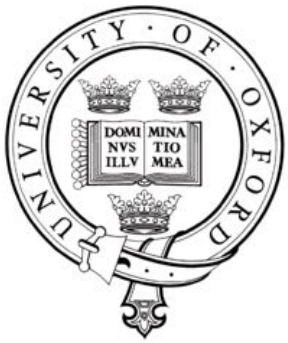
The Role of Cortical Oscillations in the Control and Protection of Visual Working Memory

Nicholas E. Myers
Merton College, Oxford

D.Phil. Thesis, Trinity Term, 2015





MERTON
COLLEGE
OXFORD

The Role of Cortical Oscillations in the Control and Protection of Visual Working Memory

Nicholas E. Myers

Merton College, Oxford

Trinity Term, 2015

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Oxford

Abstract

Visual working memory (WM) is the ability to hold information in mind for a short time before acting on it. The capacity of WM is strikingly limited. To make the most of this precious resource, humans exhibit a high degree of cognitive flexibility: We can prioritize information that is relevant to behavior, and inhibit unnecessary distractions. This thesis examines some behavioral and neural correlates of flexibility in WM.

When information is of particular importance, anticipatory attention can be directed to where it will likely appear. Oscillations in visual cortex, in the 10-Hz range, play an important role in regulating excitability of such prioritized locations. Chapter 4 describes how even spontaneous fluctuations in 10-Hz synchronization (measured by electroencephalography, EEG) before encoding influence WM. Chapters 2 and 3 describe how attention can be directed retrospectively to items even if they are already stored in WM. Chapter 3 discusses how retrospective cues change neural synchronization similarly to anticipatory cues. Behavioral and neural measures additionally indicate that the boosting of an item through retrospective cues does not require prolonged deployment of attention: rather, it may be a transient process.

The second half of this thesis additionally examines how items are represented in visual WM. Chapter 5 summarizes a study using pattern analysis of magnetoencephalographic (MEG) and EEG data to decode features of visual templates stored in WM. Decoding appears transiently around the time when potential target stimuli are expected, in line with a flexible reactivation mechanism. Chapter 6 further examines separate cortical networks involved in protecting vs. updating items in WM, and tests whether task relevance changes how well WM contents can be decoded. Finally, Chapter 7 summarizes the thesis and discusses how attentional flexibility can merge WM with a wider range of sources of behavioral control.

This thesis was funded by a 4-year Wellcome Trust studentship.

Approximate word count: 55,000 words.

Acknowledgments

I owe a large debt of gratitude to my two advisors: to Mark Stokes, for making this thesis longer, and to Kia Nobre, for making it shorter. Needless to say, this project could not have been completed without their guidance and support. Their rigor and nuance, their wholesome skepticism and wide-ranging enthusiasm, are qualities to which any scientist should aspire.

In addition to my advisors, I have had the good fortune to collaborate with a number of inspiring researchers. George Wallis, always a source of brilliant and challenging ideas on working memory, could also be trusted to commiserate when, inevitably, I got stuck battling MEG data. Gustavo Rohenkohl and Valentin Wyart could be relied upon to think of new experiments and to suggest clever and overcomplicated analyses – if I have the occasional good idea, usually it was inadvertently stolen from them. Theresa Wildegger, Rob Mok, Julian de Freitas, and Michael Wolff were kind enough to let me join in on some of their projects – not only did they allow me to learn about new subjects, but they also provided me with fresh perspectives on my own work, for which I am very grateful.

Some of the experiments in this thesis were conducted in collaboration with three undergraduate students: Lena Walther, James Baker, and Sammi Chekroud. It was a pleasure to work side by side with them, to share the fun and excitement of posing new questions and seeing answers slowly emerge, often to our puzzlement.

Much of the work presented here depended heavily on the genius for analysis and the teaching skills of the MEG analysis group, and especially on Mark Woolrich, Giles Colclough, Henry Luckhoo, Adam Baker, and Andew Quinn. Their insights, support, and good cheer helped me through many moments of desperation in trying to understand MEG.

Thanks are due to Debbie Clarke, who I could turn to when I was lost in the academic bureaucracy (which was usually at the last minute before a deadline) – like many students she has supported, I would often have faltered without her help and her tireless work on the Neuroscience programme. Likewise, Andy King's guidance as programme director and college advisor are much appreciated. I am also grateful to the Wellcome Trust and to Merton College for funding my research.

It has been a pleasure to be part of Kia's and Mark's labs. Their members (and the members of the Rushworth lab) confirmed that science is a place for good people, and that scientists are capable of having fun. For innumerable instances of help and kindness I am grateful to: Zita Patai, Alexandra Murray, Ian Gould, Riam Kanso, Bo-Cheng Kuo, Kathryn

Atherton, Josh Chauvin, Simone Heideman, Celine Gillebert, Nahid Zokaei, Natalie Nelissen, Alexander Luettich, Malcolm Proudfoot, Freek van Ede, Rocio Silva Zunino, Nora Roüast, Maryann Noonan, Ben Crittenden, Janina Jochim, Eelke Spaak, Lev Tankelevitch, Jacqueline Ding, Nils Kolling, and Jacqueline Scholl.

I would like to thank my family, Ina, Steve, Siobhán, Tommy, and Olivia, for encouraging me at every turn, and for making it possible for me to get more degrees than a thermometer. Finally, I owe heartfelt thanks to Laura Theis, who has tolerated me over the years and kept me from turning into a curmudgeon.

Table of Contents

Chapter 1: Neural Mechanisms of Maintenance and Flexibility in Working Memory 1

Chapter 2: The Fate of Prioritized and De-prioritized Items in Working Memory 1

Chapter 3: Temporal Dynamics of Attention During Encoding vs. Maintenance of Working Memory 84

Chapter 4: Oscillatory Brain State Predicts Variability in Working Memory 134

Chapter 5: Testing Sensory Evidence Against Mnemonic Templates 168

Chapter 6: Neural Mechanisms of Updating vs. Protecting the Contents of Working Memory 225

Chapter 7: General Discussion 268

Bibliography 280

Appendix to Chapter 5 306

Chapter 1: Neural Mechanisms of Maintenance and Flexibility in Working Memory

Chapter summary

Visual working memory is often thought of as the temporary storage of a sensory imprint in mind, a quickly fading snapshot that is kept to hand for later reference. Early neurophysiological studies were interpreted along these lines: memory cells in prefrontal cortex remain active as a form of representing stimuli in working memory. Subsequent studies clashed with this notion, supporting instead the view that these frontal cells were involved in a form of sustained attention towards representations kept in sensory cortex. The sustained internal attention view emerged as a prevalent model of how working memory is maintained in the brain. In the last ten years, neuroimaging and behavioral studies have begun to challenge this model too. They find that working-memory representations are much more flexible, and much more independent of sustained attention, than the sustained attention model would suggest. While attention can be deployed to improve the quality of remembered items at various stages from encoding to readout, attentional control operations seem to be transient, rather than sustained. Possibly, memory representations themselves are otherwise kept in a quiescent state through short-term modifications of a plastic network, requiring elevated activity only when new information requires their modification, recall, or their inhibition. This thesis will examine some of the neural and behavioral correlates of this surprising degree of flexibility in working memory.

Neural models of working memory maintenance – persistent activity, synaptic change, and matched filters

Working memory is the retention of a limited amount of information that will be useful to cognition or action in the near future. As such, working memory permits the representation of information over short delays without requiring the sustained presence of the original stimulus. Working memory is goal-dependent (Goldman-Rakic, 1987; Fuster, 1993), meaning that it represents information specifically for the task at hand: if information is currently irrelevant, it is ignored or banished from working memory, and information can be manipulated if it is not already presented in a format that can guide action or cognition. For example, integers can be maintained and added in working memory to determine whether they exceed a target number. This definition implies that working memory has some degree of flexibility in determining what gets encoded and how it is maintained. A prominent model has proposed that the working memory system can be separated into two parts: one part involved in the maintenance of information, and one part responsible for ensuring that information suits the current task (Baddeley et al., 1986). Much work on the neural basis of working memory has focused on the former aspect: how and where in the brain WM information is stored.

Early studies suggested that there might be a dedicated neural system for working memory, focused on the lateral prefrontal cortex (PFC). Much like the importance of the medial temporal lobes for *long-term* memory became apparent from the behavior of patients with lesions

(Scoville and Milner, 1957), behavioral deficits in patients or experimental animals with damage to the lateral PFC provided some of the first evidence that this brain area is necessary for working memory (Bianchi, 1895; Jacobsen, 1935; Luria, 1966). More precise lesion studies in macaques later confirmed that lesions to dorsolateral PFC impaired behavior in a delayed-saccade task (Funahashi et al., 1993) and in remembering hidden food locations (Butters and Pandya, 1969; Goldman et al., 1971; Passingham, 1985). Prefrontal lesions have also been shown to impair visual working memory (Pasternak et al., 2015).

Studies in humans elaborated that damage to lateral PFC may impair working memory control, rather than simple stimulus maintenance (Petrides and Milner, 1982; Tsuchida and Fellows, 2009; Barbey et al., 2013; Szczepanski and Knight, 2014). Passingham (1985) conceded that, while 'damage to frontal cortex impairs working memory [, …] it is not yet clear whether it does so by impairing the function of temporary stores or by disrupting central executive mechanisms' (p. 19). Earlier work had already supported a similar interpretation. For example, one study found that patients with PFC lesions showed intact memory performance when stimuli did not repeat across trials (Milner, 1964; cited in Postle, 2015a). However, when stimuli were reused, the same patients showed a deficit, consistent with the notion that PFC is involved in resolving interference from previously relevant stimuli (or distractor suppression more generally), but not in the mere short-term retention of the stimulus. Reusing stimuli across trials also discouraged the use of long-term memory strategies

(such as familiarity), which may have been used when stimuli were not repeated (see also Chapter 7).

Persistent activity as a mechanism of WM maintenance

Building on the lesion evidence, many studies have focused on the responses of single cells in the lateral PFC of macaques performing WM tasks. When monkeys held a location in mind for several seconds before a reach movement, many of these cells showed persistent increases in firing rate, especially at the onset of the delay (Fuster and Alexander, 1971). Since the existing lesion evidence attributed mainly attentional control functions to the PFC (Luria and Homskaya, 1964), the authors interpreted the persistent delay-period firing of PFC neurons as 'specifically related to the focusing of attention by the animal on information that is being or has been placed in temporary memory storage for prospective utilization' (p. 654, cited also in Postle, 2015). Another study published in the same year (Kubota and Niki, 1971) used a delayed alternation task in macaques, and also found a subpopulation of PFC cells that fired during the delay. However, they too were cautious not to interpret this activity as evidence of stimulus representation in PFC.

In spite of the caution of these two studies, subsequent authors increasingly interpreted delay-period firing of PFC cells as a neural signature of stimulus maintenance (Goldman-Rakic, 1987; Goldman-

Rakic et al., 1992). Early studies, however, contained a possibly important confound, in that most tasks allowed for WM information to be coded as saccade plans (Funahashi et al., 1989), or as sustained spatial attention. The supposedly mnemonic function of PFC firing, therefore, could simply reflect the preparation of an upcoming reach (Fuster and Alexander, 1971) or saccade (Funahashi et al., 1989; 1990). A second wave of experiments controlled for this by teaching animals to remember arbitrary visual information in a delayed match-to-sample task (Miller et al., 1996a). Presumably, this precluded any item-specific movement preparation because the animals could not prepare a specific response for a given memory stimulus. Nonetheless, PFC neurons showed persistent item-specific activity in the delay. Importantly, delay activity returned even after the presentation of non-target distractors, demonstrating that activity was task-dependent. By contrast, activity specific to the remembered item in inferotemporal neurons does not return after presentation of a distractor (Miller et al., 1993; 1996b), suggesting that a distractor-resistant representation is preserved in PFC even when other brain areas revert to processing new input (Suzuki and Gottlieb, 2013). Neuroimaging studies, using fMRI, later confirmed that distractor-resistant delay-period activity also occurs in the lateral PFC of humans (Courtney, 1998; D'Esposito et al., 1999; Sakai et al., 2002).

Subsequent studies further established that PFC populations encode information about multiple items remembered simultaneously (Warden and Miller, 2007; Buschman et al., 2011a). This is a crucial

finding, given that humans (Luck and Vogel, 1997) and monkeys (Buschman et al., 2011a; Lara and Wallis, 2012) can remember more than one item at a time. Any brain region that is a candidate for the storage of WM items should therefore be able to retain information about more than one item. It also speaks to the possible confound that PFC is involved mainly in WM control, not maintenance: it seems unlikely that prospective attention or top-down control would be directed towards multiple items in WM simultaneously. Nonetheless, this notion cannot be ruled out entirely, as discussed in the next section.

Persistent activity in PFC as sustained attention to WM representations in sensory cortex

While some work cited in the previous section argued emphatically that the prefrontal lobes store working memory *contents*, at least when the task does not require the memorization of precise sensory details, an alternative account for the role of PFC in many working memory tasks emphasizes its role in control. Instead of storing contents, it may be responsible for the effective management of WM, by controlling which items need to be encoded or discarded, which parts of sensory cortex are currently engaged in maintenance, and which item is of particular relevance to upcoming behavior. Under this scheme, the task of storing actual content is relegated to sensory brain areas.

A number of frameworks have argued that at least some of the PFC activation found in WM delay periods is involved in such top-down

control (Petrides and Baddeley, 1996; Petrides, 2002; Curtis and D'Esposito, 2003; Awh et al., 2006; Sreenivasan et al., 2014a). Under this reframing, PFC activity reflects a number of control operations that are executed during the delay. For example, some authors have noted the striking overlap in the prefrontal and parietal brain regions that are involved in the maintenance of spatial locations in WM and the control of visuo-spatial attention in perceptual tasks (Awh et al., 1998; 2006). This has led to the proposal that spatial WM information is maintained (or rehearsed) via sustained attention (Kiyonaga and Egner, 2013), under prefrontal (and parietal) control. Some evidence for this account comes from dual task studies showing that, when a location is held in working memory, distractors presented there during the delay receive preferential processing (Awh et al., 2000). The converse of this argument is that the contents of working memory always guide attention towards matching sensory input (Chelazzi et al., 1993; Duncan and Desimone, 1995; Soto et al., 2008).

By extension, others have argued that PFC activity 'is an extra-mnemonic source of top-down biasing control over posterior regions that actually store the representations' (Curtis and D'Esposito, 2003). Curtis and D'Esposito also pointed out that dlPFC, with its extensive connections to premotor cortex and frontal eye fields, is likely involved in motor planning and action selection. An even broader view has been made for the function of PFC delay activity (Postle, 2006), arguing that WM is an emergent property of the cooperation of sensory and motor

areas involved in maintenance of stimulus information, and of fronto-parietal control areas involved in attending to that information, protecting it from interference, and manipulating it to suit the task. Importantly, the functions served by the PFC during the delay can also be applied in non-WM contexts. Further, WM content maintenance here is solely done in the sensory or motor brain area that is specialized for responding to the kind of information that is being remembered.

These alternative models of PFC delay activity necessitate evidence that WM information can be represented in sensory and heteromodal cortical areas outside the PFC. The bulk of the evidence for this comes from studies on the representation of sensory aspects of memoranda in corresponding sensory cortex (Pasternak and Greenlee, 2005). Again, early evidence for this proposal came from delay-period activity in visual cortical neurons of the macaque (Fuster and Jervey, 1981; Miller et al., 1993; Miller and Desimone, 1994; Chelazzi et al., 1998; 2001). Subsequently, human neuroimaging studies have shown that information about visual stimuli retained in WM can be read out from delay-period activity in early visual areas (Harrison and Tong, 2009; Serences et al., 2009; Sneve et al., 2012), and that the magnitude of BOLD activity (Sligte et al., 2009) or the strength of item decoding in visual cortex correlates with an observer's recall accuracy (Emrich et al., 2013). Correspondingly, disruptions to visual cortex via transcranial magnetic stimulation early in the delay decrease the precision of recall from WM (Harris et al., 2002; Silvanto and Cattaneo, 2010; van de Ven et

al., 2012), although this may have happened because of interference with visual processing during encoding, rather than with maintenance itself. A more recent study (Zokaei et al., 2014) found that disrupting area MT during the retention interval disrupts memory of a motion stimulus that is in the focus of attention (see below).

Outside primary sensory areas, load-dependent or content-specific activity has also been observed in human (Todd and Marois, 2004; Vogel and Machizawa, 2004; Xu and Chun, 2006) and macaque parietal cortex (Friedman and Goldman-Rakic, 1994; Constantinidis and Steinmetz, 1996; Pesaran et al., 2002). The parietal cortex has also been identified as the main source (McCollough et al., 2007; Becke et al., 2015) of contralateral delay activity (CDA). In WM delay intervals requiring the maintenance of stimuli in one visual hemifield, EEG voltages contralateral to the attended side are consistently more negative during the delay (Vogel and Machizawa, 2004). The CDA has been shown to increase linearly with load up to an observer's capacity limit (Vogel and Machizawa, 2004; McCollough et al., 2007), making it a strong predictor of individual task performance (Vogel and Machizawa, 2004; Vogel et al., 2005; Luck and Vogel, 2013). This indicates that parietal cortex plays an important role in WM maintenance. Parietal areas might serve as an intermediary between top-down mnemonic control in PFC and maintenance in early sensory areas (Jacob and Nieder, 2014; Mendoza-Halliday et al., 2014). Nevertheless, an alternative account of the parietal involvement, and of the CDA in particular, could be its function in guiding

spatial attention. The intraparietal sulcus is known to contain spatial priority maps that are important for sustaining covert visual attention and for planning saccades and other movements (Bisley and Goldberg, 2010). Therefore, these measures might be non-mnemonic, and simply reflect anticipation of target stimuli, or sustained attention in the absence of cues directing attention elsewhere (Stokes, 2011).

Functional connectivity and neural synchronization as mechanisms of WM maintenance

Postle's model (2006) already implied, but did not spell out, that prefrontal control areas and sensory cortex form a temporary network to solve the current WM task. If task demands changed, the network would quickly adapt to incorporate relevant areas. A more recent revision of the model by D'Esposito and colleagues (Sreenivasan et al., 2014b) has elaborated that PFC may use top-down connectivity to sensory areas involved in WM maintenance to bias their activity in favor of task-relevant stimuli (and in order to suppress distractors). Evidence for this proposal comes from neuroimaging studies that measure functional connectivity between PFC, parietal areas, and sensory cortex (Sakai and Passingham, 2002). These have shown that connectivity, particularly between lateral PFC and the sensory region sensitive to the remembered stimulus (e.g., the fusiform gyrus for face stimuli), increases during the maintenance delay (Gazzaley et al., 2004; Gazzaley and Nobre, 2012). When cues prior to encoding indicate which stimuli are to be encoded into WM, the

strength of this connectivity is predictive of visual cortical BOLD responses to relevant stimuli during encoding (Bollinger et al., 2010). When cues presented during maintenance direct attention to one item in WM, connectivity is predictive of behavioral accuracy (Kuo et al., 2011).

In line with this view, causal manipulations using TMS have been shown to influence task-selective activity in sensory cortex (Feredoes et al., 2011; Zanto et al., 2011). Similarly, an earlier study of macaques performing a delayed-saccade task showed that temporary inactivation of dlPFC led to a reduction in delay activity in location-coding parietal neurons (Chafee and Goldman-Rakic, 2000). Similarly, humans with prefrontal lesions show reduced ipsilesional visual responses in a working memory task (Miller et al., 2010; Voytek and Knight, 2010).

The neural basis of this connectivity increase is poorly understood, but a plausible explanation is that neural populations in the two areas synchronize. For example, this can be achieved by locking spike times to specific phases of neural oscillations that are present in the local field potentials of both areas, and which are inter-regionally coherent. Such synchronization between prefrontal and visual cortex has been demonstrated for tasks requiring sustained top-down attention (Miller and Buschman, 2007; Gregoriou et al., 2009; Bastos et al., 2015). Evidence for neural synchronization during WM delays is scarce, but has been observed in macaques, where prefrontal and visual areas synchronize in the theta-range (Liebe et al., 2012). Interestingly, the pattern of beta-band (15-30 Hz) coherence between lateral PFC and lateral parietal cortex in

the macaque has been shown to encode remembered visual object categories (Salazar et al., 2012). Such patterns have been shown to reflect (using patterns of BOLD correlations across areas) the nature of the current task (Cole et al., 2013; 2014), but evidence for their modulation in relationship to the maintained content remains to be tested thoroughly.

In addition to changes in long-rang synchrony, some brain areas also change the degree of local synchronization during WM maintenance. Most notably, theta-band oscillations increase in power over midline prefrontal cortex (Jensen and Tesche, 2002; Jensen et al., 2002) during the retention interval, possibly reflecting the continued recruitment of these areas for top-down control. Interestingly, posterior visual areas tend to increase or decrease the degree of alpha-band (8-12 Hz) synchronization depending on whether the task requires ignoring (Jensen et al., 2002; Tuladhar et al., 2007; Haegens et al., 2012) or maintaining of visual information (Jokisch and Jensen, 2007; Medendorp et al., 2007).

