. I first tested for any significant correlations using cluster-corrected t-tests over the post-cue time-window (0-1 s; Figure S1A). While there was an increase in beta weights following cue-onset, this was not significant.

I took the mean decoding accuracy for eye-position and raw EEG across the post cue time window (0-1 s) for each subject and ran a general linear model to test if the mean eye-position decoding within the time window of interest predicted the raw EEG decoding within the same window and tested for a significant correlation. There was no significant correlation between time-averaged eye decoding and time-averaged EEG decoding in the post-cue window as revealed by a one-sample t-test (one-tailed, >0; t_{29}= 1.45, p=.079).

However, as there was a trend for a correlation, I next tested whether EEG decoding remained significant after regressing out decoding from eye-position and whether residual decoding was significantly worse than original decoding accuracy (Figure S1B). In the time-window of interest following the cue, residual decoding of location from EEG

remained significant from 0.081 to 1.25 s (p<.001) and there was no significant reduction in decoding accuracy compared to the original decoding values. Similarly, for the time-averaged regression, residual decoding accuracy remained significant (t_{29}= 5.60, p<.001, d=1.02) and there was no significant reduction in decoding accuracy as compared to original decoding (p=.157).

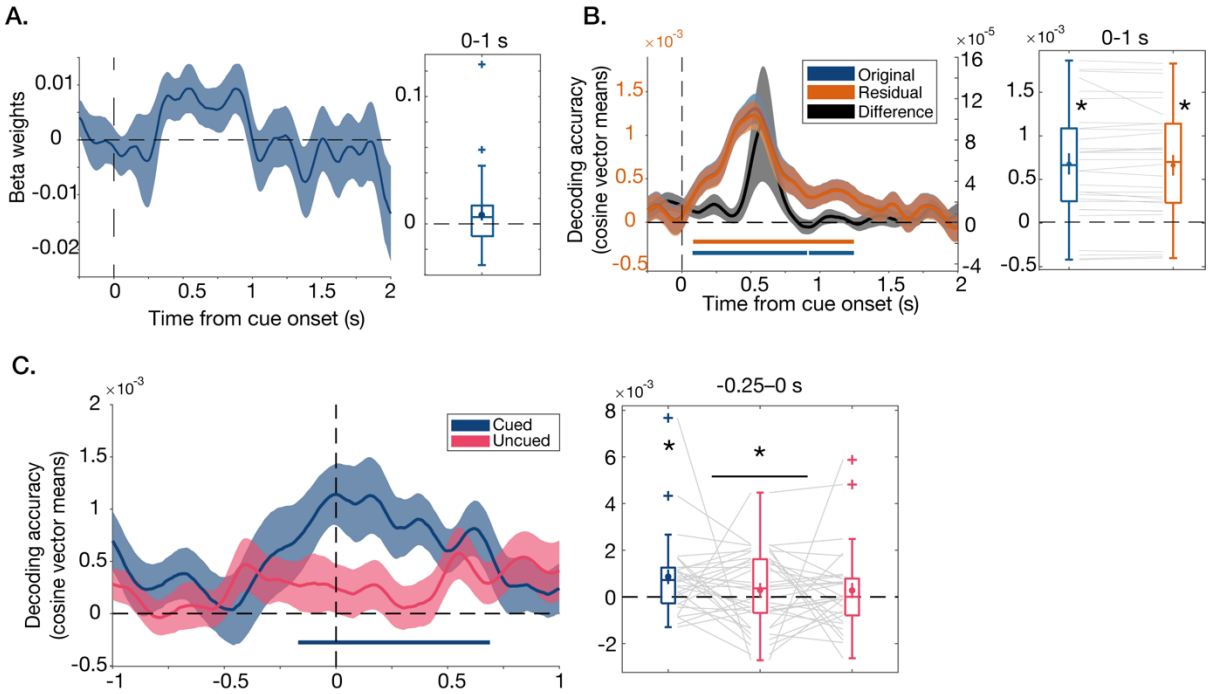

Figure S1. Location decoding from EEG and eye-position. **A.** Left: Mean beta weights from time-resolved regression analysis between trial-wise location decoding from EEG and from eye-position. Shaded areas show standard error of mean (SEM). Right: Boxplot of beta weights for regression analysis across time-window of interest (TOI) following cue onset (0-1s). Mean and SEM overlaid. **B.** Original location decoding from EEG (blue) and residual location decoding after regressing out location decoding from eye-position (orange). Left: Difference between original and residual location decoding is overlaid with separate y-axis (black, note the difference in scale). Right: Blue and orange bars show clusters of significant decoding. Boxplot shows mean original EEG location decoding (blue) and residuals (orange) across the post-cue TOI (0-1s). Grey lines show individual participants. **C.** Mean location decoding from EEG for cued (blue) and uncued (pink) items time-locked to the first microsaccade post-cue (t=0). Blue bar shows cluster of significant decoding of the cued item. Boxplot shows mean location decoding across the time-window before microsaccade onset (-0.25-0s). *p<.05.

Location decoding from EEG time-locked to saccade onset

Next, to test whether location decoding from EEG was time-locked to the onset of microsaccades post-cue, trial-wise EEG data were re-epoched to the onset of the first micro-saccade after the cue on each trial (see Chapter 2 for micro-saccade detection

methods). Post-cue micro-saccades were not present on every trial and trials with no saccades detected were excluded. Across participants the mean number of usable trials for this analysis was 431.93 (min=168, max=686). Otherwise the location decoding procedure followed the same steps as described in Chapter 2. The analysis was focused on the time-window from -1 to 1 second relative to the onset of the first post-cue micro-saccade on each trial.

Figure S1C shows time-resolved location decoding (Gaussian smoothing kernel SD=12.5 ms). Cluster corrected t-tests from -1 to 1 s revealed a cluster of significant decoding for the cued item's location from -0.67 to 0.69 s (p=.001). There was a trend for a significant difference between cued and uncued location decoding from -0.08 s to 0.17 s (p=.085). To ensure pre-saccade decoding was not due to smoothing, I also took the mean unsmoothed decoding accuracy from -0.25 to 0 s before saccade onset. In the pre-saccade window, the location of the cued item could be decoded (t_{29}=2.72, p=.006, d=.50), but the location of neither of the two uncued items could be decoded (uncued1: p=.170, uncued2: p=.205). A one-way ANOVA did not reveal any difference in decoding accuracy between the items (p=.280). Pairwise comparisons between the cued and uncued items, showed no significant difference between cued item location decoding and one of the uncued items (p=.114), but there was a significant difference between the cued item and the other uncued item (t_{29}=1.71, p=.049, d=.313). There was no significant difference between the two uncued items (p=.963).

In summary, while eye-movement related artefacts may contribute to WM location decoding from EEG, onset of location decoding for the cued item from EEG preceded the first post-cue micro-saccade.

Alpha lateralization as a function of relative position of memory items

To explore alpha lateralization as a function of the relative position of the items in memory, the alpha lateralization index was split according to the original position of the two uncued items in memory in relation to the *cued* item (Figure S2A). This resulted in three conditions: both uncued items were in the same hemifield as the cued item (load 2), one uncued item was in the same hemifield as the cued item (load 1), or the two uncued items were in the opposite hemifield (load 0). The results of the statistical analyses are presented in Chapter 2 and in Figure S2. In summary, the reversal in alpha lateralization after the completion of the incidental task (1-1.8 s) was only significant when at least one of the other memory items was presented in the opposite hemifield (load 0 and load 1), suggesting the reversal may reflect re-orienting of spatial attention toward the other items in memory. However, the results are interpreted with caution as the analyses were post-hoc and trial-numbers were not matched between conditions.