Stimulus representation in PFC versus sensory cortex may depend flexibly on task demands

Given that visual distractors (Miller et al., 1993) can eliminate delay-period activity in visual cortex, it is not clear how sensory areas could maintain distractor-resistant WM representations (Super et al., 2001). Perhaps sensory cortex delay activity is epiphenomenal. That is, persistent activity in sensory cortex may simply reflect the last stimulus

until it is overwritten by the processing of new input. This kind of stimulus-dependent hysteresis has been observed across many brain areas (Hasson et al., 2015) and could form one neural basis of short-term memory when distractor-resistance is not required. More generally, it seems plausible that sensory cortex may be sufficient for the maintenance of fragile WM representations that are not distractor-resistant, while PFC is needed (either to represent content or to override the distraction in sensory cortex) for the robust maintenance of distractor-resistant WM (Sligte et al., 2008; 2009; Clapp et al., 2010). In further support of this, persistent representation of stimulus information in sensory areas occurs only when maintaining it is necessary for the task (Riggall and Postle, 2012). While task-specificity may be necessary, it is not sufficient to show that visual cortical activity is necessarily relevant to WM maintenance. Instead, it could simply reflect anticipatory attention towards the expected features of the probe stimulus, or mental imagery of the remembered item (Stokes, 2011).

An alternative explanation for delay-period activity in sensory areas has been proposed: whenever the task demands precise memory for particular sensory details that cannot be easily abstracted, this information is stored in sensory cortex, using the same populations that are sensitive to these stimulus properties during encoding (Pasternak and Greenlee, 2005). Some early behavioral evidence supported this view (Magnussen, 2000). For example, when observers maintain precise visual information in WM (such as the spatial frequency of a grating), then visual

distractors presented during the delay have a stronger effect on performance if they match the remembered item along a task-relevant dimension (Magnussen et al., 1991; Magnussen and Greenlee, 1992; Zaksas and Pasternak, 2001). This proposal has been substantiated by human neuroimaging evidence (Lee et al., 2013). The authors showed that BOLD levels were elevated in lateral PFC and in object-selective lateral occipital cortex (LOC) while observers maintained visual objects in WM. However, the object identity could be decoded from LOC only when a precise visual decision was required at the end of the delay. Conversely, PFC encoded object identity only when the (verbalizable) object identity needed to be maintained. While this proposal has not been tested in detail, it seems possible that the precise nature of the task might also influence the degree of delay-period connectivity between WM-relevant brain areas.

WM maintenance via short-term plasticity

The preceding sections have attempted to highlight that sustained delay-period activity in PFC, parietal cortex, or sensory cortex need not be an essential substrate of WM maintenance, even where it correlates strongly with behavior. This is because almost all cited measures could be confounded with anticipatory attention, imagery, or motor preparation (which is not surprising, given that many definitions of WM draw parallels with these functions). In many other cases, correlations between delay activity and behavior are rather weak. V1 pattern decoding, for instance,

seems to correlate mainly with subtle fluctuations in WM precision, rather than whether an item is entirely remembered or forgotten (Emrich et al., 2013). Therefore, a proposal has come up that delay-period activity is not necessary for successful maintenance, but that it may, in some circumstances, be epiphenomenal (LaRocque et al., 2014; Stokes, 2015). Some initial evidence for this view comes from a number of studies requiring memory for more than one item. Importantly, they show that delay period activity for an item can be abolished even when it is remembered accurately. For instance, Watanabe and Funahashi (2014) used a dual-task paradigm in which monkeys remembered a saccade target while simultaneously attending a different location for a secondary task. PFC activity coding for the remembered saccade location emerged only after the secondary task had been completed, half-way through the delay interval. This demonstrated that, at least in PFC, delay-period activity is not strictly necessary for successful memory. This finding seems to be at odds with a study showing that persistent activity in PFC is resistant to distractors (Miller et al., 1996). However, even in that study, distractors caused a transient, memorandum-independent increase in PFC firing rates that made spike rates in response to preferred memoranda indistinguishable from spike rates to non-preferred memoranda. Nonetheless, firing rates quickly returned to coding the remembered stimulus, perhaps because attention could be withdrawn from the distracting input after it was no longer on the screen. Using human imaging, Postle and colleagues (Lewis-Peacock et al., 2012;

LaRocque et al., 2013; see also Zokaei et al., 2014) have similarly argued that items in WM may be retained without persistent activation.

If WM contents can be retained without sustained firing, then a likely alternative storage candidate is that they are retained in temporary changes in synaptic weights. Longer-lasting changes in synaptic weights are generally considered to be the neural correlate of long-term memory (Hebb, 1949; Hopfield, 1980; Bliss and Collingridge, 1993). Short-term synaptic plasticity may generally underlie memory over a variety of time scales (Hasson et al., 2015), from long-term memories to those lasting only a few seconds. Biophysical mechanisms ranging from temporary increases in calcium concentration at stimulated synapses (Mongillo et al., 2008) to temporary synchronization of content-specific neural ensembles (Buschman et al., 2011; Salazar et al., 2012) have been proposed, but their function is always similar: an otherwise flexible neural network must be modified in such a way that subsequent processing of new input reflects the task-relevant information laid down in working memory.

In contrast to the newer proposals of how silent memory may work *in the brain*, this readout-focused view at the level of *behavior* has been present (but somewhat neglected) for decades. For instance, Heath (Heath, 1982) proposed a computational model of working memory centered on the engineering concept of a matched filter. In his framework (based on an original proposal by Anderson, 1973), memories need not be remembered veridically, as an internally sustained percept. Rather, the

functional importance of the contents of WM is that they process any new stimulus so that it leads to an appropriate behavioral response. In other words, probe stimuli must be filtered by a network that effectively compares their task-relevant aspects to those of the remembered stimulus. Matched filters achieve exactly this by responding more strongly to inputs that are sufficiently similar to the 'template' encoded in the filter. In addition, matched filters have the desirable property that any *deviations* between the template and the probe stimulus can be read out as a mismatch response (Rao and Ballard, 1999). Knowledge about the exact nature of the mismatch is required in many situations. For instance, the direction of the mismatch between a probe and a remembered stimulus needs to be judged when visually identifying the ripest pieces of fruit in the supermarket (or, in psychophysics, when making delayed discrimination judgments).

Neural evidence for mechanisms of working memory that do not depend on persistent activity is still limited. Some indirect evidence for 'silent' WM maintenance via changes in to the hidden state of a network comes from studies in animal models. A study in rodents found that pairwise connectivity between neurons in the prefrontal cortex changed temporarily during the trial as a function of the animal's remembered action plans (Fujisawa et al., 2008; Harvey et al., 2012). Similar, indirect evidence for such flexible changes in connectivity comes from recordings in the PFC of macaques (Crowe et al., 2010; Stokes et al., 2013). The latter studies inferred changes in the hidden state (possibly via short-term

plasticity) by observing that content-specific coding in the PFC population changed dynamically between encoding and maintenance. Evidence for similar changes to the response properties of the network has been observed in visual cortex, during the processing of visual stimuli (Meyers et al., 2008; Sugase-Miyamoto et al., 2008; Nikolić et al., 2009). As with many other studies of sensory cortex, it is not clear whether the encoding of previous stimuli in visual cortex was truly task-dependent (and therefore a reflection of working memory), or simply a passive reflection of stimulus history. For instance, Nikolić and colleagues found that stimulation evoked neural responses in V1 that were patterned by the preceding stimulus (as in a matched filter), even though the test subjects were anaesthetized at the time.

In sum, there seems to be ample evidence for persistent coding for WM contents, and at least some indication that WM contents can also be maintained 'silently', e.g., in temporary synaptic weight changes. Instead of supposing that only one of these mechanisms is in fact used by the brain, the multiple mechanisms might reflect different representational states (LaRocque et al., 2014; Stokes, 2015). This thesis returns to the issue of silent coding of WM contents in chapter 5. I will describe an experiment using magnetoencephalography in humans to decode the memorized visual template used in a detection task. The results are consistent with the temporary reactivation of information that is otherwise stored in a silent code, consistent with the idea that WM does not always depend on persistent activity.

Neural mechanisms of selection into working memory

The previous section explored the neural basis of continued maintenance of items in working memory. It did not address one of WM's chief virtues: flexibility. It is becoming increasingly clear that attentional functions have a large degree of control over which items are encoded in WM, and which items are prioritized during maintenance. Some behavioral research described visual WM representations as relatively inflexible: an item was either encoded or not, but once encoded, the representation was fairly fixed (Luck, 2008). Work by Sperling (1960) did show that observers have some degree of control over which items to encode when cues are presented within a few hundred milliseconds after the disappearance of the memorized stimulus array. However, since the ability for cue use was so short-lived, this storage system was described as a separate system, *iconic memory* (Neisser, 1967), from which only a limited amount of information can be transferred into the more robust but less flexible working memory format. Building on these early insights, more recent research has shown that top-down attention can be used in the service of improving WM, similar to how it is used to improve perception.

Perceptual Effects on WM Selection – Early Gating

Much of what is known about the role of PFC and parietal cortex in top-down control comes from studies of covert attention. Anticipatory cues indicating that a particular location in space is likely to contain a

visual target (Kastner et al., 1999; Bressler et al., 2008), or that a particular visual feature is relevant (Saenz et al., 2002; Boynton, 2005), consistently activate a set of heteromodal brain areas, including the dorsolateral and ventrolateral PFC, the intraparietal sulcus, and the superior parietal lobule (Corbetta et al., 1993; Nobre et al., 1997; Gitelman et al., 1999; Corbetta and Shulman, 2002).

It stands to reason that the same network of lateral prefrontal and superior parietal regions might also be important for the selection of items into working memory (Gazzaley and Nobre, 2012). Functional imaging studies in humans have shown that lateral prefrontal areas (Gazzaley et al., 2007; Zanto et al., 2010; Bollinger et al., 2010b) activate when task-relevant items (e.g., images of faces) need to be picked out from a stream that includes distractors (outdoor scenes for example). Importantly, while there seems to be some overlap between frontal regions recruited for perception and for WM, the overlap is not total. For instance, Gazzaley and colleagues consistently find that a region in ventrolateral PFC is involved in anticipatory control of encoding into WM, whereas this area is not typically involved in the anticipatory control of visual attention. Disrupting ventrolateral PFC during encoding via TMS leads to decreases in WM performance when distractors are present (Zanto et al., 2011). These prefrontal effects have been shown to modulate sensory cortex during encoding, by increasing the BOLD response in feature-selective areas (such as the parahippocampal place area when images of outdoor scenes need to be encoded), and decreasing the response to distractors

(Gazzaley et al., 2007; Zanto and Gazzaley, 2009). This kind of biasing mechanism has been proposed as a way of preferentially encoding relevant items as they compete with distractors for the limited space in working memory (Bundesen, 2005). In chapter 3, the question of preparatory cueing for the preferred encoding of items into visual WM is examined using EEG.

In attention for visual perception, modulation of sensory cortex has been shown to be sensitive to subtle modulations in task relevance. That is, attention does not act as a binary on/off switch that biases processing resources either towards or away from a category. Instead, for example, spatial attention can also be graded, reflecting different levels of certainty that a relevant item will appear at a certain location (Gould et al., 2011). In this context, modulations in alpha-band (8-12 Hz) oscillations have been shown to influence the excitability of visual areas (Worden et al., 2000; Bollimunta et al., 2011; Gould et al., 2011). The amplitude of alpha oscillations in sensory cortex is usually interpreted as a marker of where attention is deployed (Sauseng et al., 2005). When amplitudes are low, firing rates of visual neurons go up (Bollimunta et al., 2011; Haegens et al., 2011), and the likelihood of perceiving low-threshold stimuli increases (van Dijk et al., 2008; Lange et al., 2013).

Therefore, the state of alpha oscillations in the visual systems has been interpreted as a robust marker of ongoing fluctuations in visual excitability. In contrast to this broad range of responses observed in perception, working memory has until recently largely been described as

an all-or-none phenomenon: items are either encoded into working memory, or they are forgotten. Recent work using statistical models to evaluate the variability in recall from visual WM, however, has found evidence that the fidelity of WM representation may vary more from trial to trial than would be expected by chance. Neuroimaging studies showing that the stimulus-specificity of activation patterns in visual cortex correlates with the precision of visual WM (Emrich et al., 2013; Ester et al., 2013) are consistent with this notion. However, whether the fidelity of WM representations might vary continuously, as a function of the excitability state of visual cortex during encoding, has not been thoroughly explored at the neural level. Chapter 4 examines this question in more detail.

Importance of PFC and Basal Ganglia for Selection – Late Gating into WM

Studies examining top-down effects on sensory cortex during WM encoding usually argue for an early selection mechanism into WM. An alternative approach has argued that visual processing reaches a late stage before selection, or gating, into working memory needs to occur (Braver et al., 1999). In general, these studies have emphasized the distractor-resistance of working memory, stressing the importance of gating as a means of protecting relevant contents from irrelevant interference. Gating is hypothesized to result from the phasic release of dopamine into prefrontal cortex. By temporarily lowering the threshold needed for input to cause sustained responses in PFC, dopamine toggles

whether the current WM representation should be replaced by new information (updating), or whether it should be protected from interference. A commonly used paradigm to investigate this hypothesis is the continuous performance task: a stream of letters appears, one by one, and must be evaluated according to a stimulus-response rule that is maintained in working memory (e.g., 'letter X is the target, but only if it immediately follows letter A'). Some letters in the stream, while they are not targets themselves, act as cues to change the current stimulus-response rule. For instance, after seeing the letter 'A', a letter 'X' is a target, whereas it is a non-target following any other letter. Presumably, changing the current rule in response to a stimulus relies on retrieval of that rule into WM from some longer-term memory store that retains all possible rules for the current task. Foregrounding of latently activated rules into the focus of attention may also be interpreted as a kind of internal attention shift (see below).

Importantly, the representations of such rules in WM may not be equivalent to the representation of other content (such as the shape of visual stimuli). For instance, Braver and colleagues (1999) argue that rule representation (or, in their phrase, 'context representation') is only 'one component of WM. Specifically, context can be viewed as the subset of representations within WM that govern how other representations are used. Representations of context are particularly important for situations where there is strong competition for response selection' (p. 315). Under this description, context might be represented within the lateral PFC.

Some evidence for this model comes from an fMRI study showing activation of dopaminergic brainstem during context updates that led to increases in dlPFC activity (D'Ardenne et al., 2012).

One problem with the theory that dopamine controls the gate into WM representation in PFC is that it is non-selective. Since dopaminergic nuclei project diffusely into large areas of cortex (Beckstead et al., 1979; Swanson, 1982), they seem ill-suited to targeting specific sub-regions when only one out of several WM representations needs to be replaced (Badre, 2012). Item-specific gating mechanisms might exist in the connections between the basal ganglia (BG), the thalamus, and the PFC (Frank et al., 2001). These projections to PFC are still fairly coarse, though, and therefore still unlikely to target item-specific subpopulations if they are closely overlapping with populations representing other information. This is further complicated because many dlPFC neurons exhibit mixed selectivity, responding to multiple items in often complex interactions (Mante et al., 2013; Rigotti et al., 2013).

There is some evidence that behavioral rules at different hierarchical levels are represented by distinct populations in PFC, with rules pertaining to immediate behavior being represented more caudally, close to the primary motor cortex, while more abstract rules are represented higher up in the hierarchy, in rostral PFC (Koechlin et al., 2003; Badre and D'Esposito, 2009; but see Crittenden and Duncan, 2014). There is some evidence that structural connectivity to the basal ganglia is distinct for these clearly separated PFC regions (Verstynen et

al., 2012), and some functional imaging studies find similar dissociations between lower-level rule updates activating posterior PFC and higher-level updates activating anterior PFC (Nee and Brown, 2012a; 2012b). Basal ganglia mechanisms seem to be useful for item-specific gating *only* when items are clearly separable roles in the prefrontal hierarchy. When multiple items in WM are comparable, a more sophisticated mechanism might be necessary.

In sum, while the BG seem to be important to the updating of some rules in WM, they appear to be only one part of a wider network that includes dopaminergic nuclei, the PFC, and the parietal cortex. Converging evidence for the involvement of this wider set of regions comes from the rule switching literature. Although these studies did not explicitly examine WM updating, they were nonetheless strikingly similar. For example, a common rule-switching task requires that observers monitor a stimulus stream and evaluate each stimulus according to one dimension (e.g., its color), until a cue instructs them to switch to an alternate, previously learnt stimulus-response rule. These switch-cue events presumably recruit a similar network to that involved in updating the context in the continuous performance task. Unsurprisingly, switches tend to activate similar lateral prefrontal areas (Rushworth et al., 2002a; 2002b; 2005). Whether similar networks are recruited in the updating of visual content into WM is examined in chapter 6.

However, they also show activation in the anterior cingulate and pre-supplementary motor area, two regions that are often found in

response to motor control in the face of conflicting behavioral rules (Botvinick et al., 2004). Importantly, these studies have highlighted a large degree of conflict in overcoming the previously active rule after a switch. This proactive interference leads to behavioral costs that have been interpreted as a form of task set inertia (Allport et al., 1994). In other words, proactive interference arises because conflict between old and new rules is only completely resolved at the time when the new rule is acted on for the first time. Whether similar inertia exists in working memory, and how proactive interference could speak to the organization of WM, is a relatively unexplored domain (Jonides and Nee, 2006). This topic is revisited in the general discussion in chapter 7.

Input Gating and the Functional Role of WM Representations in PFC

In both early updating and late updating frameworks, even though PFC is interpreted as a controller, it must maintain some content-specificity (e.g., whether faces or houses are currently relevant, or whether X's or Y's are the current target). This rule or category selectivity in PFC might correspond to the successful decoding of WM contents from PFC recordings in macaques. The reason that PFC populations in the latter studies might be so selective is that the animals are extensively trained on the particular stimulus materials (often, naturalistic images) that they need to retain. This practice might have led to a reorganization of the arbitrary stimulus into an abstracted categorical code (if that well-known image comes up again, respond with a match response). It would be

interesting to see whether content representation in PFC is eliminated for stimuli that are harder to abstract. In the case of color, this has recently been found (Lara and Wallis, 2014; but see Mendoza-Halliday et al., 2014). The authors trained monkeys to remember several colors and make a fine-grained color discrimination after a delay. They found that, while neural population activity in visual cortex maintained significant color information in the delay, PFC populations did not (even though they did represent the locations of the memorized color patches). This indicates, together with the fMRI evidence cited above (Christophel et al., 2012; Riggall and Postle, 2012; Lee et al., 2013) that PFC may only be suitable to representing category-level information about stimulus material. In other words, dorsolateral PFC can only encode information if it can be represented as a more abstract, rule-like representation.

Changing the representational status of items within working memory

There were some early contradictions of the prevailing model that WM, once encoded beyond the iconic stage, was relatively inflexible. For example, some broader proposals for the structure of higher-order cognition acknowledged the occasional need to interrupt a train of thought in order to complete a more pressing task. Their models needed to allow for the temporarily backgrounded thought or action plan to return to the foreground once the interrupting task was completed (Atkinson and Shiffrin, 1968; Baddeley et al., 1986).

The more recent model by Nelson Cowan (Cowan, 2000) built on this notion by supposing that WM contents can exist in two different representational states: mere maintenance requires them simply to be part of activated long-term memory, whereas a subset of WM items can inhabit the focus of attention, a low-capacity region that contains items more robustly and for immediate use in a task. Importantly, behavioral studies based on this framework emphasized the flexibility in shuttling items between different states during the delay. For instance, Klaus Oberauer (Oberauer, 2002; 2005) has shown that some items in WM can be temporarily foregrounded, resulting in speeded reaction times when tested on those items. These studies drew the additional conclusion that only one of these items can be kept in a prioritized format because of its immediate relevance to behavior (Oberauer and Hein, 2012). This item is functionally distinct from those retained in the background in that it

guides attention to matching items in the environment (Olivers et al., 2011). Internal attention may play an important role in determining which item is currently selected for prioritization. As discussed below, the distinction between attended and unattended items in WM might map onto the distinction between active and silent modes of neural representation. Sustained attention, therefore, may explain many of the findings of persistent delay-period activity in sensory areas and in areas usually associated with top-down attention control (Postle, 2015b).

Neural control of attention to WM representations

Much like cues direct attention to improve the perception of an upcoming stimulus, cues presented retrospectively during the maintenance delay can change the priority status of an item in WM. By indicating that it is more likely to be probed they increase the utility of remembering that item. Early studies found that cues presented during the maintenance delay, but after sufficient time for iconic memory effects to have subsided (Sperling, 1960; Phillips, 1974), increased accuracy and decreased reaction times in response to a probe that needed to be judged as a match or mismatch (Griffin and Nobre, 2003; Landman et al., 2003). Even incidental reminders to 'refresh' a representation whose behavioral relevance is otherwise unchanged can increase explicit long-term memory (Johnson et al., 2002), although they lead less consistently to improvements in the short term (Johnson et al., 2002; 2013).

Attention to representations in WM activates brain areas that are also relevant to the top-down control of attention towards external stimuli (Nobre et al., 2000; Nobre, 2004; Chun and Johnson, 2011; Chun et al., 2011), such as the lateral prefrontal cortex and the intra-parietal sulcus (Nobre et al., 2004; Lepsien et al., 2005; 2007; Yeh et al., 2007; Bledowski et al., 2009; Kuo et al., 2014; Raye et al., 2002; Johnson et al., 2005; Roth et al., 2009). Retrospective cues also activate additional areas that are not activated by externally-guided attention shifts, predominantly in the frontal cortex. Additionally recruited areas are fairly wide-spread, and include the anterior insula, anterior cingulate, and the pre-supplementary motor area, or pre-SMA (Nobre et al., 2004; Johnson et al., 2005; Lepsien et al., 2005; Roth et al., 2009; Henseler et al., 2011; Higo et al., 2011; Kuo et al., 2014), as well as lateral prefrontal cortex (Nee and Jonides, 2009; Henseler et al., 2011). Some studies have also implicated the frontal pole in either the guidance of attention to WM representations (Henseler et al., 2011) or as a 'gateway' for switching between externally- and internally-guided attention (Burgess et al., 2007; Benoit et al., 2012). Conversely, some studies have found that externally-guided attention more strongly activates the visual cortex and the human analogue of the frontal eye fields (Nee and Jonides, 2009; Roth et al., 2009).