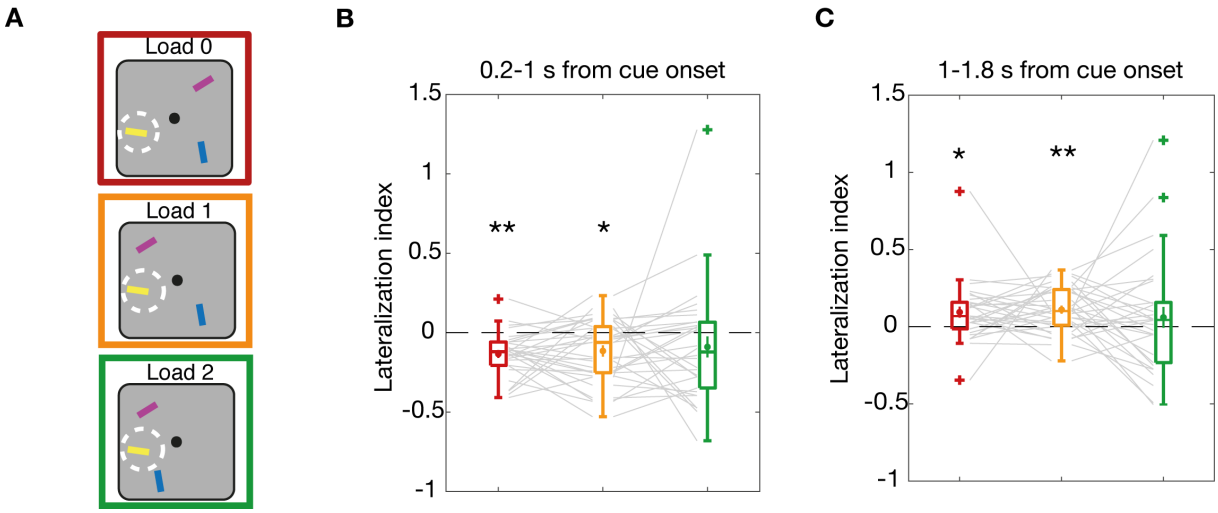

Figure S2. Alpha lateralization as a function of relative position of memory items. **A.** Example of the three possible configurations of uncued items relative to the cued item in memory. Load 0: neither of the uncued items are in the same hemifield as the cued item. Load 1: one other item is in the same hemifield as thre cued item. Load 2: both uncued memory items are in the same hemifield as the cued item. **B.** Mean alpha lateralization across the initial delay following cue onset (0.2-1s) depending on load condition: load 0 (red), load 1 (orange) or load 2 (green). **C.** Mean alpha lateralization across the late memory delay from 1-1.8 s following cue onset. Otherwise same as B. *p<.05, **p<.001.

Alpha lateralization predicts RTs on incongruent, colour-probe trials

The results of the regression analysis between alpha lateralization and absolute error are presented in Chapter 2. Below I present the results of the regression analysis between alpha lateralization and RT (Figure S3).

I performed one-sample t-tests on the β weights to test whether alpha suppression contralateral to the cued item's position in the time-window of interest (0.2-1 s post cue) was associated with faster RTs on congruent trials (one-tailed>0) and slower RT on incongruent trials (one-tailed<0). There was no evidence of a relationship between alpha lateralization and RT on congruent trials, whether orientation recall was probed by colour (p=.861) or location (p=.592). For incongruent trials, contralateral alpha suppression predicted slower RTs on colour probe trials (t_{29}=-1.99, p=.028, d=-.363), but not on location probe trials (p=.506). A 2×2 repeated measures ANOVA with congruency (congruent vs incongruent) and probe type (colour vs location) as factors revealed no significant main effect of congruency (p=.900), no main effect of probe type (p=.178) and no interaction (p=.906).

For completeness, the regression model was fit over time to obtain β weights for each time point during the delay following cue onset (Figure 5D). For colour probe trials, negative β values for incongruent trials in the post-cue delay indicated that contralateral alpha suppression relative to the cued item predicted slowed RT on incongruent trials, particularly toward the end of the delay. However this relationship was not significant when correcting for multiple comparisons across the post-cue epoch (0-2 s). There were also no significant clusters for congruent colour probe trials or congruent or incongruent location probe trials.

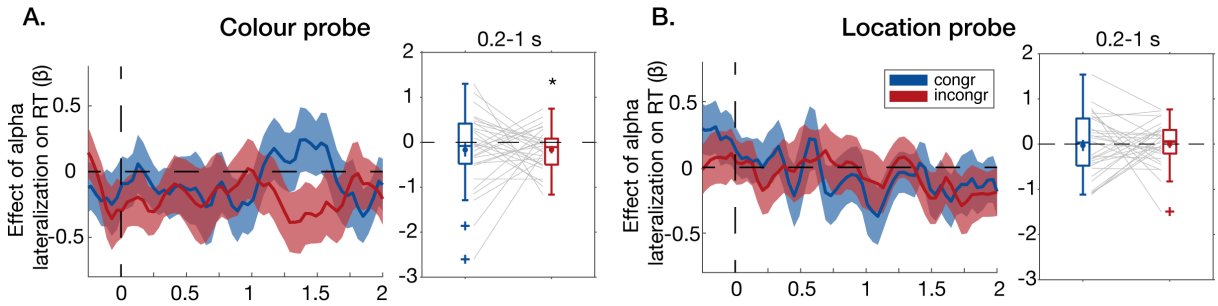

Figure S3. Relationship between alpha lateralization and reaction times. **A.** Colour probe trials. Left: mean beta coefficients from time-resolved regression analysis between trialwise lateralized alpha suppression and RTs for congruent (blue) and incongruent (red) trials. Shaded area shows standard error of mean (SEM). Right: boxplot of beta weights from regression analysis between mean alpha lateralization across TOI (0.2-1 s) and RTs. Mean and SEM overlaid. Grey lines show individual participants. **B.** Location probe trials. Otherwise same as A. *p<.05.

No significant correlation between location decoding from EEG and RT

The results of the regression analysis between location decoding from EEG and absolute error are presented in Chapter 2. Below I present the results of the regression analysis between location decoding from EEG and RT (Figure S4).

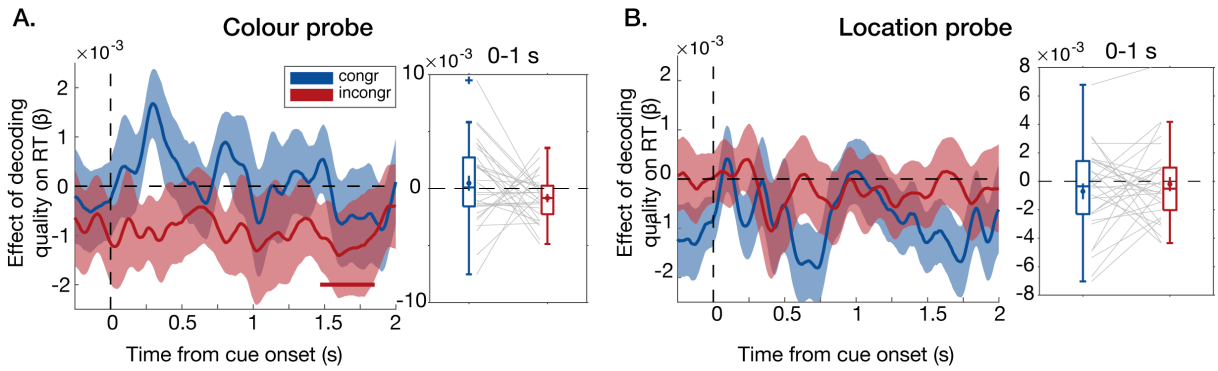

Figure S4. Relationship between location decoding from EEG and reaction times. **A.** Colour probe trials. Left: mean beta coefficients from time-resolved regression analysis between trialwise location decoding and RTs for congruent (blue) and incongruent (red) trials. Shaded area shows standard error of mean (SEM). Right: boxplot of beta weights from regression analysis between mean location decoding across TOI (0-1 s) and RTs. Mean and SEM overlaid. Grey lines show individual participants. **B.** Location probe trials. Otherwise same as A.

I performed one-sample t-tests on the β weights to test whether decoding of the cued item's original location was associated with lower RT on congruent trials (one-tailed<0) and potentially higher RT on incongruent trials (one-tailed>0). There was no evidence of a relationship between decoding accuracy for the cued item's location and RT on

congruent (colour-probe: p=.756; location-probe: p=.111) or incongruent (colour-probe: p=.986; location-probe: p=.651) trials. However, a post-hoc two-tailed t-test for incongruent colour-probe trials revealed a relationship but in the opposite direction than predicted, such that greater decoding of the cued item's location predicted faster RTs on incongruent trials (t_{29}=-2.31, p=.028, d=-.421). A 2×2 repeated measures ANOVA with factors congruency (congruent vs incongruent) and probe-type (colour vs location) showed no main effect of congruency (p=.490), no main effect of probe-type (p=.562) and no congruency by probe-type interaction (p=.099). For completeness, the regression model was fit over time to obtain β weights for each time point during the delay following cue onset. Cluster-corrected t-tests across the incidental cue epoch (0-2s post cue) revealed a significant cluster of negative beta-weights for incongruent colour-probe trials from 1.47 to 1.85 s (p=.022).

In contrast to what was expected, if this is a genuine effect it would suggest that greater decoding of the cued item's location predicts faster RTs on incongruent trials, when recall is probed by colour. It is possible that when the cued item's location is more strongly represented in memory, participants initially start to dial up the *cued* item's orientation. This may result in faster RTs specifically when probed by colour and not when probed by location where there would presumably be a stronger error signal to indicate that a different item was being probed. However, it is not clear why a stronger location representation would not similar encourage faster responses on congruent trials. Thus, I interpret this effect with caution as it is in the opposite direction to what was expected.

No significant correlation between gaze bias and error

The results of the regression analysis between eye-position decoding and RT are presented in Chapter 2. The results of the regression analysis between eye-position decoding and absolute error are presented in Figure S5.

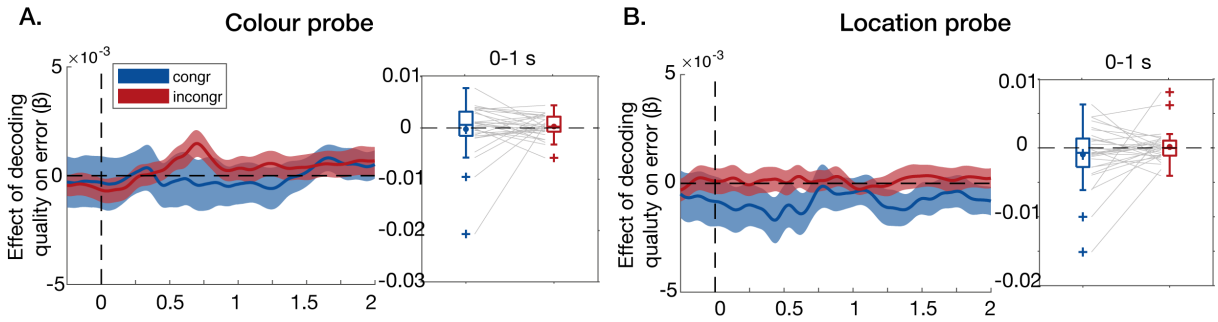

Figure S5. Relationship between location decoding from eye-position and absolute error. **A.** Colour probe trials. Left: mean beta coefficients from time-resolved regression analysis between trial-wise location decoding from eye-position and absolute error for congruent (blue) and incongruent (red) trials. Shaded area shows standard error of mean (SEM). Right: boxplot of beta weights from regression analysis between mean location decoding from eye-position across TOI (0-1 s) and RTs. Mean and SEM overlaid. Grey lines show individual participants. **B.** Location probe trials. Otherwise same as A.

I performed one-sample t-tests on the β weights to test whether decoding of the cued item's original location from eye-movements was associated with lower error on congruent trials (one-tailed<0) and potentially higher error on incongruent trials (one-tailed>0). There was no evidence for a significant relationship between location decoding from eye-position and behavioural error on congruent trials (colour-probe: p=.405, location-probe: p=.122; Figure S5) or incongruent trials (colour-probe: p=.240, location-probe: p=.380), and neither when averaging over probe-condition (congruent: p=.177, incongruent: p=.199). A 2×2 repeated measures ANOVA with factors congruency (congruent vs incongruent) and probe-type (colour vs location) showed no main effects (congruency: p=.256, probe-type: p=.467) and no significant interaction (p=.670). For completeness, the regression model was fit over time to obtain beta weights for each time point during the delay following cue onset. Cluster-corrected t-tests across the incidental cue epoch (0-2 s) revealed no significant clusters for either condition.

Appendix B. Supplementary material for Chapter 3

Orientation decoding from eye-position

I repeated the same Mahalanobis distance decoding procedure, described in Chapter 3, to decode memorised orientations from eye-position (Figure S1). One additional subject was excluded from this analysis due to missing data.

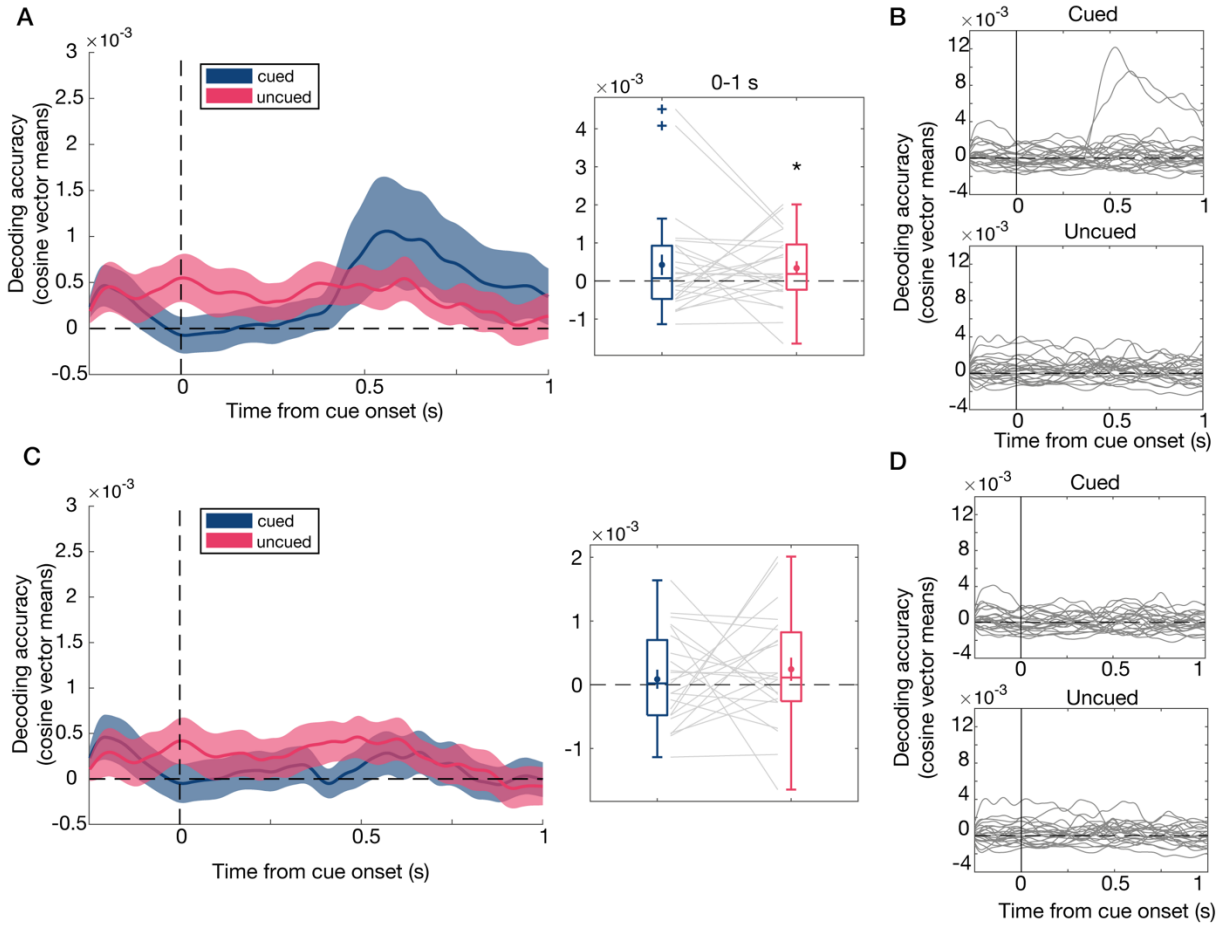

Figure S1. Decoding memorised orientation from eye-position. **A.** Left: Decoding accuracy over time following cue onset. Shaded area shows SEM. Right: Boxplot shows mean decoding across TOI 0-1s from cue onset. Overlaid mean and SEM. Grey lines show individual participants. *p<.05. **B.** Decoding accuracy over time for individual participants (grey lines) for cued (top) and uncued (bottom) item. **C.** Same as A, but after excluding participants with high eye-position orientation decoding from analyses. D. Same as B, after excluding participants with high eye-position orientation decoding from analyses.