Sensory areas specialized for the currently remembered content (such as the fusiform face area and the parahippocampal place area) are also modulated by refreshing cues (Johnson et al., 2007; Yi et al., 2008; Park et al., 2010) and by retrocues (Lepsien and Nobre, 2007; Sligte et al.,

2009; Lepsien et al., 2011; Munneke et al., 2012; Kuo et al., 2014). In fact, presenting a second cue prompting attention to switch to another category or item leads to a switch in activity in category-selective visual cortex (Lepsien and Nobre, 2007; Lewis-Peacock et al., 2012; 2014). It stands to reason that items that have been prioritized are still retained in the area of sensory cortex that is also involved in processing it in other contexts. Prefrontal areas, particularly the frontal operculum, seem to exert top-down internal attention through connectivity to the relevant sensory brain area (Higo et al., 2011; Nelissen et al., 2013).

In contrast, the temporal dynamics of activation to prospective and retrospective cues have not been explored as extensively. In particular, while alpha-band lateralization in anticipation of a stimulus is well understood, its role in internal attention shifts is fairly unknown. It seems reasonable to expect some differences in timing, because in internal attention shifts the cued information is immediately accessible (since it has already been encoded in WM), whereas the former requires waiting for the future arrival of relevant input. This question is further explored in chapter 3.

Behavioral benefits of retrospective cueing

Behavioral studies have begun to look more closely at the nature of the retro-cue benefit. A number of explanations have emerged. They include the protection of cued items from decay or overwriting in fragile memory, reduced susceptibility to interference, increased anticipation of

the probe item, and the allocation of additional mnemonic resources to the cued item. In all likelihood, several of these mechanisms are at play, and the exact cause of the benefit will vary from task to task.

In favor of the item-protection idea, one study used discrimination judgments to construct psychometric curves, permitting the dissociation of possible retro-cueing effects on the slope or the lapse rate of the curve (Bays and Husain, 2008; Murray et al., 2013). The slope of the psychometric curve has been interpreted as an index of the fidelity of a representation because it measures sensitivity to only small feature changes. Lapse rate, conversely, has been taken to measure the level of random responding, indicating on what proportion of trials an item was completely forgotten. The study found that retro-cues primarily reduce the lapse rate. By extension, retro-cues decrease the likelihood of forgetting an item without boosting its fidelity. Murray and colleagues (2013) also compared retro-cue trials with cue-absent short trials, on which the probe was presented at the same delay as the cue. They found that cues increased the likelihood of recall even beyond that of the short trials, and concluded that information can be 'rescued' from a representational store that is normally inaccessible. While protection from decay over time cannot account for the benefit seen on those trials, it has been proposed as an additional mechanism of boosting the representation of the cued item (Pertzov et al., 2013). Incidentally, this study found that benefits (as well as costs to invalid items) increased with

increasing delay. This means that the cueing benefit cannot merely be the effect of increased expectation of the location of the probe stimulus.

Similarly, an early study (Sligte et al., 2008) proposed that retro-cueing benefits may occur only when fragile items in VSTM, which would otherwise be subject to decay by interference, are transferred into a more capacity-limited, robust memory store that is distractor-resistant. The authors used pattern masks that were presented either before or after a retro-cue and found that cues benefitted performance only when they appeared before the masks, presumably when fragile VSTM had not yet been overwritten by the masks. However, subsequent studies have shown that cues can improve performance even after presentation of masks (Makovski, 2012). Therefore, cueing effects cannot be accounted for only with a transfer of information from fragile to robust VSTM. One explanation for the discrepant masking effects may have been the confusability between the masks and the WM items in the former study (Sligte et al., 2008), which could have led to erroneous recall of mask orientations.

Another possible mechanism is that retro-cues reduce random responses by allowing observers to anticipate more accurately the probe stimulus. Anticipatory feature-based attention has been shown to increase the neural (Maunsell and Treue, 2006) and behavioral (Carrasco et al., 2009) fidelity of matching responses to a stimulus. Unfortunately, increased probe anticipation is difficult to examine experimentally because features of the cued item and of the probe are usually

correlated, leading to a possible confound in many designs. However, one proposal suggests that retro-cueing allows the observer to protect the cued item from interference caused by the probe itself (Makovski et al., 2008).

Finally, retro-cues may also be the result of a reallocation of resources. WM capacity is clearly limited, and may consist of a shared resource pool that is distributed over all currently remembered items (Bays and Husain, 2008; Ma et al., 2014). If an item in WM increases in relevance, and therefore demands a larger share of the available resources, does this come at a cost to other items? Evidence is conflicting. Early studies found no decrease when a second, uncued item was probed on a retro-cue trial, compared to a neutral trial (Landman et al., 2003). Subsequent studies using invalid cues, however, did find substantial costs in remembering uncued items (Astle et al., 2012). Some even found that, for large array sizes, costs occur when there are no measurable benefits to cued items (Astle et al., 2012). These costs appear to increase with increasing delay (Pertzov et al., 2013), making it unlikely that they are related solely to the appearance of an uncued probe at an unexpected location (see also van Moorselaar et al., 2015). In contrast to invalid cue paradigms, other approaches have found negligible cueing costs, even when retro-cueing benefits were substantial. For example, a second cue reprioritizing a previously uncued item in WM can either cancel out the prior cueing costs (van Moorselaar et al., 2015), or lead to a cueing benefit of similar magnitude to a single cue trial (Rerko et al.,

2014). In both of these cases, observers knew that uncued items might become relevant over the course of the trial, which could have led to a strategy that minimizes forgetting. In contrast, Williams and Woodman (Williams et al., 2013) attempted to reduce the relevance of uncued items to zero (by only probing them once at the end of the experiment, or by probing them very infrequently). They were able to show that, in this context, uncued items can be forgotten entirely, even when initial arrays are below WM capacity (i.e., two items). Overall, resource redistribution cannot easily account for all of these effects.

Is retrospective cueing similar to feature-based attention, boosting the fidelity of the cued memory? While initial studies aiming to disentangle fidelity from recall likelihood (Murray et al., 2013) did not find any changes in fidelity, in part this null result may have been caused by the assessment methodology. Forced-choice change discrimination may not have been sensitive enough to derive accurate estimates of representational fidelity. An alternative mode of memory assessment is delayed estimation (i.e., cued recall) – observers indicate the feature they remember seeing, without any systematic input at the time of the probe. Delayed estimation methods also show noticeable decreases in guess rate after a retro-cue (van Moorselaar et al., 2014). Importantly, they additionally show small but significant improvements in memory fidelity (that is, a reduction in Gaussian noise in the random error of the responses, Williams and Woodman, 2012).

More generally, is sustained internal attention (whether feature-based or spatial) necessary for behavioral benefits? Somewhat surprisingly, benefits of retro-cues seem not to depend on sustained attention on the cued item until it is probed. For example, an uncued item can be probed before probing the cued item without reducing the magnitude of the cueing benefit (Rerko and Oberauer, 2013). Also, attention can be diverted to a second task (Janczyk and Berryhill, 2014; van Moorselaar et al., 2015) without abolishing the retro-cueing effect. It is only around the time of cue presentation itself that diverting attention has a detrimental effect (Janczyk and Berryhill, 2014), indicating that a central attentional resource may be necessary for the act of cue processing or re-prioritization, but not in the subsequent delay leading up to the probe.

Candidate neural mechanisms of re-prioritization in WM

While all putative mechanisms explain some of the retro-cueing literature, none can account for all of it. To summarize, benefits cannot be explained solely by transfer from fragile to robust memory. They do not require sustained attention on the cued item, and they are not accounted for solely by increased probe anticipation. However, a combination of these mechanisms could account for benefits in most paradigms. But is there a more parsimonious mechanism?

One possibility that has been raised in the context of the PFC / basal ganglia model of WM updating is that retro-cueing controls the

gating of one item in WM into an action-oriented buffer (output-gating, Chatham et al., 2014, Chatham and Badre, 2015). This action-oriented buffer could be equivalent to the focus of attention described in some models of working memory. For example, Oberauer and Hein (2012) describe the focus of attention as serving to 'select one item or chunk as the target of the next cognitive operation' (p. 164). The system involved in response preparation is then ready to enact the rule encoded by the cued item. Here it helps to consider the similarity between encoding the current task rules in WM and encoding visual objects in VSTM. Typically, we think of task rules as being stimulus-response mappings that are held online by activity in the lateral prefrontal cortex (Braver et al., 1999; Cohen et al., 2002). Typical VSTM representations act similarly, even when their maintenance is usually conceptualized as a sensory representation (Luck and Vogel, 1997; Pasternak and Greenlee, 2005; Ester et al., 2009). The same visual memorandum, given a task structure known to the participant, might be remembered as a new stimulus-response mapping, or rule: if the probe is a red square in the top left corner, respond 'match', otherwise respond 'mismatch'. In this framework, retro-cued working memory representations might act as matched filters (Anderson, 1973; Heath, 1982), laid down in a neural network that can process arbitrary input and give task-appropriate output by 'filtering' the input to calculate its similarity with the memorandum.

What behavioral hallmarks could provide evidence that this mechanism exists? Some testable predictions come to mind. Firstly,

output gating changes the representational status of the cued item. Therefore, it need not necessarily draw on the conventional pool of limited WM resources. As a consequence, memory for uncued items, retaining the same amount of WM resource as before the cue, would not diminish. However, uncued items *should* be less likely to occupy the special status in the output gate. Therefore, response times to cued items should be faster than those to neutral and uncued items. This is because, in the latter two cases, the output filter can only be set up (or reconfigured) after the onset of the probe. Secondly, *any* input should be processed according to the output filter. When an unexpected invalid probe appears, observers might then sustain a bias towards the retro-cued item, if responses are initiated before the filter can be completely reconfigured. Conversely, on valid trials, the setup of the appropriate stimulus-response filter before probe onset should reduce interference from uncued items. In addition, any other influences on behavior should also diminish, such as the long-run statistics of the environment, or the average of the ensemble currently held in mind (see also Discussion). These questions are further explored in chapter 2.

How does this translate into neural predictions? In particular, is it sensible to think of cued items as occupying a state that is distinguished by persistent activity (in the prefrontal cortex or, more likely, sensory cortex)? If that is the causal mechanism of improved memory, rather than an incidental correlate, then abolishing persistent activity should eliminate the retro-cue benefit. This has not been tested conclusively, but the

evidence (from imaging studies on multiple cueing, see LaRocque et al., 2014) suggests that uninterrupted delay activity is not a requirement for improved performance. One fMRI study in particular found evidence for reactivation of coding (Lewis-Peacock et al., 2012). The authors used multi-voxel pattern analysis to decode the contents of working memory in brain areas with elevated delay-period BOLD responses. Remembered categories could not be decoded in the initial delay. After a cue, however, information for the cued category emerged. Crucially, other items were not decodable in this interval, even though they were remembered by the participants. When a second cue redirected attention to a previously uncued item, decoding for that category emerged. This can be interpreted as the transient activation of WM contents that are otherwise laid down in a code that is invisible to fMRI pattern analysis.

If persistent activation is not necessary, how else could an active code improve behavior over and above a silent (output gating) mechanism? A possibility might be that firing increases the fidelity of a matched filter (David et al., 2008; Hayden and Gallant, 2013), so that it responds to the probe more accurately. Alternatively, elevated firing rates might be an epiphenomenal but revealing result of a temporary reconfiguration of the memory network: After a retro-cue, the network responsible for coding that item increases its susceptibility to new input by increasing the gain of neurons coding for the remembered item. Recurrent connections in the network could then lead to increases in firing rate (which are also more reflective of the cued item), even during

the delay. Additionally, such a mechanism could help explain the finding that prioritized items in WM (but not other items) capture attention automatically (Olivers et al., 2011).

Scope of the thesis

This thesis examines some behavioral and neural consequences of flexible attention shifts in working memory, and what this might reveal about the neural representation of information in WM. Chapter 2 revisits the issue of how retro-cueing benefits arise, arguing that cueing benefits can occur without costs to uncued items. Chapter 3 examines the temporal profile of oscillatory responses to retrospective and prospective cues in visual cortex. A role of alpha oscillations in selecting retro-cued and pre-cued items is identified. Following on, chapter 4 tests whether uncued, spontaneous variability in pre-stimulus visual excitability, as measured by alpha-band oscillations, is also predictive of the quality of encoding into WM. Chapter 4 also examines whether pre-stimulus excitability is indicative of successful inhibition of irrelevant visual distractors.

The second section uses pattern analysis methods to examine the representation of items in working memory. In chapter 5, I will describe results that are consistent with a silent representation of visual target templates in WM, which is reactivated as needed around the time of a decision. In chapter 6, brain networks related to flexibility inhibition or updating of items into WM are identified using fMRI. Chapter 6 also uses pattern decoding of MEG data to trace the consequences of updating on the representation of items in the neural response. Finally, chapter 7 discusses the findings from this thesis in a broader context. It attempts to

situate the WM systems examined here among a number of other forms of memory, all of which interact to optimize behavior.

Chapter 2: The Fate of Prioritized and De-prioritized Items in Working Memory

Chapter summary

Most recent models conceptualize working memory as a continuous neural resource that is divided up according to task demands. When an increasing number of items need to be remembered, each item receives a smaller chunk of the memory resource. These models predict that attentional benefits are a zero-sum game: improvements in remembering cued items come at the expense of uncued items because resources are dynamically transferred from an uncued to a cued representation. The current chapter provides empirical data challenging this model. Two precision visual WM experiments assessed memory for cued and for uncued items on every trial. This permitted a test for trade-off of the memory resource. While there were robust improvements in WM performance for an item cued as relevant during the retention interval, this benefit came at no cost to uncued items that were probed afterwards, increasing the net capacity of WM. In addition, when the order of recalling the cued and uncued items was unknown, recall of the uncued orientation was biased away from the orientation of the subsequently probed retrocued item – an example of retroactive interference. These results prompt a re-evaluation of some current proposals for the architecture of WM.

Introduction

Working-memory (WM) capacity limits (Luck and Vogel, 1997) can be at least partly overcome by retrospective cues presented in the delay period (Griffin and Nobre, 2003; Landman et al., 2003). These retrocues could confer benefits in several ways: by reallocating mnemonic resources to the cued item (Pertzov et al., 2013), by increasing expectation towards the likely location or features of the probe (Lepsien and Nobre, 2007; Lepsien et al., 2011; Kuo et al., 2012), or by placing the cued item in a privileged state such as the focus of attention (Griffin and Nobre, 2003; Rerko and Oberauer, 2013; LaRocque et al., 2014; Rerko et al., 2014; Zokaei et al., 2014). The latter proposal of a state change is similar to results suggesting that the cued item might be made more accessible to recall or to guiding behavior by occupying an output-driving state (Chatham et al., 2014, Chatham et al., 2015, Wallis et al., 2015). There is additional evidence that cues may also protect WM from interference (van Moorselaar et al., 2015; Makovski and Jiang, 2007; Hollingworth and Maxcey-Richard, 2013, Rerko et al., 2014; Zhang and Luck, 2008; but see Janczyk and Berryhill, 2014).

Previous studies have shown costs for invalid retrocues (Griffin and Nobre, 2003; Astle et al., 2012; Pertzov et al., 2013; Gözenman et al., 2014; but see Berryhill et al., 2012). Some models of working memory propose that costs arise because a single continuous resource needs to be shared among all remembered items (Bays and Husain, 2008; Bays, 2014; Ma et al., 2014; Bays, 2015; Matthey et al., 2015). Cueing would

therefore lead to a transfer of resources to the cued representation, so that uncued items are remembered less accurately or are forgotten more easily (Pertzov et al., 2013). However, other explanations of the cueing costs exist. For one, even if there is no redistribution of memory resources, anticipating the correct probe feature might yield benefits by improving the perceptual processing of the probe stimulus. Conversely, on invalid trials, anticipation of an incorrect feature could cause interference that leads to costs.

Beyond alternative explanations, not all experiments find significant costs for uncued items. For example, multiple-cueing paradigms have provided a different view on retrocueing (Landman et al., 2003; Rerko and Oberauer, 2013; van Moorselaar et al., 2015; see also Lepsien and Nobre, 2007). In these experiments, a first cue indicates a to-be-prioritized item, but can be superseded by a subsequent cue. The typical finding is that re-directing attention with a second cue still leads to significant performance improvements, sometimes (Rerko and Oberauer, 2013) on a par with the benefit from a single cue. This argues against a model of WM where, once resources have been withdrawn from an item, they can never be returned to improve that item's representational fidelity. Nonetheless, anticipation of the correct probe location or feature could (partially) explain performance improvements in multiple-cueing studies. Some important questions remain: do retrocues improve the representation of memory items themselves? If so, does this improvement come at a cost to uncued items?

In order to investigate what retrospective selection can reveal about WM architecture, we need to probe cued and uncued items, while controlling for probe anticipation. We probed how well cued and uncued WM items could be retrieved in two double-probe experiments. Observers encoded four oriented bars into WM, and on some trials one of the items was cued during the retention interval. Observers then recalled, in sequence, the orientation of two of the four items. On cue trials, the cued item was always probed, along with one of the other three items. By design, irrespective of whether a cued item or a neutral item was probed first, the conditional expectation of the second probe was identical (i.e. 1 in 3 of the remaining, unprobed items). Therefore, responses to uncued items on the second probe, when compared to neutral trials that did not contain a cue, would reveal any costs that are independent of probe expectation or response preparation. In brief, there was no evidence for behavioral costs, in spite of robust benefits in remembering the cued item.

While this demonstrated that cueing benefits can occur without costs when the uncued item is recalled last, there was still a performance decrease when the uncued item was recalled first, before the cued item. In addition, the recall of this uncued item was biased away from the cued item. A second experiment tested whether advance knowledge about the order of recall (recall cued item first, or recall an uncued item first) could reduce or eliminate these costs. Knowing the recall order did not entirely eliminate costs during the first response. However, it did eliminate the

biasing effect of the cued item. The set of results may be explained by retrocues changing the functional state of the cued item so that it carries a heavier weight in driving behavior.

Experiment 1

Materials and Methods

Participants

The experimental procedures were approved by the Central University Research Ethics Committee of the University of Oxford. Twenty observers (eleven females, mean age 24.5 years, range 18-35 years) performed a sequential visual working-memory task. All had normal or corrected-to-normal vision. Participants provided written consent in accordance with the University ethics guidelines, and were compensated at a rate of £10 per hour. Data from one participant were excluded because of poor performance (no evidence for memory-guided behavior, tested against Rayleigh distribution, $p > 0.01$).

Apparatus

The task was created using the Psychophysics Toolbox (Brainard, 1997) and custom scripts written in Matlab (The Mathworks). Participants sat in a dimly lit, sound-attenuated booth, at a distance of 74 cm from the monitor (22-inch Samsung Syncmaster, 1680*1050 pixels, 60 Hz refresh rate, 47 cm screen width). Binocular gaze locations were monitored online using an infrared video-based eyetracker (Eyelink 1000, SR Research) sampling at 500 Hz. A head rest was used to reduce head motion and improve the accuracy of the eyetracker. Participants were instructed to maintain fixation throughout the trial, except during the response and inter-trial periods. Responses were made using a mouse.

Task

On each trial, observers had to remember the orientations of four peripherally presented bars (see Fig. 1). Two of these were then probed in sequence. On no-cue trials, these items were randomly selected. On retrocue trials, one of the two probed items was cued during the delay. Orientation memory was probed by means of continuous recall estimates. The experiment involved three blocks of practice and eight blocks of the main task. Each block consisted of 60 trials (20 neutral, 20 cued item probed first, 20 cued item probed last), resulting in a total task length of 480 trials.

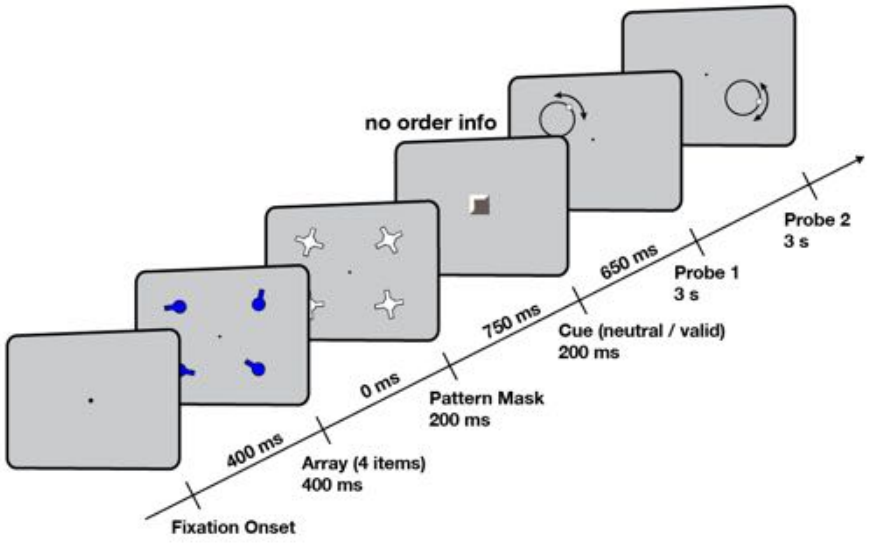


Figure 1. Trial Structure for Double-Probe Experiment 1. Each trial began with the appearance of a central fixation dot. After 400 ms, an array of 4 oriented bars appeared for 400 ms, followed immediately by pattern masks for 200 ms. Observers were instructed to remember the orientations of all stimuli. A cue appeared 750 ms later (for 200 ms). On

retrocue trials (2/3 of all trials), the cue indicated one of the two items that would be probed (by pointing to the location where that item had been presented). Uninformative (neutral) cues appeared in the remaining 1/3 of trials. After a further delay (650 ms) the first probe appeared (a black outline in the location of one of the presented items), prompting observers to recall that item's orientation. Observers had 3 s in which to respond to the first probe. After the full 3 s, a second item was probed, with the same time constraint. After responding about both items, observers saw the correct orientations superimposed over their response orientations for the two probed items (not shown).