In the time window of interest from cue onset (0-1 s), decoding the cued item's orientation from eye-position was at a trend-level (t_{24}=1.55, p=.067, d=.311) and the

orientation of the uncued item could be decoded (t_{24}=1.85, p=.038, d=.370). There was no difference in decoding strength between the cued and uncued item (p=.386). While there were no significant clusters of time-resolved decoding post cue, there was an increase in decoding of the cued item's orientation peaking at ~0.5 s after cue onset. Inspecting the boxplot and time-course of decoding for individual participants (Figure S1B), this was driven by strong cued orientation decoding for two participants. Consequently, I excluded these two participants from the analysis resulting in non-significant decoding from eye-position for both memory items (cued: p=.289, uncued: p=.102) and no difference between them (p=.726).

Next, I tested whether excluding these two subjects, in addition to the subject with missing eye data, affected orientation decoding from EEG (Figure S2). Using the standard decoding procedure, orientation decoding following cue onset (0-1s) was no longer significant for the cued item, though there was a trend (t_{22}=1.62, p=.060, d=.337). Decoding remained non-significant for the uncued item (p=.215) and there was no significant difference between cued and uncued item decoding (p=.616). Cluster-corrected t-tests across the time-window of interest post cue onset (0-1 s) still revealed a cluster of significant decoding for the cued item from 0.38 to 0.54 s (p=.047). Spatiotemporal decoding of the cued item orientation remained significant after excluding the three subjects (t_{22}=3.43, p=.001, d=.715) and the orientation of the uncued item could not be decoded (p=.144). Decoding strength was significantly higher for the cued item compared to the uncued item (t_{22}=1.82, p=.041, d=.379).

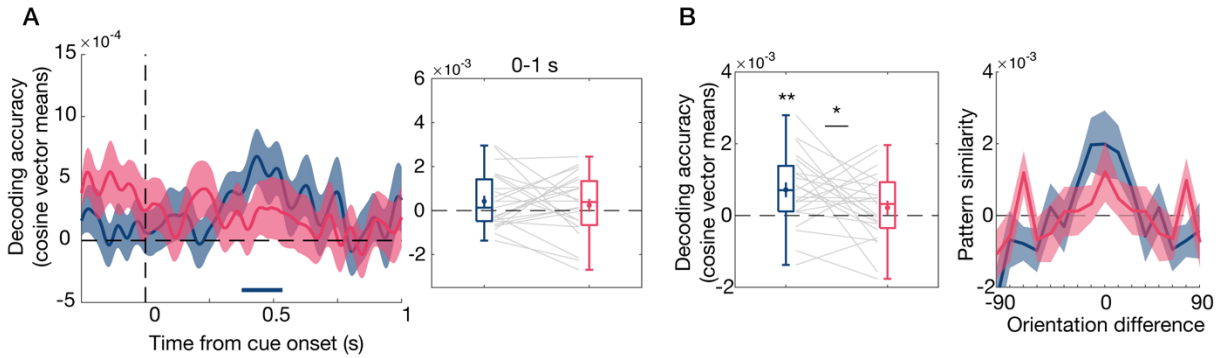

Figure S2. Orientation decoding from EEG after excluding participants with high eye-position orientation decoding from analyses. **A.** Left: Decoding accuracy over time following cue onset. Shaded area shows SEM. Bars show clusters of significant orientation decoding. Boxplot shows mean decoding across TOI 0-1s from cue onset. Overlaid mean and SEM. Grey lines show individual participants. **B.** Spatiotemporal decoding. Left: Boxplot of decoding accuracy following cue (100-500 ms). Overlaid mean and SEM. Grey lines show individual participants. Right: Mahalanobis distance pattern similarity as a function of orientation difference. Shaded area shows SEM. *p<.05, **p<.001.

Gaze bias and RT

The results of the regression analysis between gaze bias and absolute error are presented in Chapter 3. Below, I present the results of the regression analysis between gaze bias and reaction times (Figure S3).

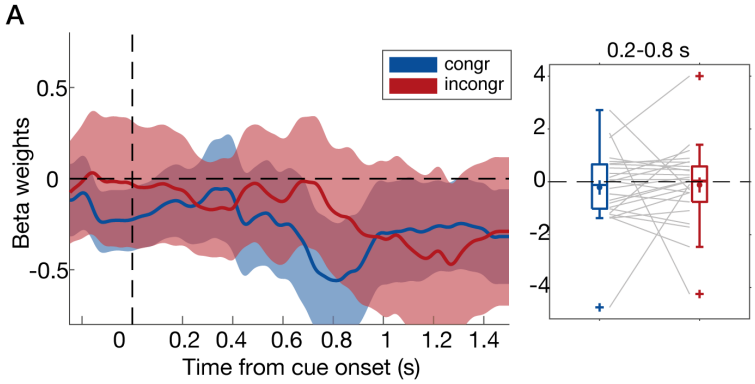

Figure S3. Gaze bias and reaction times. **A.** Beta weights from trial-wise regression with time-resolved gaze bias and reaction times for congruent (blue) and incongruent (red) trials. Shaded area shows SEM. Boxplots show beta weights from regression with mean alpha lateralization across post-cue TOI (0.2-0.8 s) and reaction times. Mean and SEM overlaid. Grey lines show individual participants.

I performed one-sample t-tests on the beta-weights to test whether gaze bias toward the cued item's original position in the time-window of interest (0.2-0.8 s post cue) was associated with lower RT on congruent trials (one-tailed<0) and higher RT on

incongruent trials (one-tailed>0). There was no evidence for a relationship between gaze bias and RT for either congruent (p=.225) or incongruent (p=.656) trials, and there was no difference between them (p=.798). Fitting the regression for each time-point also revealed no significant clusters for either congruent or incongruent trials.

Orientation decoding does not predict orientation recall performance

There was no evidence of a relationship between orientation decoding and reaction time performance (Figure S4A & C) or absolute error (Figure S4B & D). The outcomes of the significance tests are presented in Chapter 3.

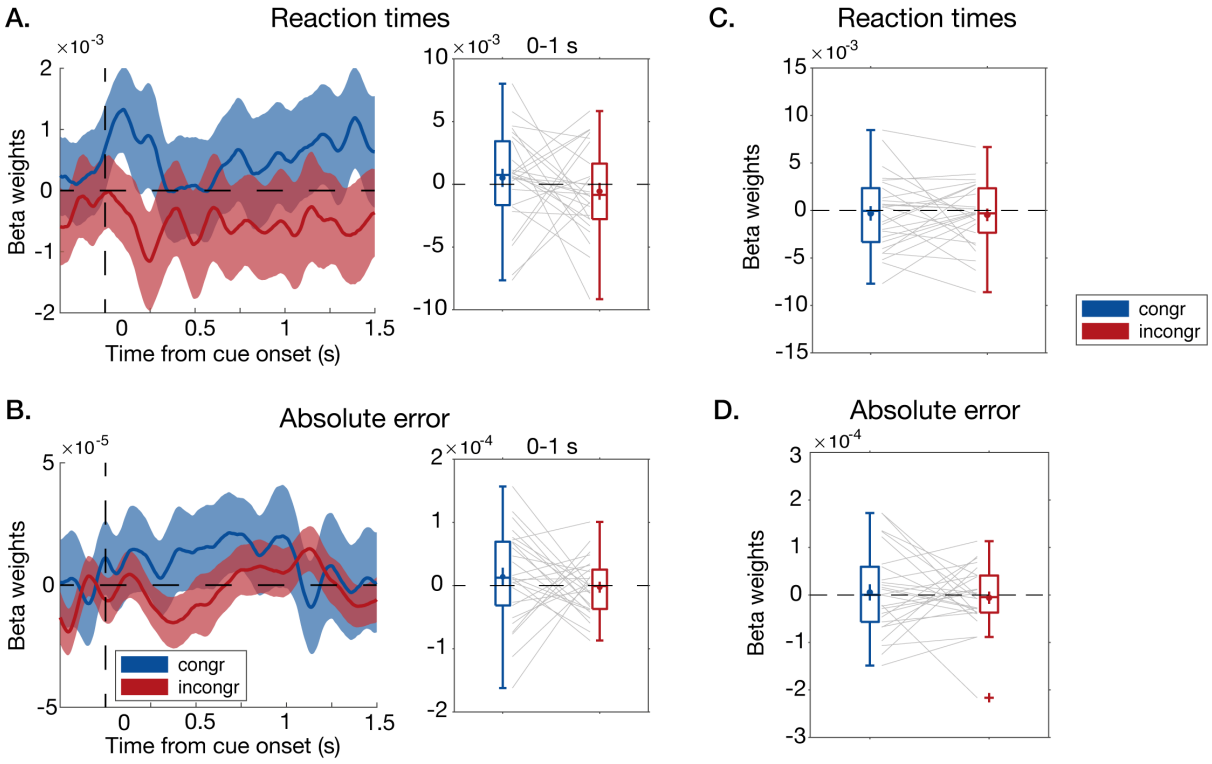

Figure S4. Orientation decoding from EEG and behavioural performance (RT and absolute error). **A.** Beta weights from trial-wise regression with time-resolved orientation decoding and reaction times for congruent (blue) and incongruent (red) trials. Shaded area shows SEM. Boxplots show beta weights from regression with mean orientation decoding across post-cue TOI (0-1s) and reaction times. Mean and SEM overlaid. Grey lines show individual participants. **B.** Same as A, but for regression with absolute error. **C.** Boxplot of beta weights from trial-wise regression with spatiotemporal orientation decoding post-cue (100-500 ms) and reaction times. Mean and SEM overlaid. Grey lines show individual participants. **D.** Same as C, but for absolute error.