On each trial, participants saw a stimulus array (presented for 400ms) consisting of four oriented bars around a central fixation dot (a black square with 0.20° visual angle edge length, on a grey background). Oriented bars (2.60°x0.30°, with a disc of 0.70° diameter at one end to denote the axis of rotation) appeared at the center of each quadrant, at a radial eccentricity of 5° from fixation. Stimuli were immediately masked (for 200 ms) to minimize the contribution of iconic memory on performance. Masks consisted of randomly oriented white crosses (created by overlapping two oriented bars at 90° from one another). After a delay of 750 ms, a central cue appeared for 200 ms (0.40° edge length). Neutral cues (1/3 of trials) were black squares. Retrocues (2/3 of trials) contained a colored corner (blue), indicating which of the four items would be probed. After a 650-ms delay, a probe appeared at a previous

stimulus location. Probes consisted of a black circle with a radius equal to the length of the bar. After participants started moving the mouse to respond to the orientation, a small tracking disc appeared at a random location on the circle. Participants were instructed to move the mouse to rotate the tracking disc to the remembered orientation of the bar at the probed location. They locked in the current orientation with a mouse click. Observers had a 3,000-ms window in which to respond. The first probe remained on the screen for 3,000 ms, irrespective of response time, to avoid any speed-accuracy trade-off between responses to successive items and to ensure that the memory delay for the second item did not depend on response time to the first.

Analysis

For each response, we calculated the deviation (in degrees) of the recalled orientation from the actual orientation of the probed stimulus, generating a distribution across trials. Recall accuracy was defined as the inverse of the circular standard deviation of the response distribution (Bays et al., 2009), separately for each condition. Repeated-measures analysis of variance (ANOVA) with the factors cue (neutral, cued, and uncued) and response (first, second) was used to compare accuracy between conditions. Benefits and costs were also established separately by comparing only cued and neutral or uncued and neutral trials. This was a crucial analysis to establish, especially for the second response, whether uncued items carried a cost.

We also fit a mixture of two distributions to observers' responses:

$$p(x|\theta) = P_{MEM} * VonMises(x,0,\kappa) + (1 - P_{MEM})/2\pi, \text{ (Equation 1)}$$

where x is the response angle, θ is the presented angle, *VonMises* denotes the Von Mises distribution, i.e. the circular analogue of the normal distribution, and $(1-P_{MEM})/2\pi$ is the height of the uniform distribution arising form randomly distributed guesses. The two variable parameters, P_{MEM} and κ, define the two distributions: the probability of recalling an item (P_{MEM}), and the concentration parameter, or precision, of remembered items (κ). This model was fit using maximum-likelihood estimation (MLE, Myung, 2003), separately for each condition. Model parameters for different conditions were compared using ANOVA, as above. Previous studies have found that retrocues affect mainly the likelihood of recall (P_{MEM}), having little or no effect on the precision (κ) of memory representations (Murray et al., 2013; Wallis et al., 2015; but see Williams et al., 2013). We therefore focused on how cues changed the recall rate.

Because the results revealed a number of non-significant effects, we used Bayes Factors (Dienes, 2011; Kass and Raftery, 2012) to establish whether the lack of significance indicated a genuine absence of difference, or merely a lack of sensitivity. For each effect of interest, we compared two alternative models: the null model assumed a difference of 0, and fit only the standard deviation across participants (using MLE). The

alternative model fit both mean and standard deviation to the distribution across participants. The Bayesian Information Criterion (BIC) was used to compute the fit of each model (accounting for the difference in degrees of freedom) and to compute the corresponding Bayes Factor (BF):

$$BF = \exp([BIC_1-BIC_0] /2), \text{ (Equation 2)}$$

where BIC_1 is the BIC for the alternative model and BIC_0 is the BIC for the null model. Bayes factors above 3 generally are interpreted as substantial evidence in favor of the alternative model, while values smaller than 1/3 favor the null model (i.e., an absence of difference). Following this general guideline, Bayes factors below 1/3 were interpreted as genuine null effects.

Experiment 1 Results

Accuracy

There were robust cueing effects on accuracy (Fig. 2a, 2x3 ANOVA, main effect of cue: $F_{2,36}$=32.16, p<0.001), as well as a significant effect of response order ($F_{2,36}$=19.31, p<0.001). In addition, there was a significant interaction ($F_{2,36}$=5.20, p=0.010). The main effect of cueing could arise from cueing benefits (i.e., cued items were remembered better than neutral), cueing costs (uncued items worse than neutral), or both. Therefore, it was crucial to examine these effects separately via 2x2 ANOVAs that included either the cued or uncued trials (together with the neutral condition). Excluding uncued trials still resulted in an effect of cueing (2x2 ANOVA with cued and neutral responses, main effect of cue: $F_{1,18}$=21.27, p<0.001). There was also a main effect of order ($F_{1,18}$=15.03, p=0.001). However, here there was no interaction with the cue effect ($F_{1,18}$=2.83, p=0.110), indicating that cueing benefits were comparable on the first (t_{18}=2.97, p=0.008, BF=18.745) and second responses (t_{18}=5.14, p=0.0001, BF=$1.22*10^6$). In contrast, there was a strong interaction when comparing neutral to uncued items (both main effects: $F_{1,18}$>35, p<0.001, order by cue interaction $F_{1,18}$=11.76, p=0.003), owing to the strong cueing cost during the first response (t_{18}=-5.20, p=0.0001, BF=$1.69*10^6$) in the absence of a significant cueing cost during the second response (t_{18}=-0.18, p=0.86, BF=0.23). This pattern of results confirmed our hypothesis: when an uncued item is probed first, and is therefore unexpected, accuracy is worse than on neutral trials. However, after the cued item has

been recalled, accuracy for the uncued item is no worse than on neutral trials.

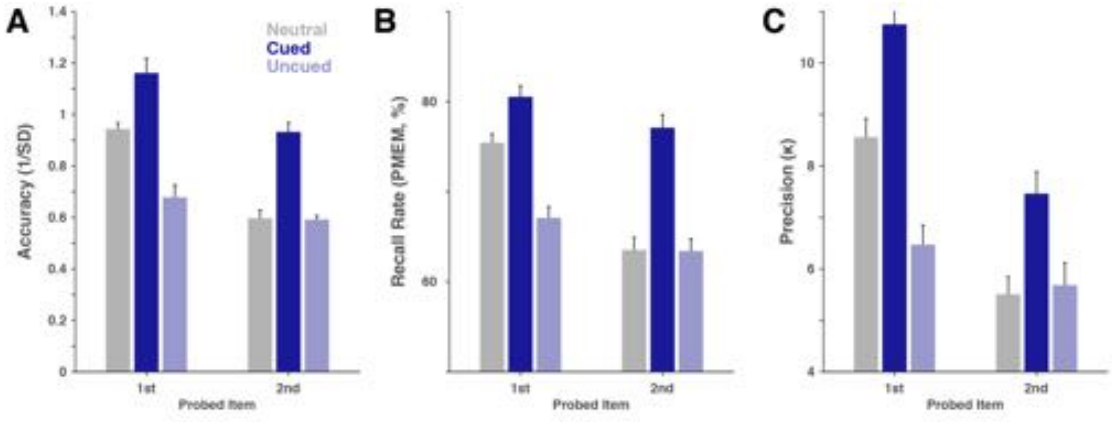


Figure 2. Behavioral Performance in Experiment 1. A. Precision was modulated both by cue type and by probe order. Importantly, during the second response, uncued items had equivalent accuracy to neutral items, indicating that there are no lingering costs to improving the accuracy of a retrocued item. Error bars denote within-observer standard error of the mean. B. Recall rate for the same conditions. C. Precision parameter estimates from the same model fit.

This pattern of results was also reflected in the recall rates, as estimated from the mixture model. Again, there were main effects of cueing (Fig. 2b, $F_{2,36}$=34.63, p<0.001) and of response order ($F_{1,18}$=25.63, p<0.001), with a significant interaction ($F_{2,36}$=5.37, p=0.009). Cues improved recall rates for both responses (main effect in 2x2 ANOVA: $F_{1,18}$=23.91, p<0.001, cueing benefits during the first response: t_{18}=2.81, p=0.012, BF=12.05, second response: t_{18}=5.43, p<0.0001, BF=5.82*10^5).

Although recall rates, in general, decreased for the second response (main effect of order, $F_{1,18}$=52.46, p<0.001), the cueing benefit increased (interaction with probe order: $F_{1,18}$=15.92, p=0.001).

In addition to the cueing benefits, there was also a significant reduction in recall rate when uncued items were probed (main effect of cue, $F_{1,18}$=10.92, p=0.004, main effect of order, $F_{1,18}$=22.55, p<0.001). This was caused by a significant reduction in recall rate during the first response (t_{18}=-4.39, p=0.0004, BF=3.51*10^4). By contrast, there was no reduction in recall rate when uncued items were probed second (t_{18}=-0.05, p=0.96, BF=0.23).

Although our main hypothesis focused on benefits or costs in the recall rate, similar effects were also visible in the precision parameter of the mixture model (Fig. 2b, main effect of cue: $F_{2,36}$=15.28, p<0.001, main effect of order: $F_{1,18}$=23.78, p<0.001, interaction: $F_{2,36}$=5.23, p=0.010). As with recall rate, cued and uncued items were significantly different from neutral items (all t>3, p<0.011, BF>14), with the exception of uncued items probed during the second response (t_{18}=0.27, p=0.79, BF=0.24).

The main results of experiment 1 confirmed that there is a trade-off between cued and uncued accuracy only during the first response. When a retrocued item has already been recalled, the uncued items are not remembered worse than neutral items. Therefore, cueing benefits do not arise from a permanent redistribution of a memory resource.

Trial-by-trial and between-observer correlations

While uncued items carried no overall costs (at least after the cued item had already been recalled), it is possible that there was a more subtle trade-off in accuracy on a trial-by-trial basis, such that trials in which the cued item was remembered better than average led to reduced accuracy for uncued items. The absolute angular error for the first and second response on each trial was used to calculate the Pearson correlation coefficient across trials, separately for each condition (neutral, cued item probed first, and cued item probed last). If there was any trade-off, there should be a negative correlation between responses in the cueing conditions. Averaged across all conditions, there was a small but significant positive correlation (mean Pearson r = 0.033±0.013, t_{18}=2.66, p=0.016) of accuracy for responses 1 and 2 across trials. Nonetheless, between-trial correlations were weak, with no consistent relationship on neutral trials (r = 0.009±0.021, t_{18}<1). There was a positive correlation when the cued item was probed first (r = 0.067±0.026, t_{18}=2.58, p=0.019), although there was also no significant correlation when the cued item was probed second (r = 0.024±0.017, t_{18}=1.40, p=0.18). Therefore, there was no evidence that resource redistribution might lead to an accuracy tradeoff across trials.

While there was no evidence of a trade-off across trials (in any of the conditions), there could have been a trade-off between benefits and costs across participants. When testing for correlations across observers, a significant positive correlation emerged between the cueing benefit on responses 1 and 2 (Pearson r_{18}=0.49, p=0.031). In addition, there was

also a correlation between cue benefits and cue costs when cued items were probed second (r_{18}=-0.59, p=0.0075). This could indicate strategy differences, where some participants focused more exclusively on accurate recall of the cued item. However, the reverse correlation appeared when cued items were probed first (r_{18}=+0.545, p=0.016). In sum, there was no consistent pattern of trade-offs. Possibly, cueing benefits came at the cost of other items only when the cued item was recalled last, and therefore had to be kept in memory during recall of an uncued item.

Experiment 2

A second experiment attempted to replicate the findings of Experiment 1 and to assess whether knowledge of probe order had an influence on the benefits and costs of cueing (Fig. 3).

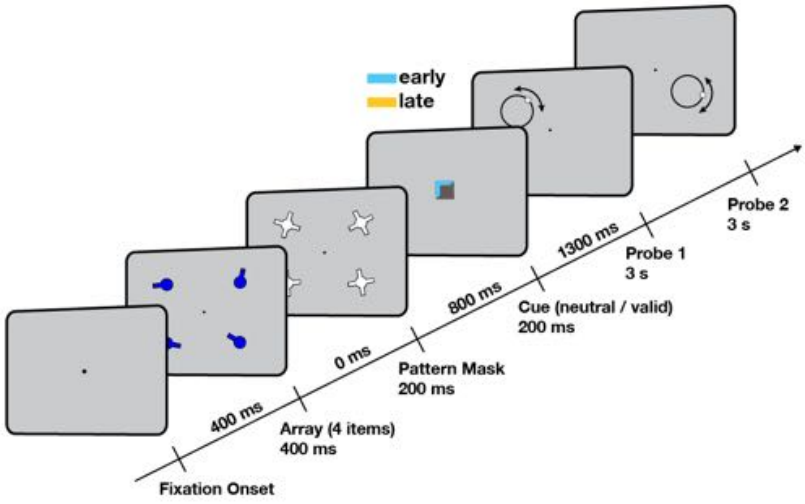


Figure 3. Trial Structure in Experiment 2. The structure was identical to Experiment 1, with the exception that the cue color indicated whether that item would be probed first or second. The delay between the cue and the first probe was longer (1,300 ms) to give observers more time to parse the more complicated cue.

Materials and Methods

Participants

Twenty new observers (nine females, mean age 24.15 years, range 19-31 years) participated in Experiment 2. All had normal or corrected-to-normal vision. Participants provided written consent in accordance with

the University ethics guidelines, and were compensated at a rate of £10 per hour. One participant was excluded because of poor performance.

Task

The task was identical to Experiment 1, with the exception that retrocues also gave information about probe order. On each cue trial, the color of the cue indicated whether the cued item would be probed first or second. Cue colors (orange and blue) were equiluminant, and were counterbalanced across participants. Participants were instructed about the cue-order mapping before the experiment started. Because of the increased demands in parsing the cue, the delay between cue offset and onset of the first probe was increased to 1,300 ms. Other timings were kept the same (with the exception of a minor increase in the delay between array offset and cue onset from 750 to 800 ms). Analyses were identical to those in Experiment 1.

Recall Bias Analysis

Experiment 2 was designed to test whether cueing costs were still present with prior knowledge of the probe order. However, in addition to influencing recall accuracy, retrocues could also have more subtle effects on the mutual interactions between items represented in WM. For instance, recall of uncued items might be attracted toward or repulsed from the cued feature. This was tested by looking for biases arising from the cued item on the recall of the uncued item. For the within-trial

analysis, the first response from each trial was binned (into 64 bins, each containing 1/4 of all trials, see also Cravo et al., 2013) according to the relative orientation of the second probed item with respect to the first probed item. The average response bias was calculated for each bin, yielding a bias curve as a function of relative orientation. If there was a consistent attraction effect, then response bias should be largely negative for negative relative orientations, and positive for positive relative orientations. A repulsive (orthogonalizing) effect would create the opposite pattern: a positive bias for negative relative orientations, and vice versa. To quantify whether there was any influence, the area under the bias curve was integrated, separately for negative and positive angular differences. If the average area for positive relative orientations was smaller than for negative orientations, this indicated a *repulsive* effect of the unprobed stimulus, whereas a positive area difference would indicate an *attractive* bias. Significant bias was assessed using a one-sample t-test on the area under the curve. This analysis was repeated for the second response (with respect to the item probed during the first response), separately for each cue condition.

While this experiment was mainly aimed at the consequences of retrocues on the current trial, cueing additionally permitted addressing the separate question of how inter-trial biases arise. As outlined in the introductory chapter, WM recall is generally biased towards features encountered on the previous trial (or trials). If this bias occurs during encoding, then WM representations should be shifted from the beginning

of the delay, and retrocues should have no influence on the magnitude of the bias. However, if the bias emerges (at least partially) during a later stage in the trial, such as during recall (as could be predicted by a Bayesian account of WM recall that combines current and prior information), then retrocues could have an influence: By increasing the reliability of information from the current trial, the influence of previous trials should decline. To address this question, the bias analysis was repeated, but both items probed on the current trial were sorted with respect to the orientations probed on the preceding trial (separately for both responses). Otherwise, bias estimation was identical to the within-trial analysis.

Experiment 2 Results

Accuracy

Accuracy of recall (as measured by the inverse of the standard deviation) was calculated for all three cue conditions (cued, neutral, and uncued), and for both responses (to probe 1 and 2). Retrocues had a significant influence on accuracy (Fig. 4, main effect across all three cue types, $F_{2,36}$=30.92, p<0.001). Accuracy also declined for the second response on each trial (main effect of probe order, $F_{1,18}$=68.13, p<0.001). There was a significant interaction between cue type and probe order ($F_{2,36}$=4.44, p=0.019).

As with Experiment 1, the effect of cue type could have arisen from improved recall of the cued item or impaired recall of the uncued item. Importantly, when examining only neutral and cued responses, there was still a significant main effect of cue type ($F_{1,18}$=33.53, p<0.001) and of order ($F_{1,18}$=49.33, p<0.001), with a small but significant interaction ($F_{1,18}$=4.49, p=0.048). The interaction indicated that the magnitude of the retrocueing benefit was slightly higher when the cued item was probed first (t_{18}=5.86, p=1.5*10^{-5}, BF=6.53*10^{6}), compared to when it was probed second (t_{18}=4.62, p=2.1*10^{-4}, BF=9.94*10^{4}).

The same was not true for the costs to uncued items. When comparing uncued to neutral responses, there was still a significant effect of order ($F_{1,18}$=41.08, p<0.001), but only a trend towards an effect of cue ($F_{1,18}$=3.48, p=0.079), with no interaction between order and cue type ($F_{1,18}$=2.07, p=0.168). While keeping in mind the difficulty of interpreting

post-hoc effects in the absence of a significant interaction, the marginal costs to uncued items, if there were any, arose only during the first response (t_{18}=-1.88, p=0.076, BF=1.35), while on the second response neutral and uncued trials were indistinguishable (t_{18}=-0.17, p=0.87, BF=0.23). This null effect again confirmed the result predicted in the introduction. However, given the trend for a cost to uncued items during the first response, it cannot be ruled out that cues decreased accuracy of recall, even when observers know the order of recall.

In general, this set of findings is again consistent with the hypothesis that retrocue benefits should always occur, while retrocue costs (to the uncued items in the array) occur primarily when the cued item is anticipated (that is, during the first response), but not when the cued item is no longer behaviorally relevant (during the second response). More broadly, the detrimental effects of having anticipated the cued item when the uncued item is probed could reflect incorrect preparation for memory-guided action. A closer look at the trials could also reveal behavioral biases induced by keeping the cued item in an action-oriented state.

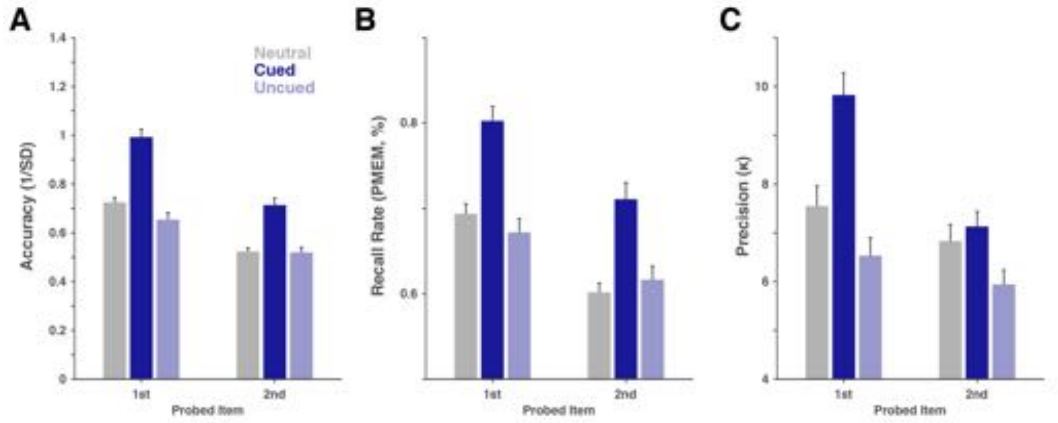


Figure 4. Behavioral performance in Experiment 2. A. Recall accuracy (1/SD of the response distribution) was modulated both by cue type and by probe order. Importantly, during the second response, uncued items had equivalent accuracy to neutral items, indicating that there are no lingering costs to improving the accuracy of a retrocued item. Error bars denote within-observer standard error of the mean. B. Recall rate parameter (from the mixture model fit) for the same conditions. C. Precision parameter estimates from the same model fit.

Mixture-Modeling Results

The next question was whether the behavioral dissociation between the effects on cued and uncued items was also reflected in recall rate, as assessed with a mixture model. Again there was a significant main effect of cue type ($F_{2,36}$=16.78, p<0.001) and of order ($F_{1,18}$=77.10, p<0.001), with no significant interaction when looking only at neutral and cued items ($F_{1,18}$<1, p=0.99). Retrocues increased the recall rate on both the first (t_{18}=4.46, p=0.0003, BF=4.80*10^4) and second response (t_{18}=4.04, p=0.0008, BF=806). Conversely, uncued items did not

lead to significant reductions in recall rate (main effect of cue: $F_{1,18}$=0.05, p=0.83, first response: t_{18}=-0.97, p=0.35, BF=0.37; second response: t_{18}=0.76, p=0.46, BF=0.31). Again, this analysis confirmed an asymmetric pattern of significant benefits at no cost. Additionally, it appears that recall rates for uncued items did not suffer for either response when cues indicated the recall order.