Residual probe-target distance decoding after regressing out reaction times

RTs varied as a function of probe-target distance, with slower RTs when the probe-target distance was small (Chapter 3; Figure 6C). Consequently, the probe-target distance decoding analysis may have been confounded by a temporal signal of how recently the probe appeared. To control for this possibility, I computed the residual decoding values after regressing out trial-wise RT for each subject (Figure S5A).

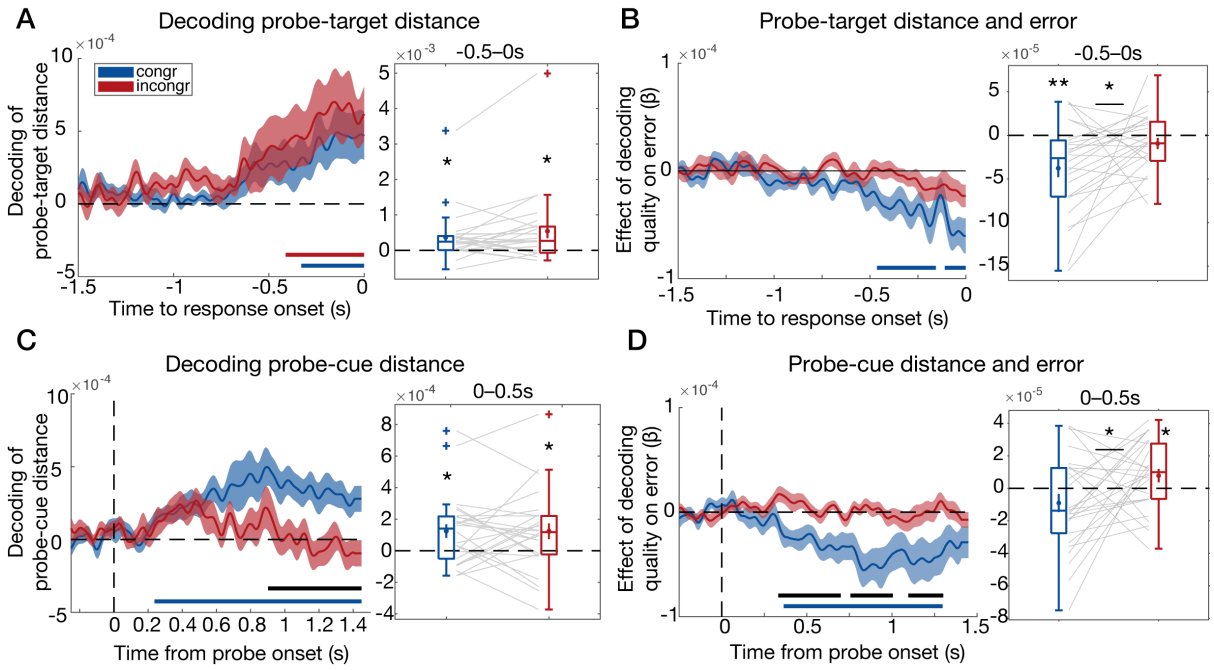

Figure S5. Residual probe-target/probe-cue distance after regressing out RTs. **A.** Residual decoding of probe-target distance (absolute distance between random probe orientation and orientation of probed item in memory) for congruent (blue) and incongruent (red) trials time-locked to response onset. Shaded area shows SEM. Bars show significant decoding. Boxplot shows mean probe-target distance decoding before response onset (-0.5-0 s). Mean and SEM overlaid. Grey lines show individual participants. **B**. Beta weights from trial-wise regression with residual time-resolved probe-target distance decoding and absolute error for congruent (blue) and incongruent (red) trials. Bars show non-zero beta weights. Boxplots shows beta weights from regression with mean decoding in TOI (-0.5-0s) and absolute error. **C.** Residual decoding of probe-cue distance (absolute distance between random probe orientation and orientation of cued item in memory) for congruent (blue) and incongruent (red) trials time-locked to probe onset. Boxplot shows mean decoding strength in TOI after probe onset (0-0.5 s). **D.** Beta weights from trial-wise regression with residual probe-cue distance decoding and absolute error. Otherwise same as B. *p<.05, **p<.001.

Although residual decoding values were lower, the probe-target distance could still be decoded prior to response onset on both congruent and incongruent trials (TOI:-0.5-0 s; congruent: t_{25}=2.64, p=.007, d=.518; incongruent: t_{25}=2.67, p=.007, d=.523) and there was no difference between conditions (p=.135). Cluster-corrected t-tests across the time window prior to response onset (-1.5-0 s) revealed clusters of significant decoding on congruent trials from -0.33 to 0 s before response onset (p=.002) and on incongruent trials from -0.41 to 0 s (p<.001). Thus, the probe-target distance was decodable from EEG before response onset even after controlling for differences in RTs.

Figure S5B shows β weights from the regression analysis between the strength of residual probe-target distance decoding and absolute error. Broadly consistent with the main results, β weights were significantly below zero for congruent trials (t_{25}=-3.62, p<.001, d=-.710) and though not significant, there was a trend for negative β weights on incongruent trials (t_{25}=-1.43, p=.082, d=-0.281; Figure S5B). As before, decoding-strength predicted error to a greater extent on congruent than incongruent trials (t_{25}=-2.33, p=.028, d=-.457). Cluster-corrected t-tests across the time window prior to response onset (-1.5-0 s) revealed significant clusters of negative beta weights for congruent trials prior to response onset from -0.46 to -0.16 s (p=.003) and again from -0.11 to 0 (p=.024). There were no clusters of significant correlations between probe-target distance decoding and error on incongruent trials and no significant differences between congruent and incongruent trials.

Residual probe-cue distance decoding after regressing out reaction times

As for probe-target distance decoding above, I computed the residual probe-cue distance after regressing out RTs (Figure S5C). Again, the results broadly confirmed the main analysis. Residual probe-cue distance could be decoded on both congruent and incongruent trials following probe onset (TOI:0-0.5 s; congruent: t_{25}=2.84, p=.004, d=.557; incongruent: t_{25}=2.44, p=.011, d=.479). There was no difference in decoding

strength between the conditions in the initial time-window after probe onset (0-0.5s; p=.978). Cluster-corrected t-tests across the time window after probe onset (0–1.45 s) revealed a cluster of significant decoding on congruent trials from 0.24 to 1.45s (p<.001) but no significant time points for incongruent trials. There was a significant difference in decoding between congruent and incongruent trials from 0.90 to 1.45 s (p=.001).

Figure S5D shows β weights from the regression analysis between the strength of residual probe-cue distance decoding and absolute error. While there was still a negative relationship between residual probe-cue distance decoding and error on congruent trials, as in the main results, this was no longer significant (t_{25}=-1.57, p=.065, d=-.307). However, the trend for a positive relationship between probe-cue distance decoding strength and error on incongruent trials, reported in the main results, was now significant after regressing out RTs (t_{25}=1.85, p=.038, d=.364). The significant difference between congruent and incongruent trials remained (t_{25}=-2.22, p=.036, d=-.435). Cluster-corrected t-tests across the time window after probe onset (0–1.45 s) still revealed no significant clusters for incongruent trials. There was a significant cluster of negative beta weights for congruent trials from 0.36 to 1.30 s (p=.001) and significant differences between beta weights for congruent and incongruent trials from 0.33 to 0.70 s (p=.01), 0.76 to 1.01 s (p=.022) and 1.10 to 1.30 s (p=.040).