One possibility for the asymmetric pattern in the recall rate could be that cueing benefits impacted mainly the recall rate, while costs to uncued items appeared in a reduction in their precision (estimated from the von Mises distribution fit). As before, there were main effects of cueing ($F_{2,36}$=11.46, p<0.001) and of recall order ($F_{1,18}$=14.72, p=0.001), with a significant interaction ($F_{2,36}$=3.80, p=0.032). However, this pattern broke down into a significant increase in fidelity for the cued item on the first response (t_{18}=2.88, p=0.0099, BF=14.69), with no strong effects on fidelity for the cued items on the second response (t_{18}=0.52, p=0.611, BF=0.26), and no significant costs to uncued items (first response: t_{18}=-1.61, p=0.124, BF=0.84, second response: t_{18}=-1.60, p=0.126, BF=0.83).

Trial-by-trial and between-observer correlations

As with Experiment 1, there were modest but significantly *positive* correlations for all three cueing conditions (mean±s.e.m. for neutral: r=0.055±0.025, t_{18}=2.18, p=0.043, cued first: 0.073±0.019, t_{18}=3.94, p=0.001, cued second: 0.062±0.021, t_{18}=2.92, p=0.0091). Correlations did not diminish in magnitude on cued trials compared to neutral trials (cued

first: t_{18}=0.73, p=0.47, BF=0.30, cued second: t_{18}=0.22, p=0.83, BF=0.23). This again contradicts the hypothesis of any trade-off measured across trials.

Similarly, there were no significant correlations across observers between cueing benefits (accuracy on cued minus neutral trials) and costs (uncued minus neutral, all pairwise $p>0.45$, all $r<0.19$). The only significant correlation was a positive relationship between cueing benefits on the first and the second responses (Pearson r_{19}=0.66, p=0.0023).

Behavioral orthogonalization of recalled items

In spite of the absence of a strong trade-off in accuracy, there might be more subtle effects of the cued item on the recall of uncued items. As outlined above, the accuracy improvement from cues in the current design is hard to explain via increased probe anticipation or an increased share of the memory resource. The benefit could arise because the representation of the cued item has changed (see Griffin and Nobre, 2003): After the cued item has been selected, it is stored in a state that can immediately drive behavior. As a result, unexpected probes for a *different* item could result in responses that are biased by the orientation of the cued item, since both items temporarily compete to drive behavior.

Returning to experiment 1, this was tested by binning trials by the angular difference between the orientation of the first and the second probed item. In each bin, the response bias was calculated. On trials when an uncued item was probed first (and, consequently, the cued item had to be held in mind for the second response), responses to the uncued item were significantly biased *away* from the orientation of the cued item (Figure 5, t_{18}=-4.28, p=4.47*10^{-4}, BF=2.21*10^{4}). In other words, when the cued orientation was clockwise with respect to the probed orientation, responses to the probed orientation were biased in the counterclockwise direction (and vice versa). There was no bias on neutral trials or on trials when the cued item was probed first (all p>0.52, BF<0.28, resulting in a main effect of condition, $F_{2,36}$=16.32, p<0.001).

Therefore, bias during the first response appeared only when the cue predicted the second probe.

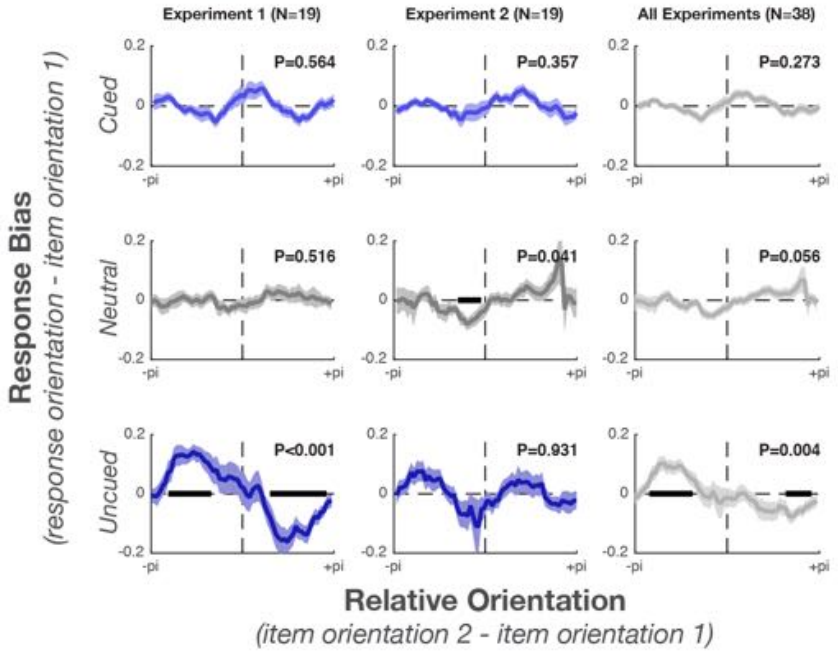


Figure 5. Response bias during serial probe recall. Response bias during the first of two responses, with respect to the orientation of the item that was probed second. Data are shown separately for all cueing conditions and both experiments. Each column shows bias curves for a different experiment (with the last column showing the aggregate across both experiments). Each row shows a different cueing condition, with the condition label (Cued, Neutral, Uncued) referring to the item that is being recalled. Each plot shows the average response bias (y-axis: negative values indicate that responses were, on average, clockwise to the correct orientation of the probed item) with respect to the angular difference between the two probed orientations (x-axis: negative values indicate that the other orientation was clockwise to the currently probed orientation).

*Shading indicates s.e.m., and black bars indicate points on the curve that showed significant bias (p<0.05, uncorrected). P-values in each plot denote whether the curve as a whole was significantly biased. The first response was generally not biased towards or away from the second item, with one exception. In Experiment 1, when the retrocued item was probed last, it exerted a significant repulsive bias on the response to the uncued item (bottom row, middle column, p=4.47*10^{-4}).*

The biasing effect of the subsequently probed, retrocued item, in experiment 1 could have appeared because observers were not cued about when the cued item would be probed. Therefore, observers might prepare to recall the cued item by default. The biasing step might have resulted from an unexpected change in recall. In experiment 2, the cue color indicated probe order: this might have allowed observers to compartmentalize the recall process, and prevent the retrocued item from driving behavior to the uncued item. In line with this prediction, there was no biasing effect in experiment 2. Here, reporting the uncued item first was unaffected by the orientation of the retrocued item (t_{18}=-0.09, p=0.93, BF=0.23). There was also no consistent biasing effect across all conditions (mean across conditions: t_{18}=1.76, p=0.094, BF=1.09, main effect of condition: $F_{2,36}$=1.31, p=0.28), although there was a marginal attraction effect for neutral trials (t_{18}=2.20, p=0.04, BF=2.58). When pooling over both experiments (right column in Figure 4), the only

significant biasing effect was again for uncued items (t_{36}=3.04, p=0.004, BF=16.3, main effect of cue: $F_{2,74}$=10.89, p<0.001).

The previous set of results established that cued items that have not yet been recalled can have a repulsive effect on the recall of other items. Importantly, cueing the order of recall abolishes this repulsion. Therefore, observers in Experiment 1 were likely preparing to recall the cued item at the first response on each trial, and this preparation induced the bias. Repulsion could arise at several processing stages. One possibility is that the effect occurs in a motor planning circuit, rather than at an earlier, perceptual stage. One reason to believe this is that the orientation of an uncued or neutral item seems not to have any repulsive effect, even though it is still held in memory. Therefore, the recall of the uncued item (rather than the representation of its visual features) could be biased away from a concurrently formed response plan (to the cued item). A similar effect has been demonstrated recently in the serial recall of items from visual WM (Kang and Choi, 2015). In that study, the second response on a trial was biased away from the item recalled first. This finding was interpreted as a readout-induced orthogonalization of representations. A similar phenomenon (proactive orthogonalization) might be occurring here. However, a more direct comparison to the study by Kang and colleagues comes from examining the *second* response on each trial. Here, recall should be biased away from the item probed first, irrespective of condition.

In both experiments, the second response was indeed consistently biased away from the item recalled first (Fig. 6). In Experiment 1, recall was biased away (average across conditions, t_{18}=-3.77, p=0.0014, BF=276), with no main effect of condition ($F_{2,36}$=1.90, p=0.16). Nonetheless, the biasing effect was statistically significant only in the cue conditions (cued item probed second: t_{18}=-3.11, p=0.006, BF=29.1, cued item probed first: t_{18}=-3.65, p=0.0018, BF=181, neutral: t_{18}=-1.21, p=0.24, BF=0.48). Likewise, in Experiment 2 the responses to the second probe were also biased away from the first probed item (average across conditions, t_{18}=-2.83, p=0.011, BF=12.5), with no main effect of condition ($F_{2,36}$=0.087, p=0.92). While the biasing effect was not significant in all cueing conditions individually in Experiment 2 (cued item probed second: t_{18}=-3.18, p=0.005, BF=36.5, neutral: t_{18}=-2.07, p=0.054, BF=1.94, cued item probed first: t_{18}=-1.69, p=0.11, BF=0.96), when pooled over experiments all conditions showed a significant repulsive bias (cued item probed second: t_{36}=-4.51, p=6.25*10^{-5}, BF=4325, neutral: t_{36}=-2.36, p=0.024, BF=2.61, cued item probed first: t_{36}=-3.48, p=0.001, BF=69.8).

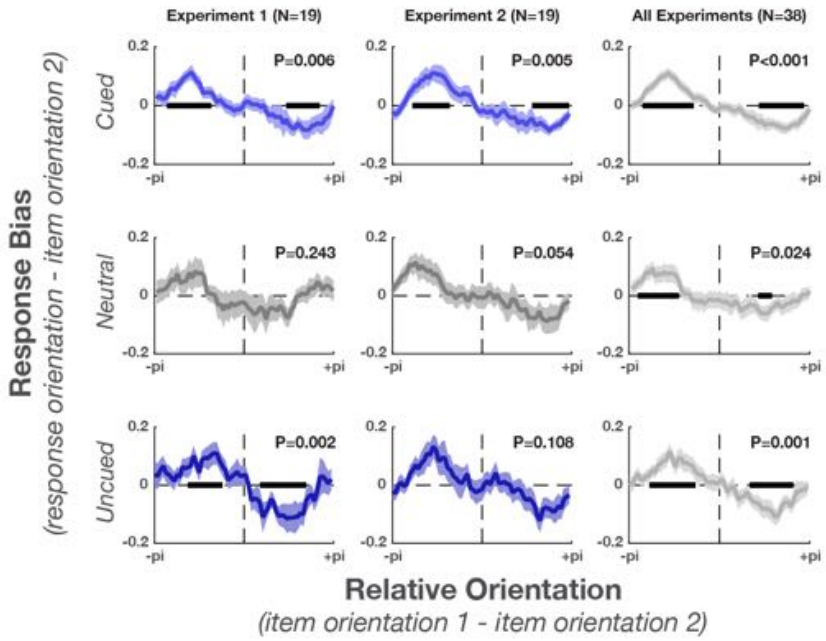


Figure 6. Response bias during the second of two responses, with respect to the orientation of the item that was probed first. Conventions are the same as in Fig. 5. Generally, responses were biased away from the orientation of the first probed item.

Retrocues reduce serial bias across trials

The previous section examined interactions between items remembered on the same trial. As outlined in the introduction, recent studies have additionally found that items probed on preceding trials can bias recall from WM (Huang and Sekuler, 2010; Fischer and Whitney, 2014; Liberman et al., 2014; Papadimitriou et al., 2015). The main effect seems to be that recall of visual features is biased towards the feature values of items on preceding trials. Some authors have argued that this bias is perceptual (Fischer and Whitney, 2014). Therefore, bias should originate during encoding, and information presented during the

maintenance delay should not reduce it. An alternative, Bayesian account argues that items from past trials act as priors whose influence on behavior is larger when the representation of the current item is noisy (Huttenlocher et al., 2000; Sims et al., 2012). Under this explanation, retrocues could reduce the reliance on prior information by increasing the amount of information about the current item.

In order to test for the effect of prior items, the bias curve analysis was repeated, but trials were sorted by the relative orientation of both items probed on the previous trial (Fig. 7a). Given that we expected no major differences between experiments for this analysis, data from all observers were collapsed. For each cueing condition (cued, neutral, and uncued), this resulted in four response bias curves (first or second response, sorted relative to the first or second probed item from the previous trial). Responses were significantly attracted toward the orientation of stimuli in the previous trial (averaging across the four bias curves) in all three conditions (Figure 7b; cued: area under the curve=12.1±2.1°, t_{37}=5.63, p=1.99*10^{-6}, BF=1.2*10^{6}, neutral: 16.4±2.2°, t_{37}=7.34, p=9.98*10^{-9}, BF=8.21*10^{10}, uncued: 21.1±2.6°, t_{37}=7.96, p=1.57*10^{-9}, BF=8.98*10^{12}). Crucially, the cueing condition had a significant effect on the magnitude of the bias ($F_{2,74}$=4.70, p=0.012), since bias was significantly smaller on cued trials than on uncued trials (t_{37}=3.21, p=0.0027, BF=28.1). There was also a positive linear trend from the cued to the neutral to the uncued condition (t_{37}=4.15, p=1.87*10^{-4}). Pairwise comparisons with the neutral condition yielded only trends (cued

vs neutral: t_{37}=-1.45, p=0.16, BF=0.46, uncued vs neutral: t_{37}=1.55, p=0.13, BF=0.54). Therefore, the effect of previous trial history can be (at least partially) modulated during the maintenance phase.

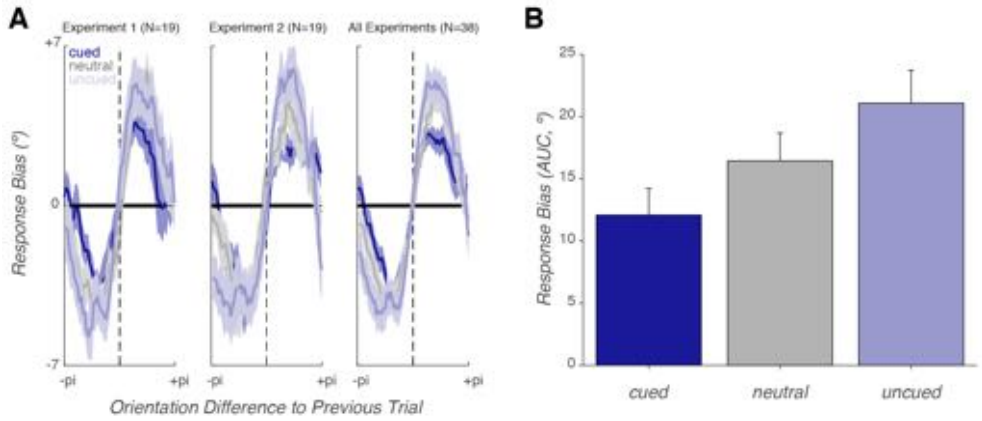


Figure 7. Retrocues reduce influence of the previous trial on recall. A. Response-bias curves, plotted separately for the three cue types. The first and second panels show response bias separately for experiments 1 and 2, and the third panel shows response bias for all participants. Error bars denote within-subject s.e.m. B. The area under the bias curve, indicating the magnitude of attractive bias, is dependent on cue condition. Error bars show s.e.m.

Discussion

This study examined how cueing attention towards an item in working memory affects recall of the cued item as well as recall of the uncued items. Specifically, the study set out to test whether memory for a cued item improves because it receives a larger portion of a shared memory resource, and whether uncued items lose resources as a consequence. Observers recalled two out of four items, one of which was sometimes cued during the memory delay. The cueing benefit persisted even when a second, uncued item was recalled first. More importantly, once a cued item had been reported, other items were recalled with the same accuracy as on neutral trials, yielding a net increase in working-memory performance. This finding seems inconsistent with a single, finite memory resource that is distributed among all items, because a finite resource account predicts a trade-off in memory between cued and uncued items. After cues, the recall of uncued items was biased away from the cued representation (if the order of recall was not known), and the recall of cued items was less biased by trial history. These biasing effects, together with the lack of a straightforward resource tradeoff, indicate that retrocues may change the representational format of the cued WM item so that it is more apt to drive recall.

To our knowledge, only one other study has tested memory for other items by probing more than one item in the recall phase (Rerko et al., 2014). The study found, in two experiments, significant costs to recalling an uncued item *after* having recalled a cued item. Two

differences in task design could account for the discrepant findings. First, Rerko and colleagues used six items per array, exceeding the average storage capacity of WM, whereas we used four items. When operating above the capacity limit, the difficulty of continued maintenance of all five uncued items might increase the likelihood of dropping them. Also, the difference in expected value between maintaining a cued item and maintaining any other item is larger for larger arrays, so strategic differences might have led observers to accept performance costs for larger arrays. Finally, in Rerko and colleagues' paradigm, mismatch probes were sometimes items from an unprobed location in the array. Possibly, the increased load could have led to swap errors on uncued trials due to a reduction in spatial resolution of feature-location bindings (Rerko et al., 2014). Given the significant costs to uncued items, Rerko and colleagues put forth two explanations for retrocueing benefits: firstly, uncued items could be (partially) forgotten, reducing interference with the cued item and leading to benefits and costs (see also Souza et al., 2014). Alternatively, benefits could be the result of moving the cued item into the focus of attention, a functional state that singles out one item in WM as the input to the next (cognitive) action (Rerko et al., 2014). The current set of experiments shows that it is possible to see benefits without costs, making the latter explanation likelier, at least for the paradigm used here.

In contrast to single-probe tasks, prompting for multiple responses can test for residual memory of uncued items while equating probe anticipation. After the first response has been made, whether that first

item probed a neutral item or a cued item, the second response probed one of the three remaining (uncued) items. Therefore, probe expectation in both conditions was equal. Furthermore, the probe was simply a circular outline presented in the location of one of the WM items. A small tracking disc indicating the response orientation appeared only once the response had begun. This procedure was designed to minimize the interfering effect of the probe stimulus, possibly further reducing any costs to behavior. Given that, under these circumstances, costs disappeared, benefits without costs presumably arose from a change in the representational format of the cued item that was independent of the maintenance of other items. As a consequence, there is likely more than one resource pool for the maintenance of information in WM. Similar accounts of the benefits of attention in the perceptual domain have been made in the past (Kinchla et al., 1995).

Overlapping representations, perhaps occurring in the focus of attention, might have been the cause of the biasing effect observed in Experiment 1. When observers did not know when the cued item would be probed, and an uncued item was probed first, then recall was biased away from the cued orientation. Interestingly, knowing the order of recall (in Experiment 2) seemed to eliminate the bias. The repulsion effect in Experiment 1 is not the result of mistakenly reporting the cued item. If the cued feature had been guiding behavior, responses should have been biased *towards* it. Instead, responses were biased away from the cued feature. This kind of interference with recall for uncued items goes against

a simpler account where the retrocued item is simply the most likely to be recalled next (see Souza et al., 2014 for a similar argument). Crucially, this bias was not visible in other cueing conditions. It appears that retrocues therefore change the quality of representation of the cued item. While its fidelity may not necessarily increase, its efficacy in changing behavior seems increased.

The repulsive bias could have occurred for two reasons. For one, the appearance of an uncued probe might have prompted observers to remove the representation of the cued orientation from the focus of attention, leading to suppression of and a consequent bias away from the cued orientation. Alternatively, the concurrent representation of two response rules during recall of the uncued item might only be possible if they are orthogonalized to some degree. Behavioral and neuroimaging studies of visual attention have argued that orthogonalization of search templates and visual distractors is optimal for the detection of targets, given a known context of distractors (Navalpakkam and Itti, 2007; Scolari and Serences, 2009; Scolari et al., 2012). Similar retroactive orthogonalization has been shown to occur in recall from visual WM (Kang and Choi, 2015), although it remains to be seen whether this is governed by a similar mechanism. In hippocampal long-term memory recall, orthogonalization is likely required for successful read-out of separable patterns (Kohonen and Oja, 1976; McNaughton and Morris, 1987; Marr et al., 1991; Leutgeb et al., 2004; and possibly also in WM, see Goldman, 2009). Pattern separation may additionally help avoid the

merging of the cued and uncued features in the output gate, or the accidental recall of the cued feature.

Going beyond the within-trial effects reported above, retrocues also affected the reliance of recall on the trial history. In all conditions, recall was biased towards the two orientations probed on the preceding trial (Huang and Sekuler, 2010; Fischer and Whitney, 2014). However, bias was reduced for cued items, compared to uncued items. This finding sheds light on the malleability of the trial history effect. Since retrocues modulated the bias, it did not arise entirely during encoding. Contrary to recent suggestion (Fischer and Whitney, 2014), bias from preceding trials is therefore unlikely to be a purely perceptual phenomenon. Instead, it may point to the adaptive use of multiple sources of memory. Under normal circumstances, recent experience (i.e. the past trials) should be informative for upcoming events, and could be used as a prior to improve recall if items from the current trial were encoded inaccurately (Huttenlocher et al., 2000; Sims et al., 2012). By increasing the accuracy of representations encoded on the current trial, retrocues could reduce the reliance on such a prior (even in circumstances when the prior is uninformative, as it was here).

In sum, this chapter paints a picture of WM as a dynamic architecture that draws on multiple mnemonic stores and that consists of more than a single resource pool. Cues appear to change the functional state of representations, without necessarily withdrawing resources from other items to do so. One possibility is that cued representations change

in quality by occupying the focus of attention, a state that is more apt to guide behavior. Whether or not this state change occurs in the same neural region (Oberauer, 2013; Lewis-Peacock et al., 2014; Zokaei et al., 2014), or is accomplished by transferring a representation to a different brain area, remains to be tested.

Chapter 3: Temporal Dynamics of Attention During Encoding vs. Maintenance of Working Memory

Chapter summary[1]

Working memory (WM) is strongly influenced by attention. In visual working-memory tasks, recall performance can be improved by an attention-guiding cue presented before encoding (precue) or during maintenance (retrocue). Although precues and retrocues recruit a similar fronto-parietal control network, the two are likely to exhibit some processing differences, since precues invite anticipation of upcoming information, while retrocues may guide prioritisation, protection, and selection of information already in mind. This chapter explores the behavioral and electrophysiological differences between precueing and retrocueing in a new visual working-memory task designed to permit a direct comparison between cueing conditions.

Marked differences appeared in event-related potential (ERP) profiles between the precue and retrocue conditions. In line with precues primarily generating an anticipatory shift of attention toward the location of an upcoming item, a robust lateralization in late cue-evoked potentials was associated with target anticipation. Retrocues elicited a different pattern of ERPs that was compatible with an early selection mechanism, but not with stimulus anticipation. In contrast to the distinct ERP patterns, alpha band (8-14 Hz) lateralization was indistinguishable between cue

[1] The experiment in this chapter has been published as: Myers NE, Walther L, Wallis G, Stokes MG, Nobre AC (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 27:492–508.

types (reflecting, in both conditions, the location of the cued item). Whereas alpha-band lateralization after a precue is likely to enable anticipatory attention, lateralization after a retrocue may instead enable the controlled spatiotopic access to recently encoded visual information.

Introduction

Visual working memory (WM) is known to be malleable by selective attention (Oberauer, 2002; Awh et al., 2006; Postle, 2006; Gazzaley and Nobre, 2011; Nobre and Stokes, 2011). In partial-report paradigms, the presentation of a selective cue preceding the onset of the memory array ('precue') effectively reduces memory load, leading to improved performance. Importantly, selective cues presented *after* encoding ('retrocues', Griffin and Nobre, 2003; Landman et al., 2003; Pertzov et al., 2012), during the maintenance delay, can also improve performance. As with precues, retrocues increase the probability of accurate recall (Murray et al., 2011; 2013; Williams et al., 2012), possibly by shifting the cued item into a more robust representational format (Sligte et al., 2008; Oberauer, 2013; Rerko and Oberauer, 2013; Rerko et al., 2014). Prioritization in WM may rely on similar mechanisms as in anticipatory attention (Chun, 2011, Chun et al., 2011). In order to chart the possible commonalities and differences in anticipatory and retrospective shifts of attention, I compared the dynamics of spatial attention mechanisms triggered by precues and retrocues using electroencephalography (EEG).

Research on the neural effects of anticipatory spatial cues has revealed a number of event-related potentials (ERPs) that reflect the orienting of attention in space. Broadly, they have been grouped into early potentials related to processing the cue and shifting attentional focus (see Methods for an introduction to the early potentials used in this study), and later potentials related to anticipation. Among later potentials

(beyond 500 ms after cue onset), the late directing attention-related positivity (or LDAP, Harter et al., 1989; Hopf and Mangun, 2000; Eimer et al., 2003; Murray et al., 2011) has been linked to anticipatory processing of the upcoming task-relevant stimulus. Similarly, the lateralization of alpha-band oscillations has been hypothesized to regulate visual excitability in specific regions of cortex so that anticipated information at an attended location can receive prioritized processing (Foxe et al., 1998; Worden et al., 2000; Jensen and Mazaheri, 2010, but see Pfurtscheller et al., 1996). One possibility is that directed alpha-band lateralization leads to the preferential routing of attended visual input into higher-level areas through changes in excitability of the underlying neural population (Jensen et al., 2012).

Compared to anticipatory cueing, the neural consequences of retrocueing are much less explored. The behavioral effects of retrocues are generally of comparable magnitude as those of precues, and retrocues activate a similar fronto-parietal control network (Nobre et al., 2004; Chun et al., 2011). Moreover, retrocues share some of the early ERP markers of precues (Griffin and Nobre, 2003), and retrospective cues have been shown to modulate activity in visual regions involved in the encoding of WM information (Lepsien and Nobre, 2007; Sligte et al., 2009; Kuo et al., 2011; Munneke et al., 2012; Kuo et al., 2014). Nevertheless, there may still be important differences in the attention-related neural mechanisms triggered by retrocues compared to precues.

In contrast to the effects on control networks and ERPs, the effect of retrocues on neural oscillations, and on posterior alpha oscillations in particular, is unknown. During WM maintenance, a number of studies have shown posterior alpha power *increases* (Klimesch et al., 1999; Jensen et al., 2002; Haegens et al., 2010; Spitzer and Blankenburg, 2012). The increase in posterior alpha power during WM delays has been interpreted as an active inhibitory process designed to limit interference from distracting visual information (Sauseng et al., 2009; Bonnefond and Jensen, 2012). Anticipatory cueing studies typically require the selection of a visual stimulus. In contrast, studies finding WM-related alpha-power increases either presented non-visual material for rehearsal (such as vibrotactile stimuli, which do not require processing in visual cortex, Haegens et al., 2010; Spitzer and Blankenburg, 2012), or they employed a verbal WM task (Sternberg, 1966), which requires the rehearsal of letter strings (Bonnefond and Jensen, 2012). In the latter task, once visual information has been transformed into a phonological code, visual cortex presumably becomes task-irrelevant. It is unclear whether alpha power also increases in a task encouraging storage of information as a visual code, and how retrocues would then modulate alpha power.

The current study was primarily interested in testing whether comparing the oscillatory responses to anticipatory and retrospective shifts of attention can arbitrate between different models of how retrocues improve WM. The improvement could be achieved, (1) by maintaining spatial attention towards the cued location, possibly in

anticipation of the probe. In this case, retrospective cues would lead to a contralateral *decrease* in alpha power, similar to what is observed after an anticipatory cue and possibly reflecting probe anticipation. Alternatively, (2) it could arise through protection from interference, by inhibiting locations where irrelevant input appears. This should lead to a relative *increase* in contralateral alpha power in order to limit interference. Finally, (3) transient access to the cued item and recoding into a preferred representational format (such as the 'focus of attention', see Oberauer, 2013, Cowan, 2000) may be the cause of retrocueing benefits. In this case, a temporary decrease in alpha power might suffice to allow access to areas storing cued information before it is transformed into a different representational state.

I examined this question using a novel task design that allowed me to address these possibilities. In the past, task design has complicated direct comparisons of the neural mechanisms involved in precue and retrocue processing, since many studies have not controlled for some important nuisance factors. For one, precues have tended to appear earlier in the trial sequence (before presentation of the memory array) than retrocues (*after* presentation of the memory array, see Griffin and Nobre, 2003; Nobre et al., 2004). Since the two cue types have appeared at different times within a trial sequence, serial position in the trial complicates the interpretation of any differences. In the current study, the task design was adapted to compare modulatory mechanisms evoked by precues and retrocues. The current design equates all aspects of

stimulus presentation between cueing conditions, so that only the nature of the attention shift (anticipatory or retrospective) differs between cue types. A spatial cue, presented at the same serial position in every trial, could increase the relevance of either, (a), one of two items presented in a previous array (in the retrocue condition), or, (b), one of two memory items from an upcoming second array (precue).

ERPs linked to attentional selection (see Methods) and to sustained anticipatory attention (LDAP) were compared between cue conditions. Firstly, early potentials linked to attentional target selection and alpha-band lateralisation were analysed in response to the different cue types. Early potentials should be similar for prospective and retrospective cues, reflecting attentional orienting to a cued spatial location. In contrast, the LDAP and alpha lateralization should show differences between conditions, since they have been associated primarily with stimulus anticipation. Prospective cues, where relevant information is still forthcoming, might lead to a stronger LDAP and alpha lateralization than retrospective cues, where relevant information has already been retrieved and must be protected from upcoming interference. These results might then inform different theories of prioritization in WM, and the possible functions of alpha oscillations more generally.

The results revealed some unexpected effects, shedding light on common as well as distinct mechanisms of anticipatory and retrospective attention shifts. ERP responses were largely as predicted: early ERP

responses were lateralized to both cue types, while an EEG marker of anticipatory attention (LDAP) robustly differentiated between retro- and prospective cues. In addition, early stimulus-evoked processing (P1) was attenuated after a retrocue and amplified after a precue, especially contralateral to the cued side. Surprisingly, this occurred even though the spatiotopic response of alpha synchronization was identical in anticipatory and retrospective attention to WM items. The unexpected alpha-band response to retrocues favors the interpretation that alpha-band lateralization during retrocueing may reflect access to spatiotopically stored information, rather than perceptual anticipatory attention.

Methods

Participants

Twenty-four healthy volunteers between the ages of 18 and 35 participated in a sequential VSTM task (Fig. 1a). Data from six participants contained too many eye movements or blinks to be analyzed (more than 50% of trials), leaving data from 18 participants (age range 19-33 years, 11 female) in the final analysis.

Precision Working-Memory Task

EEG data were collected while participants performed a precision visual working-memory task (see Bays and Husain, 2008). The task was modified to include the presentation of a spatial cue in between two serially presented memory arrays (see Fig. 1a). Participants viewed two arrays, each containing two peripherally presented oriented bars. The presentation of the arrays was separated by a spatial or neutral cue. Participants were required to judge whether a probe appearing was rotated clockwise or counter-clockwise relative to the memory item that had been in the same location. Each trial contained a total of four memory items, each presented in a separate quadrant of the screen. Spatial retrocues (towards a location in the first array) and precues (towards a location in the upcoming second array) indicated the item that was likely to be probed (with a validity of 70%). On invalidly cued trials, one of the other three items was probed (each with equal likelihood).

Each trial began with a fixation period (400 ms), followed by memory array 1 (presented for 200 ms). Each memory array consisted of two oriented bars (length 1.5° visual angle, width 0.2°, with a small disc of 0.6° diameter at the bar's center to denote the point of rotation). In each array, stimuli were centered in diagonally opposite quadrants of the screen (i.e., either the top left and bottom right quadrants, or the top right and bottom left quadrants), at 5° eccentricity from fixation. After an inter-stimulus interval (ISI) of 800 ms, a cue appeared (200 ms). Neutral cues (a dark gray square of side length 0.4° at fixation) were uninformative about which item would be probed. Spatial cues (a thin white border of 0.1° along two adjacent edges of the gray cue square, forming an arrow pointing to one of the four quadrants) indicated the location of the item that would be probed (with a validity of 70%). If the cued item had already been presented in array 1, the cue acted as a retrocue. Alternatively, the cue acted as a precue if it indicated an item to be presented in array 2 (i.e., not already presented in array 1). Memory array 2 appeared (also for 200 ms) after cue offset and another ISI of 800 ms. To prevent spatial overlap across arrays, the two stimuli were presented in the two quadrants that were unoccupied in array 1. After a further ISI of 800 ms, a probe stimulus appeared (for 200 ms) at the same location as one of the four memory items. The probe was an oriented bar similar to the item at that location, but rotated relative to the original orientation. Participants made a binary response indicating whether the probe was rotated clockwise or counter-clockwise, relative to the item at the probed

location. The rotation angle was varied between memory and probe angles between 2° and 39°. Rotation angles were binned into three categories: easy (rotation angles uniformly distributed between 21 and 39°), medium (11-21°), and hard (2-11°). The bins of the small and medium angles were narrower than the large angle bin to improve estimation of memory precision (see *Behavioral Analysis*) and to encourage participants to remember stimuli as accurately as possible. Responses were made using a mouse with a trackball that could be rotated clockwise or counter-clockwise. Participants were explicitly encouraged to make use of cues to improve their accuracy. Reaction time was not emphasized. Inter-trial intervals were uniformly jittered between 1.25 and 1.75 seconds, with an average duration of 1.5 seconds.

Participants completed the experiment in one session after two practice blocks. The task consisted of ten 7-minute blocks, each comprising 64 trials, for a total of 640 trials. Of these, 400 trials were cued (200 retrocue, 200 precue), and 240 trials were neutral (with 120 probing array 1, and 120 probing array 2), meaning that a cue was presented on 62.5% of trials. Of the 400 cued trials, 280 trials (140 retrocue, 140 precue) were validly cued. Participants sat in a dimly lit booth at a distance of 74 cm from the monitor (22 inch Samsung SyncMaster 2233; resolution: 1680 * 1050 pixels; refresh rate: 60Hz; screen width: 47cm). A chin rest was used to help participants keep their heads still. Binocular gaze locations were continuously recorded with a desktop mount video-based eye tracker at 500 Hz (EyeLink 1000, SR Research, Ontario,

Canada), allowing for offline detection of blinks and saccades. All trials in which gaze deviated from the central fixation dot by more than 1.5° were discarded. Participants were instructed to fixate on the center of the screen at all times and not to blink during trials. Eyetracker data from three participants were too noisy for analysis. In these cases the electro-oculogram was used for blink and saccade detection (see below).

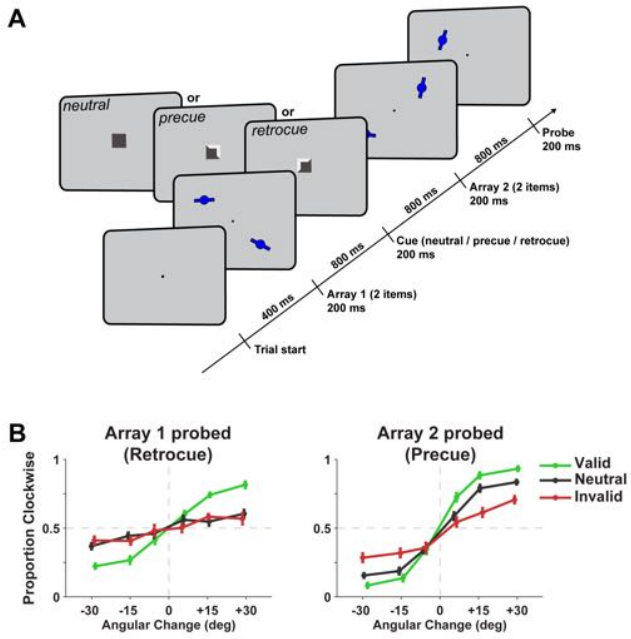


Figure 1. Task Design and Behavioral Results. A. Example Trial Structure. Each trial began with the onset of a fixation cross, followed by the display of a memory array consisting of two oriented bars. Participants remembered the orientation of each bar. After a 800-ms delay, a cue appeared. Informative cues could point to a quadrant that had contained an item in the first array (retrocues), or to a previously unoccupied quadrant (precues). The cued item was probed in 70% of trials. Uninformative (neutral) cues gave no information about which item would

be cued. Following the cue, a second array appeared, with two items in previously unoccupied quadrants. After a further delay, a probe stimulus appeared at the location of one of the previous stimuli. The probe appeared at the same location, but was rotated clockwise or counter-clockwise (relative to the probed memory item). The magnitude of rotation changed from trial to trial and ranged from 2 to 39°. Participants judged the direction of rotation, and received visual feedback in the form of a green (correct response) or red (incorrect) fixation dot during the inter-trial interval (not shown). B. Behavioral Performance. Each panel shows the proportion of clockwise responses as a function of the magnitude and direction of the relative probe rotation (angular change), separately for valid, neutral, and invalid cues (the invalid condition corresponds to trials probing an uncued item from the respective array, irrespective of whether the cue pointed to the other item in the same array, or one of the items in the other array). When array 1 was probed (retrocue trials, left panel), valid cues significantly improved accuracy, primarily by reducing guess rate (i.e. the offset between the asymptote of the sigmoid response function and 0 or 1). Memory precision (i.e., the slope of the function) was not significantly affected by cue validity. Precue trials (right panel) showed better performance overall, but the same effect of cue validity. Here, costs of invalid cues are also visible. Both plots show mean responses across all 18 observers. Error bars indicate s.e.m.

Behavioral Analysis

In a first step, a three-way repeated-measures ANOVA on response accuracy evaluated the effects of: array probed (array 1, array 2), cue validity (valid, neutral, invalid), and trial difficulty (easy (21-39° rotation angle), medium (11-21°), hard (2-9°)). In a second step, the variable trial difficulty was used to test the effects of cue validity and of array on the likelihood of correct recall and the precision of memory. Observers were assumed to respond at random on trials where they had no memory of the probed item, while on remembered trials they would respond with Gaussian variability. Therefore the behavioural data were described by fitting a mixture model (Zhang and Luck, 2008), consisting of a uniform distribution (to account for the proportion of random guesses) and a von Mises distribution (to account for the precision of remembered responses). The model consisted of two free parameters:

$$P(\text{correct}|\Delta) = (1-\mathbf{pGuess})*\mathbf{vonMises}(\Delta,\mu,\kappa) + \mathbf{pGuess}/2;$$

(Equation 1)

where *P(correct|Δ)* is the probability of a correct response given a probe that is rotated Δ degrees relative to the remembered stimulus, *vonMises* is the cumulative von Mises distribution (i.e. the circular analogue of the normal distribution, which can be evaluated numerically) with mean μ (fixed at 0) and precision κ (kappa, proportional to the inverse of the standard deviation), evaluated at Δ. *pGuess* determines asymptotic performance (i.e. the guess rate, or proportion of random responses).

Estimates of memory precision (kappa) and guess rate (pGuess) were then compared across conditions using a repeated-measures ANOVA with factors cue validity and array probed.

EEG Acquisition

The EEG was recorded continuously with NuAmp amplifiers (Neuroscan Inc., Albany, USA) from 40 Ag/AgCl electrodes. Electrodes were positioned according to a subset of the 10-10 international system (AEEGS, 1991). Recordings were taken from the following electrodes: Fz, FCz, Cz, CPz, Pz, POz, Oz, FP1/2, F3/4, F5/6, FC3/4, FC5/6, C3/4, C5/6, CP3/4, CP5/6, P3/4, P5/6, PO3/4, PO7/8, O1/2. Blinks and eye movements were monitored by deriving bipolar recordings from an electrode placed below the right eye and FP2 (vertical electrooculogram, VEOG) and from electrodes placed near the outer canthi of each eye (horizontal EOG, HEOG). The electrode in position AFz was used as the ground. The right mastoid was used as the active reference. Recordings were also taken from the left mastoid to derive an average-mastoid reference off-line. Electrical impedance was kept below 5 kΩ and activity was filtered with a low-pass filter of 300 Hz. The analogue-to-digital sampling rate of the brain activity was set at 1,000 Hz and data were recorded continuously for the entire experiment.

EEG Processing

Processing and EEG analyses were done offline, using the EEGLAB toolbox (Delorme and Makeig, 2004) and custom Matlab scripts for preprocessing. First, the EEG was re-referenced to the average of the left and right mastoid, and bandpass-filtered between 0.1 and 40 Hz (using a second-order Butterworth filter). The HEOG and VEOG data were bandpass-filtered between 1 and 40 Hz, and smoothed the filtered timeseries with a sliding boxcar window of 11 ms length to remove brief spikes that were not related to eye movements. Next, the EEG and EOG were segmented into epochs ranging from 1,600 ms before the cue (600 ms before the first stimulus array) to 3,000 ms after the cue (1,000 ms after the onset of the probe stimulus). Trials were automatically rejected if EOG amplitudes were outside ±75 μV surrounding array 1 (-200 to +400 ms), the cue (-200 to +1,600 ms), array 2 (-1,200 to +600 ms), or the probe (-200 to +600 ms). Eyetracking data (see above) and the HEOG were used to eliminate trials with saccades to the stimuli. Finally, all epochs were visually inspected for any further artifacts. For ERP analyses of cue- and array-locked potentials, the data were baseline-corrected using a short interval before the onset of the event of interest (200 ms before cue onset and 50 ms before onset of array 2). The shorter baseline period before onset of array 2 was chosen to minimize any potential effect of late, slow potentials evoked by the cue. When calculating lateralization measures, on trials with a right-side cue, the channel topography was flipped along the midline to obtain an index of

lateralization (contra- versus ipsilateral to the cue). On neutral trials, the topography was flipped if the *probed* item was on the right.

ERP Analyses

One of the earliest potentials showing cue-dependent lateralization is the early directing attention negativity (or EDAN, Harter et al., 1989). It has been interpreted as a consequence of attentional orienting towards a relevant spatial location (Harter et al., 1989; Yamaguchi et al., 1994; Hopf and Mangun, 2000; Nobre et al., 2000; Murray et al., 2011), though in some cases it may also reflect the selection of relevant features of the cue stimulus itself (van Velzen and Eimer, 2003). The latter interpretation of the EDAN parallels that of another lateralized potential – the posterior contralateral N2 (N2pc, Eimer 1996, or posterior contralateral negativity – PCN, e.g., Töllner et al., 2012). The N2pc (or PCN) is elicited when a target item is identified and selected from a spatially distributed array (Luck and Hillyard, 1994; Eimer, 1996; Woodman and Luck, 1999, 2003; Hopf et al., 2002). Much like the EDAN, it consists of an accentuated negative potential over posterior contralateral electrodes. The EDAN is usually succeeded in time by the anterior directing attention negativity (ADAN, Eimer et al., 2002; Praamstra et al., 2005; Green and MacDonald, 2006), which is another lateralized component that has been associated with the controlled deployment of spatial attention (e.g., Van Velzen and Eimer, 2003). This component is usually observed over frontal sensors,

peaking approximately between 300 and 500 ms after a cue (e.g., Kelly et al., 2009).