Decoding probe-target distance (probe-locked)

The absolute distance between the random probe and the probed memory item (probe-target distance) could be decoded in the first 500 ms after probe onset on congruent trials (t_{25}=3.74, p<.001, d=.734) and to some extent on incongruent trials, although decoding strength was not significant (t_{25}=1.67, p=.054, d=.328; Figure S6A). Decoding strength was significantly better on congruent than incongruent trials (t_{25}=2.38, p=.025, d=.466). Cluster-corrected t-tests across the time window following probe onset (0-1.45 s) revealed clusters of significant decoding on congruent trials from 0.29 to 1.45 s (p<.001)

and on incongruent trials from 0.42 to 1.45 s (p<.001) and a significant difference between congruent and incongruent trials from 0.76 to 1.45 s (p=.004). This suggests the onset of significant decoding of the comparison-signal between the probe and the target item in memory was delayed on incongruent compared to congruent trials.

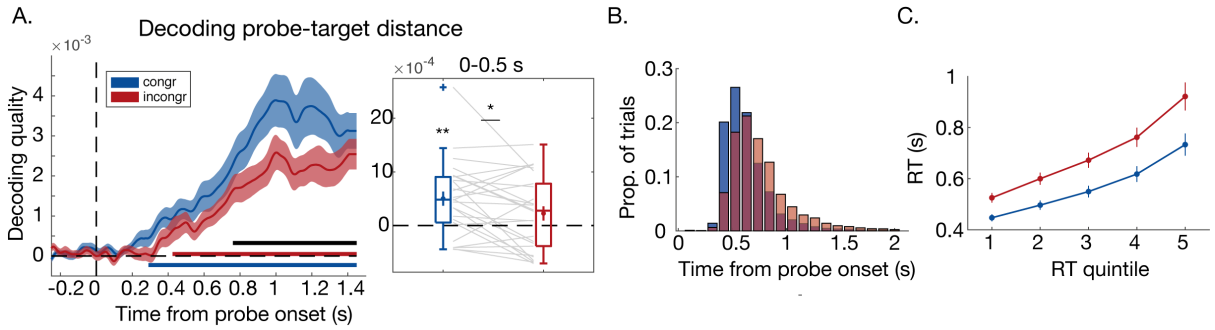

Figure S6. Decoding probe-target distance time-locked to probe-onset. **A.** Probe-target distance decoding for congruent (blue) and incongruent (red) trials. Shaded area shows SEM. Bars show significant decoding. Boxplot shows mean probe-target distance decoding 0-0.5 s after probe onset. Mean and SEM overlaid. Grey lines show individual participants. *p<.05, **p<.001; **B.** Histogram of response onset across participants for congruent (blue) and incongruent (red) trials. **C.** Mean reaction time as a function of RT quintile. Error bars show SEM.

The delay in onset of probe-target distance decoding on incongruent trials (~130ms) corresponds approximately to the difference in mean RTs between congruent and incongruent trials (120ms; Figure 2, Chapter 3). Therefore, I speculated slowed RTs may be caused by participants first comparing the random memory probe to the cued item in memory thereby delaying decision-making on incongruent trials. Two possible sources of slowed RTs include delayed onset of decision-making (non-decision time) and slowed integration of evidence toward a decision (decision-time). Delays in the onset of decision-related processes should cause a general shift in the RT distribution on incongruent relative to congruent trials, while extended decision-making processes should cause greater positive skew in the distribution (Ratcliff & Tuerlinckx, 2002).

Figure S6B shows the distribution of response onsets across participants and Figure S6C shows mean RTs as a function of RT quintile. I compared RTs across conditions and for

each quintile in a 2x5 repeated measures ANOVA. There was a main effect of congruency ($F_{1,25}$=63.08, p<.001), a main effect of quintile ($F_{4,25}$=90.85, p<.001) as well as a congruency by quintile interaction ($F_{4,25}$=39.34, p<.001). The difference in RT between congruent and incongruent conditions was larger for slower RTs (Figure S6C), suggesting decision-making might be prolonged for incongruent relative to congruent trials. However, significant differences in RT between congruent and incongruent trials at each RT quintile (confirmed with one-sample t-tests, one-tailed>0; 1st quintile: t_{25}=7.02, p<.001; 2nd: t_{25}=7.88, p<.001; 3rd: t_{25}=8.02, p<.001; 4th: t_{25}=8.57, p<.001; 5th: t_{25}=7.98, p<.001), further suggests that the onset of decision-related processes was delayed. Consequently, a combination of extended non-decision time and decision-time may contribute to slowed RTs on incongruent relative to congruent trials.

Appendix C. Supplementary material for Chapter 4

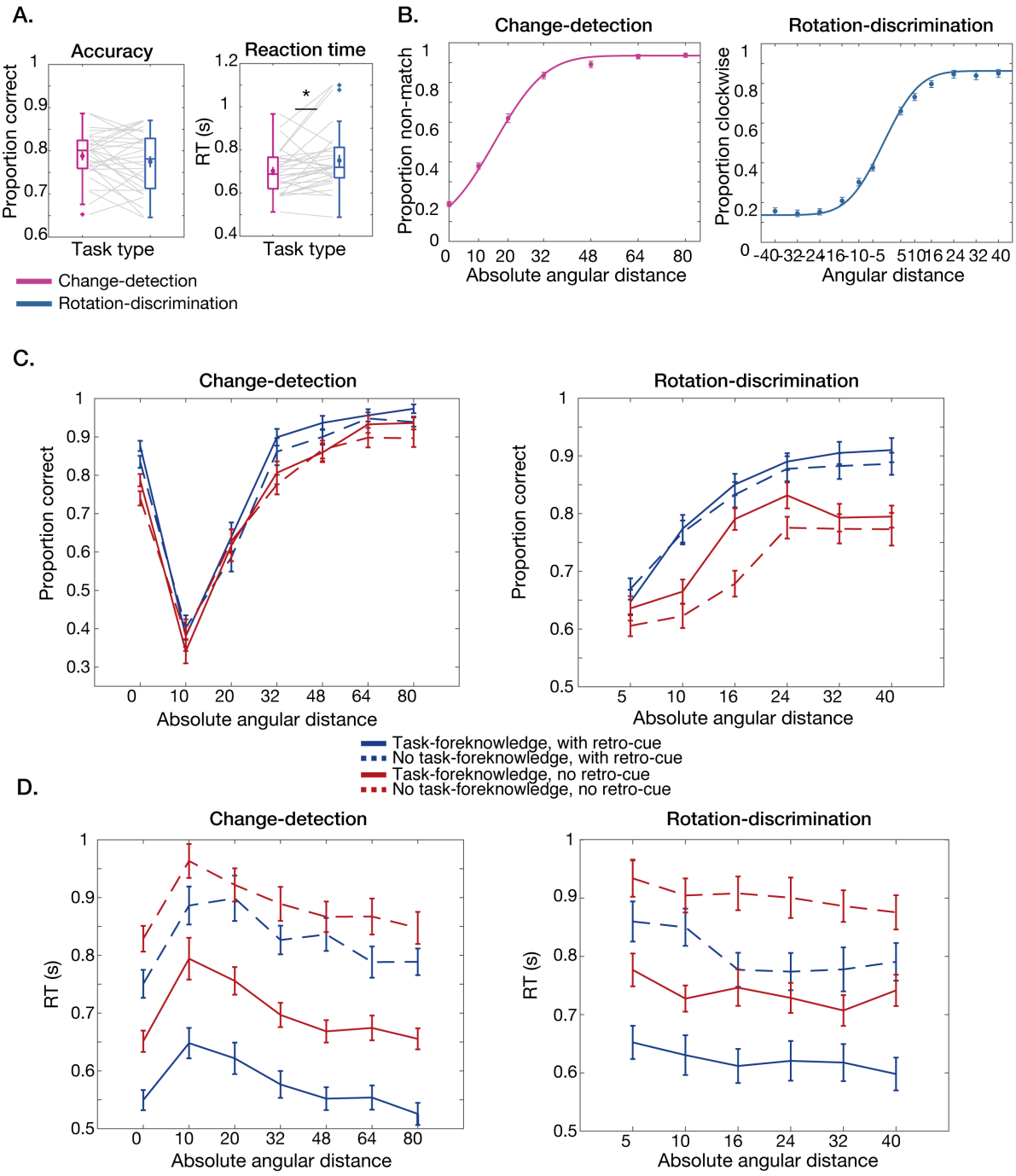

Figure S1. Behavioural performance across the two task types. **A.** Boxplots of accuracy (proportion correct; left) and median reaction times (right) for the change-detection task (pink) and rotation-discrimination task (blue). Mean and standard error of mean (SEM) overlaid. Grey lines show individual participants. *p<.05. **B.** The proportion of 'non-match' (left) and clockwise (right) responses for each degree offset between the memorized item and the probe, with a cumulative Gaussian function fit overlaid for the change-detection task (left) and rotation-discrimination task (right). Error bars show SEM. **C.** Accuracy (proportion correct) across the span of angular offsets between the memory item and the probe for the change-detection task (left) and rotation-discrimination task (right) for all conditions: task-foreknowledge, with retro-cue (solid blue line), no task foreknowledge, with retro-cue (dashed blue line), task-foreknowledge, no retro-cue (solid red line), and no task-foreknowledge, no retro-cue (dashed red line). **D.** Same as C but for median reaction times.