Trials were averaged separately for the three conditions of interest (retrocue, precue, or neutral), and 3X2 ANOVAs with factors cue type (retrocue, precue, neutral) and side (contralateral, ipsilateral) were used to test for the effects of cueing and of lateralization in a number of potentials, defined *a priori* based on previous literature testing anticipatory cueing effects (Hopf and Mangun, 2000; Nobre et al., 2000; Murray et al., 2011). Relative to the cue, the potentials of interest were the N1 (150-180 ms after cue onset, at PO7/8), the N2 (240-300 ms, PO7/8, i.e. the N2pc or PCN), the early directed attention-related negativity, or EDAN (250-350 ms, PO7/8, O1/2, Mangun, 1994; Hopf and Mangun, 2000), the anterior directed attention-related negativity, or ADAN (350-500 ms, FC3/4 and C3/4), and the late directed attention-related positivity, or LDAP (750-1000 ms, PO7/8 and O1/2). On neutral trials, lateralization was calculated relative to the side of the probed item.

Relative to array 2, the potential of interest was the early visual response as measured with the P1 (60-140 ms after array onset, tested at PO7/8 and O1/2). I also tested for more persistent effects on encoding by examining the amplitude of the N1 (150-180 ms) and P2 (300-400 ms) potentials, as well as lateralization in the N2 window (N2pc, 240-300 ms, PO7/8).

Analysis of Alpha Power Lateralization

To test for the effects of cues on desynchronization in the alpha band (8-14 Hz) and in surrounding frequencies, the broadband signal over 6 occipital-parietal sensors (O1/2, PO7/8, P5/6) was decomposed using Hanning tapers (window width 5 cycles), estimated at 0.5 Hz steps between 5 and 20 Hz (using custom Matlab scripts and the Fieldtrip Toolbox, Oostenveld et al., 2011). After log transforming ($10*\log_{10}$) the power at each frequency, single-trial data were baseline-corrected by subtracting the average power in the 200 ms preceding the onset of array 1 (i.e., 1200 ms before cue onset). Averaged power was then averaged across bins in the alpha range (8-14 Hz), and averaged separately across contralateral and ipsilateral sensors (relative to the cued side on cue trials, or relative to the probed side on neutral trials) at each time point after cue onset. Contra- and ipsilateral alpha power was then averaged between 500 and 1000 ms after cue onset (i.e., up to the onset of array 2), and tested for effects of lateralization and of cue type in a 2X3 ANOVA.

Quadrant-Specific Analysis of Alpha Power

The preceding analysis only tested for lateralization of alpha power following a cue to the left or to the right. However, cues were directed at a specific quadrant, warranting a more detailed analysis of the topographical distribution of alpha power. Specifically, the pattern of alpha desynchronization might be specific to the cued quadrant, rather than the entire hemifield. In this task, the spatial specificity of alpha

desynchronization could carry additional importance because the uncued quadrant (of the cued hemifield) always contained a (previously seen or an upcoming) distractor. To test this hypothesis, I used an encoding model to estimate the topographical pattern of alpha power desynchronization following visual stimulation in each quadrant, and applied this model to the cue-induced alpha-power topographies to estimate the specificity of desynchronization for the cued quadrant. Conceptually, this analysis is identical to encoding models that have been used to estimate information about the orientation, color, or shape of a stimulus, using fMRI (Brouwer and Heeger, 2009; 2011; Naselaris et al., 2011; Serences and Saproo, 2012) or EEG (Garcia et al., 2013, see also Chapter 5).

For each trial in the experiment (i.e., the test trial), a general linear model (fit to all other trials in the experiment, i.e., the training trials) was fit to the topographical distribution of alpha power evoked by a probe presented in each of the four quadrants (using neutral and valid cue trials together and averaging over 8 to 14 Hz and 200 to 800 ms after probe onset, when the effect was strongest). The GLM was solved via ordinary least squares:

$$W = B_1C_1'(C_1C_1')^{-1}; \qquad \text{(Equation 2)}$$

where C_1 is the design matrix (4 regressors x no. of training trials, with each regressor set to 1 for trials with a probe in the corresponding

quadrant, and set to 0 otherwise), B_1 is the training data set (probe-evoked alpha power at 35 EEG sensors x no. of training trials), and W is the estimated weight matrix (35 sensors x 4 regressors/quadrants). Given that the absolute power in a given frequency band might have differed across sensors, variability in power across the topography could have decreased the sensitivity of this analysis. However, a control analysis using a fifth, constant regressor to model out the mean at each sensor yielded nearly identical results. Therefore, differences in power between sensors were ignored in the subsequent analyses.

W was used to estimate the quadrant-wise activation in the test trial, B_2 (35 sensors x 1 trial):

$$C_2 = (W'W)^{-1}W'B_2; \text{(Equation 3)}$$

where C_2 is the 'activation' (in arbitrary units) in each quadrant. For each trial, this activation was sorted relative to the cued quadrant. This procedure was repeated (using the same weight matrix W) for each time point in the trial and each frequency (between 5 and 20 Hz) before moving to the next iteration in the leave-one-out procedure. A 2x4 ANOVA with factors cue type (precue, retrocue) and quadrant (cued, same hemifield, opposite hemifield, and diagonally opposed) was then used to test for cue-evoked quadrant-specific activation after the onset of array 1, the cue, array 2, and the probe. Further, the cued quadrant was classified on each trial, time point, and frequency (using a simple voting

procedure that assigned a quadrant label according to the regressor with the highest activation level) to test the accuracy of the model. Classification accuracy was then compared to chance (25%) using one-sample t-tests. To assess the effects of cueing, a window-of-interest approach was used by averaging the classification accuracy from 8 to 14 Hz and 500 to 1000 ms after cue onset.

Results

As a point of reference, Table 1 provides a schematic summary of all results described below.

Behavioral Results

I conducted a 3x2x3 ANOVA (Fig. 1b) on accuracy with factors cue validity (valid, invalid, neutral), array probed (array 1, array 2), and probe difficulty (hard, medium, easy). Main effects of cue validity ($F_{2,34}$=36.660, p=3.3*10^{-9}), array ($F_{1,17}$=64.076, p=3.6*10^{-7}), and difficulty ($F_{2,34}$=76.028, p=2.8*10^{-13}) were all significant; as well as significant interactions between cue validity and array ($F_{2,34}$=7.389, p =0.002), cue type and difficulty ($F_{4,68}$=7.663, p=3.70*10^{-5}), and array and difficulty ($F_{2,34}$=3.599, p=0.038). The three-way interaction between cue validity, array, and difficulty was also significant ($F_{4,68}$=3.882, p=.007).

Given the significant three-way interaction, post-hoc analyses were used to test whether both precues (array 2 probed) and retrocues (array 1 probed) individually affected accuracy. For trials on which array 1 was probed ('retrocue' trials), there were significant effects of cue validity ($F_{2,34}$=17.77, p=5.1*10^{-6}) and difficulty ($F_{2,34}$=20.40, p=1.5*10^{-6}), along with a significant interaction ($F_{4,68}$=3.50, p=0.012). This confirmed that: (a) retrocues improved accuracy and that (b) the cueing benefit was largest on the easiest trials (t-test comparing valid retrocues with neutral trials for easy probes: t_{17}=6.51, p = 5.4*10^{-6}, Cohen's d=1.82, medium probes: t_{17}=4.08, p=7.8*10^{-4}, d=1.33, difficult probes: t_{17}=1.49, p=0.153, d=0.49).

Compared to neutral cues, invalid retrocues (that is, trials probing a non-cued item from array 1) did not decrease performance (t-test neutral against invalid cues, $t_{17}<1$, $p>0.40$ for all three difficulties). However, this null-result does not imply that invalid retrocueing carries no cost, as accuracies were sufficiently low (neutral: 57.7%, invalid: 55.7%) to have produced a floor effect. A similar (but stronger, given the three-way interaction) pattern of results was present when array 2 was probed ('precue' trials; effect of cue validity: $F_{2,34}=37.72$, $p=2.3*10^{-9}$, effect of difficulty: $F_{2,34}=57.97$, $p=1.1*10^{-11}$, interaction: $F_{4,68}=8.75$, $p=9*10^{-6}$). There was a graded increase in effect size on valid trials, as indicated by an interaction between cue validity and difficulty: Valid cues showed a larger benefit on easier trials (easy: $t_{17}=7.53$, $p=8.19*10^{-7}$, $d=0.94$, medium: $t_{17}=3.87$, $p=0.001$, $d=0.83$, difficult: $t_{17}=2.79$, $p=0.006$, $d=0.70$). In addition, on precue trials there were significant costs of invalid cues (i.e., trials probing a non-cued item from array 2) on easy and medium difficulty trials (invalid vs. neutral, easy: $t_{17}=-4.13$, $p=0.0007$, $d=0.99$, medium: $t_{17}=-5.76$, $p=2.3*10^{-5}$, $d=1.79$, hard: $t_{17}=-0.71$, $p=0.488$, $d=0.22$). These results confirm that both valid precues and retrocues significantly improve accuracy, in spite of the baseline accuracy differences between arrays 1 and 2. In other words, retrocues (to items in array 1) improve memory in spite of interference (from items in array 2).

I used a model-based approach to dissociate whether cueing affected the precision or recall rate of remembered items. Cueing affected only the guess rate ($F_{1,17}=7.81$, $p=0.002$), with no significant effect on

precision (kappa, p=0.987), indicating that cues primarily improved the likelihood of recalling an item, but not its fidelity. Whether the first or second array was probed also affected guess rate ($F_{2,34}$=17.4, p<0.001), but not precision (p=0.705). However, there was a significant interaction between cue validity and probed array on precision ($F_{2,34}$=3.676, p=0.036). The interaction reflected a significant difference in precision on valid trials probing either the first or second array (t_{17}=2.635, p=0.0174) that was not present on neutral or invalid trials (p=0.380 and p=0.445, respectively). Thus, validly cued items presented in the second array were more likely to be recalled, and they were recalled with higher precision, than items in array 1. Given these differences in performance between arrays, the cueing effect was assessed separately for array 1 (retrocues) and for array 2 (precues). Retrocue validity significantly improved guess rate ($F_{2.34}$=3.98, p=0.028, t-test valid vs neutral: t_{17}=3.43, p=0.003), but not precision (p=0.256). Similarly, precues affected only guess rate ($F_{2.34}$=7.255, p=0.002, t-test valid vs neutral: t_{17}=4.17, p=0.0006) but not precision (p=0.152).

	Behavior		ERPs (cue)				Alpha-Band Effects		ERPs (array 2)		
	Recall Rate	Precision	N1 / N2pc	EDAN	ADAN	LDAP	Lateralization	Quadrant-based Desynchronization	P1	N2pc	P2
Retrocue	+	ns	+	+	+	ns (rev.)	+	+	red.	ns (rev.)	red.
Precue	+	ns	ns	+	ns	+	+	+	+	ns	+

Table 1. Summary of Behavioral and Neural Cueing Effects. + indicates a significant effect, ns indicates no effect, and red./rev. indicates a significant reduction or reversal of an effect, respectively (marginally significant effects in parentheses).

Event-related Responses to the Cue

The cue-evoked lateralization of the ERP was used as a marker of spatially directed attention to previous or upcoming information (Fig. 2), focusing on three well-known ERP markers of directed attention: EDAN, ADAN, and LDAP. In addition, I tested for modulation of the N2pc (or PCN) related to spatiotopic mechanisms of target selection.

In general, both retrocues and precues modulated markers of attentional orienting. The EDAN was present after both retrocues and precues (t_{17}=-7.06, p<0.0001 for retrocues, t_{17}=-2.23, p=0.04 for precues, p=0.28 for neutral), although it was significantly stronger after a retrocue (side*cue type interaction $F_{2,34}$=12.33, p<0.001). The intermediate stage ADAN was not strongly modulated by either cue type (t_{17}=-2.33, p<0.05 for retrocues, t_{17}=-1.44, p=0.16 for precues, p>0.90 for neutral, side*cue type interaction p>0.20). In contrast, the LDAP, immediately preceding the onset of the second array (250 ms to 0 ms pre-stimulus), showed lateralization for both retrocues and precues, but in opposite directions. While precues showed a modest relative contralateral positivity (t_{17}=2.35, p=0.031), retrocues exhibited a relative contralateral *negativity* (t_{17}=-2.417, p=0.027), resulting in a significant interaction between side and cue type in the interval from 750-1000 ms ($F_{2,34}$=11.17, p=0.004). In addition, there was a main effect of cue type ($F_{1,17}$=10.07, p=0.006). This main effect was driven by more negative potentials after precues: while retrocue and neutral trials did not differ ($F_{1,17}$=1.01, p=0.33), there was a difference in

potential between precue and neutral trials ($F_{1,17}$=11.04, p=0.004). The difference between precues and retrocues was strongest over ipsilateral sensors (t_{17}=4.21, p=0.0006), with only a trend for a difference at contralateral sites (t_{17}=2.02, p=0.06).

At the earliest processing stage, attentional orienting effects were specific to retrocues. The N1 was significantly lateralized after a retrocue (paired t-test contra- vs. ipsilateral sensors, t_{17}=-4.39, p=0.0004), but not after precues or neutral cues (p>0.38 for both, side*cue type (retro/precue) interaction $F_{2,34}$=4.45, p<0.05). Likewise, the N2pc was modulated only after a retrocue (t_{17}=-4.03, p=0.0009, p>0.20 for precues and neutral, side*cue type interaction $F_{2,34}$=6.60, p<0.05).

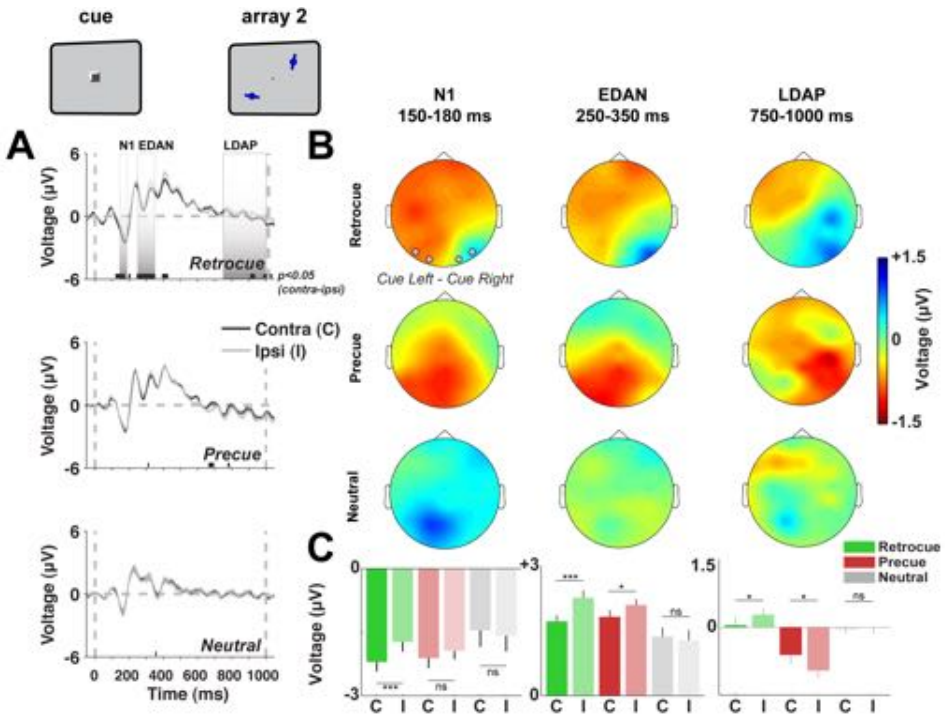


Figure 2. ERP Responses to the Cue. A. Grand average potential timecourses taken from visual sensors (O1/2, PO7/8, see top left

topography in B for locations). Visual ERPs showed modulations by both retrocues (top panel) and precues (middle panel), lateralizing in response to the cued side (contralateral sensors: dark lines, ipsilateral: gray lines). Neutral cues did not elicit any lateralization. The three shaded areas indicate time ranges used to calculate topographies and bar plots in B and C. B. Average lateralization topographies at times of interest in the N1 (left column), EDAN (middle column), and LDAP (right column) ranges. Topographies show difference between potentials after left cues and right cues (Left minus Right). C. Bar plots showing mean potentials at times and sensors (contralateral, C, and ipsilateral, I) of interest. Error bars indicate s.e.m.

Event-related Responses to Array 2

Regarding the response to the second stimulus array (Fig. 3), I predicted that both items would be attended on neutral trials, the cued item would be attended on precue trials, and no items would be attended on retrocue trials. This prediction was borne out in the magnitude and lateralization of the P1 response, occurring 60-140 ms after stimulus onset. The ANOVA revealed a main effect of cue type ($F_{2,34}$=6.74, p=0.003) and a significant interaction with side ($F_{2,34}$=7.56, p=0.002), indicating that cue types affected P1 magnitude differentially at contra- and ipsilateral sensors. While precues led to the expected contralateral P1 enhancement (paired t-test, t_{17}=2.51, p=0.022), retrocues *attenuated* the P1 contralateral to the cued side (t_{17}=-3.06, p=0.007). Neutral trials did

not show any lateralization (t_{17}=0.41, p=0.689). Overall, the P1 was smaller on retrocue trials than on precue or neutral trials (t_{17}=-3.03, p=0.008 for retrocue vs. precue, t_{17}=-3.01, p=0.008 for retrocue vs. neutral), with no significant difference between precue and neutral P1 (t_{17}=0.94, p=0.363, precue-neutral). In sum, the P1 amplitude reduction is in line with the hypothesis that retrocues suppress processing of the second array. Further, the reduction of the P1 contralateral to the cue indicates an additional, selective suppression in processing of new items in the retrocued hemifield.

How these effects carried forward was tested by analyzing the P1-N1 amplitude difference. The main effect of cue type ($F_{2,34}$=9.33, p=0.001) appeared to be driven by a reduction of the P1-N1 amplitude on retrocued trials (t-test compared to neutral, t_{17}=-3.67, p=0.002, t_{17}=-4.48, p=0.0003 on contralateral and ipsilateral sensors, respectively). The P1-N1 amplitude was not lateralized in either cue condition (p>0.30 for retrocues and precues). N1 amplitude itself (150-180 ms) was not modulated by cue type ($F_{2,34}$=1.19, p=0.318) or by side ($F_{2,34}$=0.12, p=0.730), with no interaction ($F_{2,34}$=1.72, p=0.194).

In potentials in the N2 range (240-300 ms, PO7/8), there was no strong evidence for lateralization: while there was a main effect of cue type ($F_{2,34}$=10.349, p=0.002), there was no main effect of cued side ($F_{1,17}$=2.692, p=0.119) and no interaction between side and cue type ($F_{2,34}$=1.400, p=0.260). A paired t-test between contra- and ipsilateral potentials after precues showed no significant lateralization (t_{17}=-0.289,

p=0.776). After retrocues, following the inverted P1 effect, there was a trend towards a contralateral negativity (t_{17}=-2.043, p=0.054).

Finally, the P2 (over PO7/8 and O1/2), occurring between 300 and 400ms after onset of the array, also showed a main effect of cue type ($F_{2,34}$=8.11, p=0.001). P2 magnitude over more central-parietal sensors (P3/4/5/6, CP3/4/5/6) showed the same pattern ($F_{2,34}$=6.08, p=0.006). In both cases, the magnitude was highest for neutral trials, followed by precue trials. There was no significant P2 lateralization in the central-parietal region of interest ($F_{1,17}$=1.04, p=0.323 over central-parietal cortex), but there was a main effect of cue side over visual cortex ($F_{1,17}$=5.10, p=0.037). The effect over visual cortex (PO7/8, O1/2) did not correspond with significant lateralization in any one cue condition (contra minus ipsi difference, t_{17}=-1.59, p=0.129, t_{17}=-1.20, p=0.247, t_{17}=-1.11, p=0.284, for retrocue, precue, and neutral trials, respectively), and no significant interaction ($F_{2,34}$=0.16, p=0.850).