Appendix D. Supplementary material for Chapter 5

Data simulations

We validated our analysis approach by generating simulated neural data under the assumption of either task-specific and/or task-invariant working memory (WM) codes. In order to ensure that the analysis could in principle uncover true task-specific WM codes under various possible scenarios, but also identify task-invariant coding, we generated data from several different ground truths assuming either equal (Figure S1) or unequal (Figure S2) levels of noise for the two tasks: A) only task-specific coding B) only task-invariant WM coding, C) both task-specific and task-invariant WM coding, D) task-invariant WM coding, along with task-specific WM coding for only one of the two tasks. For each of the four scenarios, we simulated 60 EEG channels over 2000 trials i for N=30 repetitions of each simulation:

$$Y_i = wX_i + e_i$$

Where Y is a $channel \times time$ matrix of the multivariate signal on each trial, X is a set of multivariate WM and task patterns, w is a binary vector of flags depending on the simulated scenario and e is a $channel \times time$ matrix of noise drawn from a normal distribution $\sim N(0, SD_i)$. SD_i is the noise level for each task, drawn at random from a normal distribution with parameters $\sim N(10, 0)$ for equal noise levels (Figure S1) or $\sim N(10, 2)$ for unequal noise levels (Figure S2). On each trial i, X is defined as:

$$X_i = P_i^1 \cos(\theta_i) + P_i^2 \sin(\theta_i) + Q^1 \cos(\theta_i) + Q^2 \sin(\theta_i) + R_i + U_i$$

Where P is a $channel \times 2$ task-specific multivariate WM signal, Q is a $channel \times 2$ task-invariant multivariate WM signal, R is a multivariate $channel \times 1$ task signal and U is a scalar task signal added to all channels. θ is the WM condition on each trial. The simulated WM conditions were orientations ranging from 0-360° chosen at random with

replacement on each trial. The task conditions were Task A or Task B, chosen at random with replacement on each trial. We simulated a time-course by scaling the signal X in steps over time from 0% to 25% to 75% to 100%. To mimic EEG data, we also introduced covariance by smoothing the simulated data over channels using a gaussian kernel (SD = 2).

For orientation decoding (Figure S1 and S2, left panel), we followed the analysis pipeline described in Chapter 5 with a few exceptions. Instead of k-fold cross-validation, we split the data in half and trained the decoder on one half of the data and tested the decoder on the other half. As the simulated data does not have a dynamic time-course, we did not use spatiotemporal decoding as described above and instead ran the decoder separately for each timepoint. To compute cross-decoding costs, we took the mean decoding values over time when the full signal strength (100%) was present in the simulated data. If within-task decoding was matched across the two tasks, we used the mean cross-generalization costs for each task as an index of use-specific WM coding. If within-task decoding was not matched for the two tasks, we instead used the cross-generalization cost for the worst task, defined per subject, as an index of use-specific WM coding (Figure S1 & S2, centre-left panel).

In addition to orientation decoding, we also decoded the current task context (Figure S1 and S2, right panel). We computed Mahalanobis distance between each trial in the test set and the mean Task A training trials and the mean Task B training trials. We took the difference between the resulting Task A and Task B distances for each trial and sign-reversed Task A test trials, so larger values reflected greater similarity between test trials and training data with matching task labels. Finally, we took the mean divided by the standard deviation over trials to get a single task decoding value for each time point for each subject.

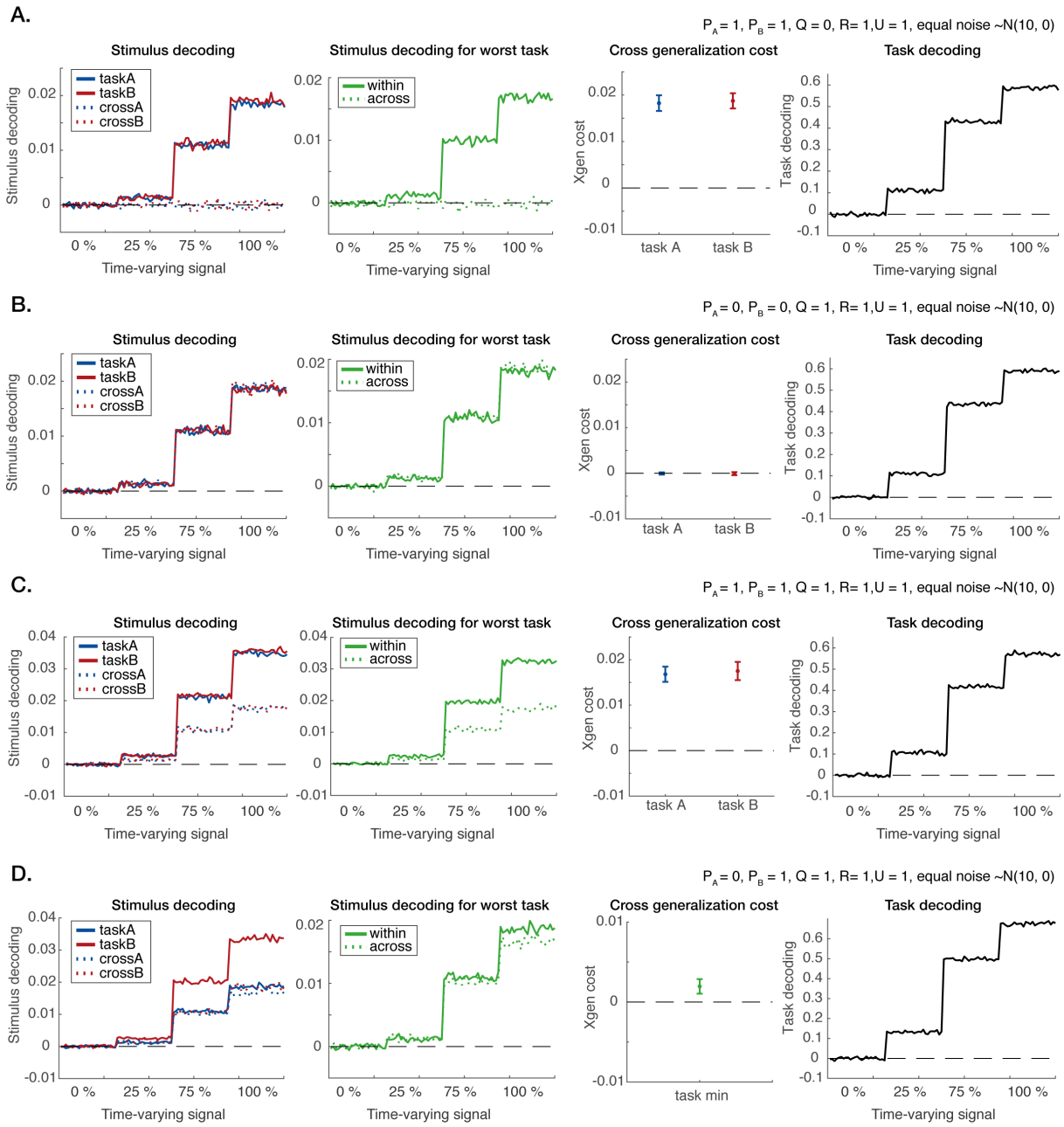

Figure S1. Decoding results on simulated neural data with equal noise levels for the two tasks. Left panel: decoding of the WM stimulus orientation. Solid lines represent within-task decoding and dotted lines represent between-task decoding. Centre left panel: decoding of the WM stimulus for the worst task, defined per subject. Solid lines represent within-task decoding and dotted lines represent between-task decoding. Centre right panel: cross-decoding cost computed separately for task A and task B on the basis of testing set (blue and red) or computed for the worst task (green). Right panel: decoding of task context **A. Task-specific coding.** Simulation with task-specific WM code for both tasks (P_A and P_B), no task-invariant WM code (Q), a multivariate task signal (R), a univariate task signal (U) and equal noise levels for the two tasks. **B. Task-invariant coding.** Simulation with no task-specific WM stimulus code for either tasks. Otherwise same as A. **C. Both task-specific and task-invariant coding.** Same as A., but supplemented by a task-invariant WM code (Q). **D. Task-invariant coding, along with task-specific coding for one task.** Task-specific WM stimulus code for task B only (P_B). Otherwise same as B.

Figure S1 illustrates the results of four simulated scenarios with equal noise levels for the two tasks. Figure S1A shows a scenario with task-specific stimulus patterns for Task A and Task B, and no task-invariant stimulus pattern, resulting in zero cross-decoding between tasks (left panel, dotted lines) and a cost of cross-generalization (centre-right panel) for each task, consistent with our main hypothesis (H1) of use-specific WM coding. Figure S1B shows a scenario with no task-specific stimulus patterns, resulting in perfect cross-decoding between tasks, consistent with the alternative hypothesis (H0). It is worth noting that perfect cross-generalization is possible, even when task context is strongly decodable from the data based on multivariate and/or univariate differences (right panel). Figure S1C shows a scenario with task-specific stimulus patterns for Task A and Task B, supplemented by a task-invariant stimulus pattern, resulting in preserved cross-decoding between tasks (left panel, dotted lines), as well as a cost of cross-generalization (centre-right panel) for each task, again consistent with the use-specific WM hypothesis (H1), which does not make assumptions about the presence or absence of an additional shared WM code. Figure S1D shows the results of a simulation with a task-invariant neural code supplemented by a task-specific stimulus pattern for Task B only. The presence of a task-specific stimulus component, even if only for one task, would be interpreted as evidence in favour of our hypothesis (H1). The differences in overall within-task decoding between the two tasks means the cross-generalization cost is computed based on the worst task (S1D, centre-right panel).

Next, we wanted to ensure the analysis pipeline could, in principle, uncover true effects or null effects in the presence of unequal noise levels between the two tasks. Figure S2 shows the results of the same scenarios as above, but with unequal noise levels between tasks. The presence of unequal noise generally produces differences in overall within-task decoding (left panel), and the cross-generalization cost is consequently computed on the worst task (centre-left and centre-right panel). In all cases, the analysis pipeline

leads us to accept the hypothesis (H1) when a task-specific WM code is present in the data (Figure S2A,C-D) and reject the hypothesis in favour of the null (H0) when the WM stimulus is coded exclusively in a task-invariant pattern (Figure S2B).

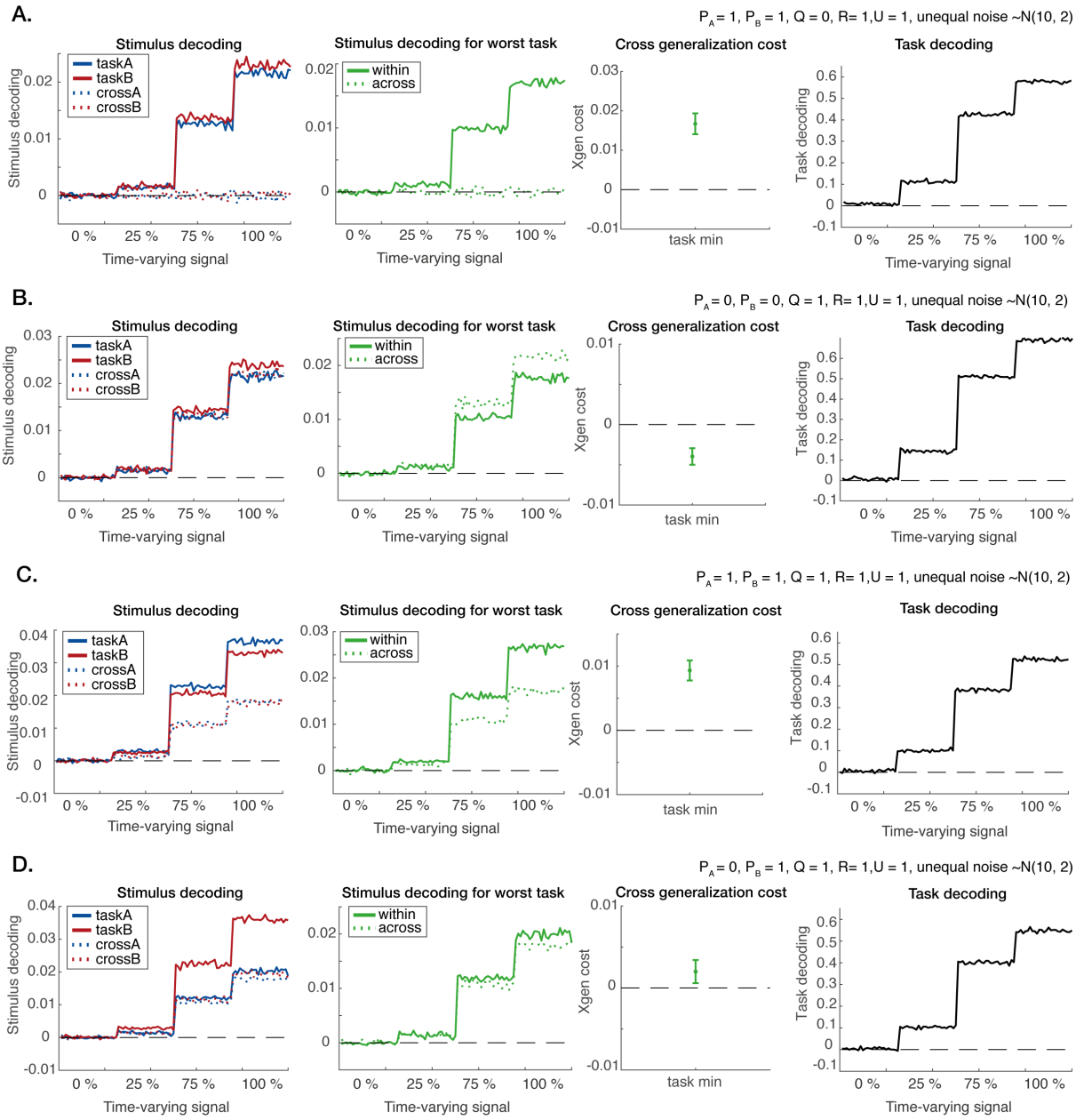

Figure S2. Decoding results on simulated neural data with unequal noise levels for the two tasks. Left panel: decoding of the WM stimulus orientation. Solid lines represent within-task decoding and dotted lines represent between-task decoding. Centre left panel: decoding of the WM stimulus for the worst task, defined per subject. Solid lines represent within-task decoding and dotted lines represent between-task decoding. Centre right panel: cross-decoding cost computed separately for task A and task B on the basis of testing set (blue and red) or computed for the worst task (green). Right panel: decoding of task context **A. Task-specific coding.** Simulation with task-specific WM code for both tasks (P_A and P_B), no task-invariant WM code (Q), a multivariate task signal (R), a univariate task signal (U) and unequal noise levels for the two tasks**. B. Task-invariant coding.** Simulation with no task-specific WM stimulus code for either tasks. Otherwise same as A. **C. Both**

task-specific and task-invariant coding. Same as A., but supplemented by a task-invariant WM code (Q). **D. Task-invariant coding, along with task-specific coding for one task**. Task-specific WM stimulus code for task B only (P_B). Otherwise same as B.

However, we note that when allowing for larger differences in noise levels between tasks (e.g. $\sim N(15, 3)$), our conservative proposal to compute cross-generalization costs based on the worst task might miss more subtle effects in scenario D, specifically when a task-specific code is present only for one task and is supplemented by task-invariant coding. This would be an unlikely combination of effects and could be diagnosed by checking for differences in within-task decoding and noise for each participant. For each subject, differences in noise levels between tasks will be quantified by computing the first eigenvalue of the noise covariance matrix separately per task and compare this to noise levels estimated from data with shuffled task labels, using a permutation test. We will test for a significant noise differences at the group level using a one-sample t-test (one-tailed, >0) on the z-values derived from the permutation test p-values. A control analysis excluding subjects with significant differences in noise levels between tasks until they are matched in the remaining sample could recover the true effect in such a scenario.