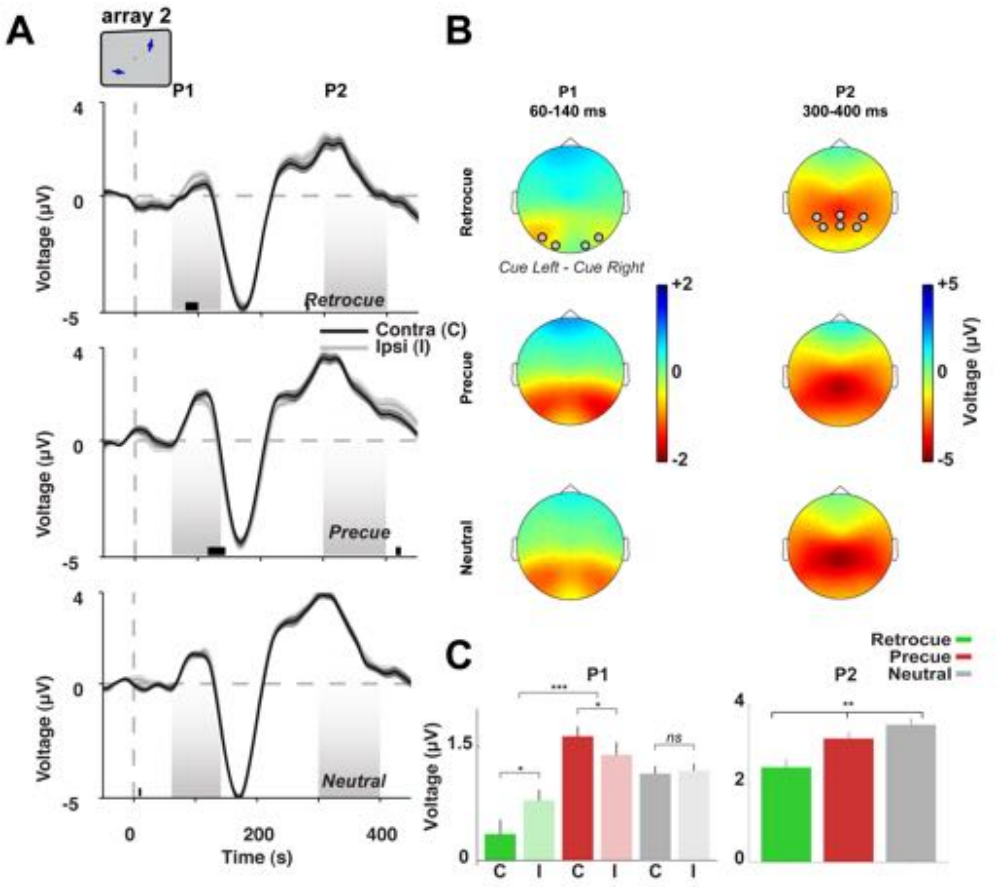


Figure 3. ERP Responses to Array 2. A. Grand average potential timecourses taken from visual sensors (O1/2, PO7/8, see top left topography in B). After retrocues, visual responses (P1 and P2) were attenuated, and P1 was lower at contralateral (dark lines) compared to ipsilateral sites (gray lines). Conversely, precues showed a contralateral increase in P1 amplitude. Shaded areas indicate the time range used to calculate topographies and bar plots in B and C. B. Average lateralization topographies at times of interest in the P1 (left column) and P2 (right column) ranges. Topographies show difference between potentials after left cues and right cues (Left minus Right) for the P1, and average potential for the P2. C. Bar plots showing mean potentials at times and sensors (contralateral, C, and ipsilateral, I) of interest. Sensors of interest

(P1: O1/2, PO7/8, P2: P3/4/5/6, CP3/4/5/6) are indicated on the topographies in the top row. Error bars indicate s.e.m.

Cue Effects on Alpha Power

Overall, the onset of pre-, retro-, and neutral cues led to a temporary desynchronization in the alpha band for roughly 400 ms (see timecourses in Fig. 4b). Between 500 ms and 1000 ms after the cue (leading up to array 2), alpha power gradually rose again. On spatial cue trials, this rebound coincided with a divergence in power between contralateral and ipsilateral sensors (Fig. 4b, main effect of side, $F_{1,17}$=13.66, p=0.002), leading to a significant lateralization both on retrocue trials (t_{17}=-3.19, p=0.005) and precue trials (t_{17}=-3.19, p=0.005), but not on neutral trials (t_{17}=0.08, p=0.93, interaction side*cue type: $F_{2,34}$=3.39, p=0.045). The relative decrease in contralateral alpha power after a precue was predicted. It was somewhat less expected to see the same pattern after a retrocue (since contralateral alpha could have *increased* to inhibit processing of the upcoming stimulus in array 2, in order to protect the cued item). This similarity of alpha lateralization stands in contrast to the marked differences between cue types found in the ERP during the same time period (i.e., the LDAP in the 500 ms leading up to array 2).

Cue-related lateralization was quenched by the onset of the second stimulus array, but then re-emerged in the 250 ms immediately before probe onset (main effect of side, $F_{1,17}$=9.336, p=0.007), between

1750 and 2000 ms after the cue. Again, both cue types were significantly lateralized during this pre-probe interval (retrocues: t_{17}=-2.36, p=0.03, precues: t_{17}=-2.81, p=0.01, neutral: t_{17}=-0.59, p=0.56). On precue trials, this effect emerged earlier, beginning 500 ms before onset of the probe (lateralization 500-250 ms before probe onset, t_{17}=-3.01, p=0.008, p>0.70 for retrocue and neutral trials). For both cue types, the cue-induced alpha-band lateralization was restricted to parietal-occipital sensors (Fig. 4a).

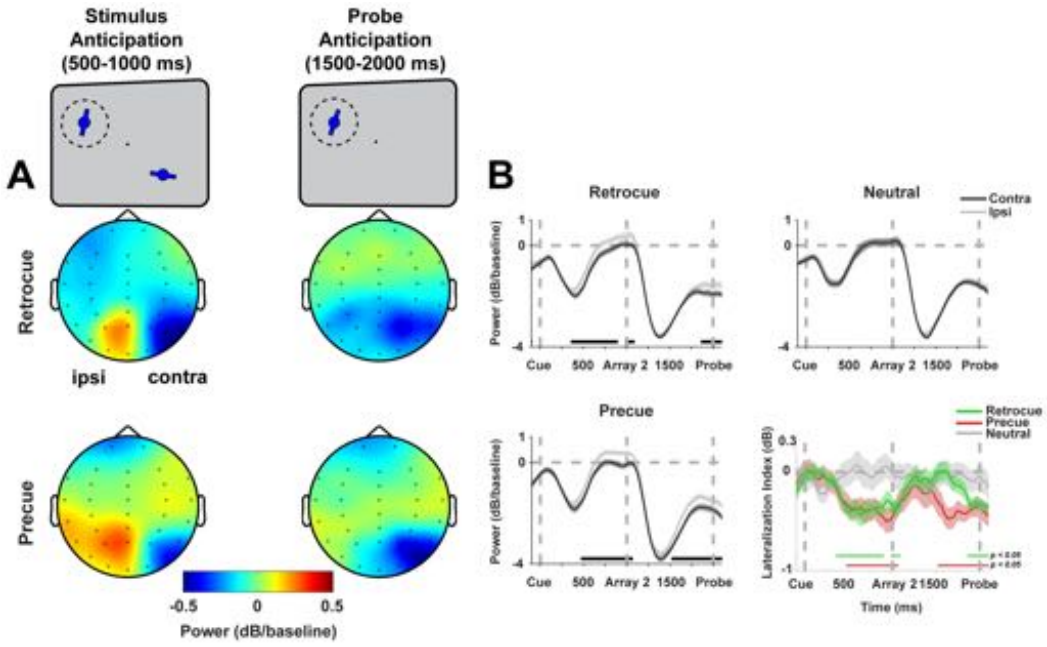


Figure 4. Cue-evoked changes in posterior alpha power (8-14 Hz) after spatial retrocues and precues. A. Topographic distribution of alpha power. Top row shows effects of retrocues after the cue (i.e., preceding the onset of array 2), and before the onset of the probe. Bottom row shows effects of precues in the same time range. Colors indicate power change (relative to prestimulus baseline, in decibels). B. Alpha power timecourses at

posterior sensors (O1/2, PO7/8, P5/6) between cue onset and probe onset. Each plot contains timecourses of contralateral (dark lines) and ipsilateral (light gray) power, separated by cue condition (retrocue/neutral/precue). Black lines at bottom indicate time points of significant lateralization (contralateral minus ipsilateral, p<0.05). The bottom right plot shows lateralization (contra-ipsi) timecourses for the three cue conditions. Green and red lines show time points of significant lateralization for retrocues and precues, respectively. Shaded areas indicate s.e.m.

I additionally examined whether cued lateralization (as measured by alpha-band lateralization or the LDAP) related to behavioral improvements after retrocues or precues. The model-estimated cueing benefit (i.e. the difference in recall rate between cueing and neutral conditions) was tested for Pearson correlations (across participants) with LDAP magnitude and the magnitude of cue-evoked alpha lateralization (contralateral minus ipsilateral power). There was a significant correlation between alpha-band lateralization after retrocues and the difference in recall rate between retrocues and neutral trials ($r=-0.474$, $p=0.0469$), indicating that when alpha power was more lateralized (more negative), the improvement in recall rate (after retrocues) was greater. However, this correlation was only significant at an uncorrected statistical level, and no other correlations with behavior were significant (all uncorrected $p>0.078$). This result should therefore be interpreted with some caution,

particularly since the sample size (N=18) may have been sufficient to show robust within-subject statistics, but fairly low for between-subject correlation.

Quadrant-specific Activation in the Alpha Band

The topographical distribution of alpha power in response to the probe (8-14 Hz, 200-800 ms after onset) reliably distinguished the spatial location of the probe stimulus in all three cue conditions (see Fig. 5a, dashed windows after probe onset, classification accuracy against chance level: $p<0.001$ for neutral ($t_{17}=5.93$), valid precue ($t_{17}=5.74$), and valid retrocue trials ($t_{17}=4.28$)). Using a leave-one-out procedure, this time (200-800 ms after probe onset) and frequency (8-14 Hz) window was used to train quadrant-specific topographies (Fig. 5a) that were then applied to the test trial at all time points and frequencies (5-20 Hz). Quadrant classification after cue onset (Fig 5c, 500 to 1,000 ms after cue onset, 8-14 Hz, see also dashed windows after cue onset in Fig. 5b) was significant on both retrocue ($t_{17}=2.41$, $p=0.028$) and precue trials ($t_{17}=2.84$, $p=0.011$), but not on neutral trials ($t_{17}=-0.06$, $p=0.953$). In addition, on precue trials the cued quadrant was decodable in the period leading up to the probe (1,500 to 2,000 ms after cue onset, $t_{17}=3.64$, $p=0.002$, $p>0.50$ on retrocue and neutral trials). The latter effect accompanies the earlier re-emergence of alpha lateralisation after array 2 on precue trials (Fig. 4b). The analysis described below, examining quadrant-specific activation, may elucidate why, on *retrocue* trials, alpha lateralization did

seem to appear just before probe onset (1,750 to 2,000 ms after the cue, Fig. 4b), even though there was no significant classification of the cued quadrant in this epoch using the decoding approach.

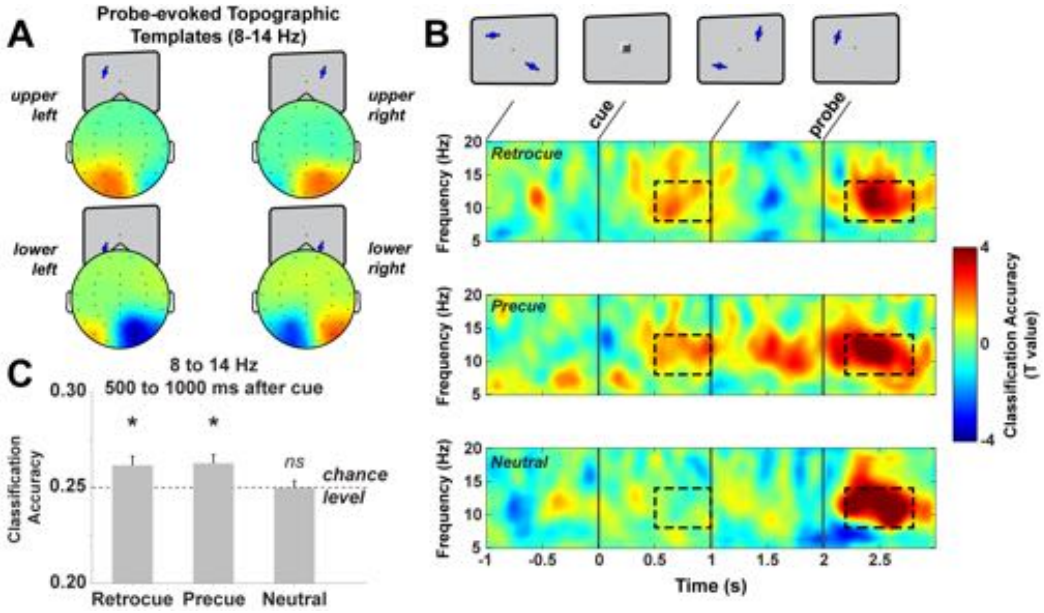


Figure 5. Quadrant Decoding using Alpha Band Power (t-test vs. chance, 25%). For each trial, the 8-14 Hz topographies evoked by probe onset were averaged over all other trials, separately for probes presented in the four screen quadrants. A. shows the topographies, averaged across all trials and all participants. For illustration, the average topography has been subtracted from each quadrant-specific topography. The topography of the left-out test trial was then calculated at each frequency and time point in the trial, and regressed against the four template topographies. The stimulus quadrant corresponding to the best-fitting topography was voted to have evoked the pattern on the test trial, and was compared to the actually cued quadrant (or probed quadrant on neutral trials) to calculate classification accuracy. B. shows t-statistic

maps of classification accuracy (one-sample t-tests against 25%, computed at each time point and frequency), smoothed using a 3-point moving average. Dashed windows after cue onset in B correspond to the data shown in C. C. Bar plots of classification accuracy per condition, averaged across 8-14 Hz, 500-1000 ms after cue onset. For both cue types, quadrant was classified above chance (: p<0.05, ns: not significant).*

To examine the quadrant specificity of the analysis further, the activation in the cued quadrant and the three uncued quadrants (in the same hemifield, the opposite hemifield, and the diagonally opposite corner) was separated on retrocue and precue trials (Fig. 6).

At a basic level, the analysis revealed predicted quadrant-specific patterns during periods of visual stimulation. In other words, the activation timecourses followed the stimulus sequence. On retrocue trials (Fig. 6a), activation in the two quadrants in which stimuli had appeared in array 1 (i.e., the cued and diagonally opposed quadrants) preceded cue onsets (when stimuli were presented in these locations), while the two remaining quadrants showed activity after onset of the second array (when stimuli were presented there). Conversely, precue trials showed the opposite activation timecourse (Fig. 6b). Here, activation in response to the second array emerged in the cued and, to a lesser extent, diagonally opposed quadrants, while the remaining quadrants showed activation preceding the cue. These timecourses indicate that the analysis was

sensitive enough to detect the stimulus sequence in each trial type. Further, they suggest why retrocued quadrants could not be decoded leading up to the probe: since the stimuli in the second array led to alpha band activation in the uncued quadrants, they may have counteracted the spatial specificity of the retrocue-based activation in the cued quadrant (without necessarily reducing the more basic lateralization effect).

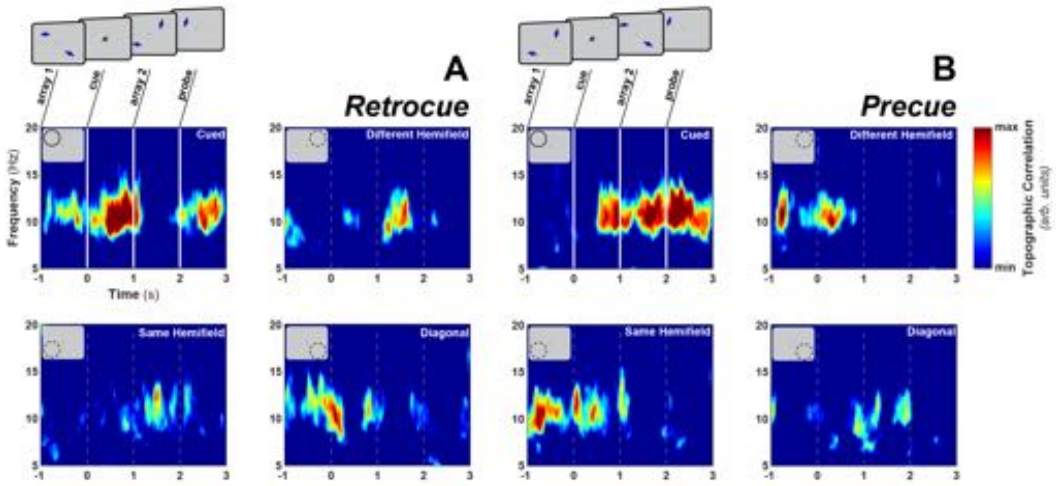


Figure 6. Quadrant-specific Activation in the Alpha Band. A. Retrocues elicit quadrant-specific shifts in alpha power. Each of the four time-frequency plots shows the correspondence of average topographies with the probe-evoked pattern, separately for each time point (from onset of array 1 to 1 s after the probe) and frequency (5-20 Hz). The top left plot shows correlations with topographies evoked by a probe in the cued quadrant, with the other three plots showing correlations with topographies evoked by probes in the three uncued quadrants (separately for probes in the cued hemifield, and in the two quadrants of the opposite hemifield). Color indicates correlation strength (in arbitrary units), with

warmer colors indicating a better fit. B. Quadrant-specific topographic patterns in response to precues.

There was again a reliable activation in the cued quadrant after probe onset and after cue onset (500-1000 ms) that was limited to the alpha band. Importantly, for the post-cue epoch (500-1000 ms), there was a main effect of quadrant ($F_{3,51}$=5.83, p=0.002), but no effect of cue type (precue/retrocue, $F_{1,17}$=0.243, p=0.63) and no interaction ($F_{3,51}$=0.216, p=0.89). This was true even when the analysis was restricted to the two quadrants within the cued hemifield (main effect of quadrant: $F_{1,17}$=10.220, p=0.005; main effect of cue type: $F_{1,17}$=0.002, p=0.97; interaction: $F_{1,17}$=0.521, p=0.48). The robust activation difference between quadrants of the same hemifield confirms that shifts in alpha power were indeed specific to a cued quadrant, rather than a cued hemifield. In addition, the analysis shows that both precues *and* retrocues led to indistinguishable quadrant-specific reductions in alpha power.

Discussion

I used a novel task in which a cue occurring within an equivalent temporal and spatial context triggers either anticipatory or retrospective shifts of attention during WM. The main finding is that retrospective attention shifts lead to a relative contralateral decrease in posterior alpha-band power. Importantly, cued alpha-lateralisation was temporarily interrupted by a second stimulus array without eliminating the retrocue-benefit on behavior. Corroborating a recent study (Wallis et al., 2015), the findings point to a mechanism of transient access to visually selective populations while selected WM information is transformed into a more robust representational state. Before discussing this mechanism, I will lay out the consequences of retrospective attention shifts on behavior and on ERP markers to constrain the interpretation of the alpha-band results.

Behavioral Impact of Cueing in Working Memory

Cues conferred behavioral benefits for both precued and retrocued items (Griffin and Nobre, 2003; Landman et al., 2003) by increasing the likelihood of recalling the correct item (also see Murray et al., 2011; 2013). The precision of memory representations was not modulated by either cue type. Prior studies using free recall of probed features have found that precues significantly improve both the rate of recall *and* precision (e.g., Bays et al., 2011; Gorgoraptis et al., 2011; Zokaei et al., 2011). The likelihood of correct recall seems to be robustly improved by both

precues and retrocues, irrespective of the mode of assessment. By contrast, the forced-choice design of the current study may not have been sensitive enough to measure cue-evoked changes in precision.

While the behavioral improvement due to precues and retrocues was similar, overall performance was significantly higher for items from the second array. Presumably, performance for array 1 was lower because of the interfering effect of array 2 (Zanto and Gazzaley, 2009). This asymmetry in difficulty may have induced a difference in motivation that could explain some of the discrepancies in the ERP results: since performance was comparatively high on pre-cue trials, participants may have used the cues less (or at least differently) and still achieved a high level of performance. Nonetheless, precues induced both accuracy benefits and costs (on infrequent invalid trials), demonstrating that observers were using them to their advantage.

ERP Markers Point to Processing Differences between Retrospective and Prospective Attention Shifts

Within the first 500 ms of cue processing, both prospective and retrospective cues led to the lateralization of potentials that have been associated with shifts of spatial attention. Retrocues had a strong effect on the EDAN (250-350 ms after the cue), and also modulated the ADAN (350-500 ms), indicating a rapid refocusing of attention (Kiss et al., 2008; Brignani et al., 2009). In addition, retrocues may also have prompted the rapid selection of the cued target from WM (Kuo et al., 2009), as indicated

by even earlier lateralizations in the N1 (150-180 ms) and N2 (240-300 ms, i.e. N2pc) ranges.

Lateralization may have occurred more quickly after retrocues than precues because retrocues allowed for immediate selection from the contents of WM. Other studies have found variability in the time course of selection from WM, with some seeing markers of access in the N2 time range (Kuo et al., 2009; Astle et al., 2010; Leszczyński et al., 2012), while others found that similar, but slightly later components responded to cues directing attention in WM (Dell'Acqua et al., 2010; Eimer and Kiss, 2010). In the current study, the time pressure to complete selection before the onset of the second array (and the interference this likely caused) may have encouraged the relatively early shifts induced by retrocues.

Similar to retrocues, precues also showed a significant lateralization in the EDAN, in line with previous findings (Harter et al., 1989; Yamaguchi et al., 1994; Hopf and Mangun, 2000; Nobre et al., 2000; van Velzen and Eimer, 2003). However, the magnitude of the EDAN was significantly larger after retrocues than precues. Precues may have shown attenuated (and delayed) lateralization in this study for a number of reasons. First, observers may have used them less, since performance was already better on precue trials (see above). Second, a precue in the current study required more operations than in typical studies of anticipatory attention, since participants first had to parse whether the cue pertained to the previous or the upcoming array, and possibly

needed to suppress information about the first array before shifting attention.

Interestingly, ERP responses to precues and retrocues began to diverge between 500 and 1000 ms after the cue, leading up to the onset of the second array. While precues evoked the typical lateralized anticipatory response (LDAP), retrocues did not, and even led to a brief but statistically significant *reversal* of this potential. The LDAP has been associated with the top-down deployment of anticipatory spatial attention, which should have improved encoding of relevant information on precue trials. On retrocue trials, none of the upcoming information was relevant (since the cued information had already been presented), and thus one may have expected no anticipatory lateralization at all. The reversal of the LDAP may indicate a shift of attention *away* from the location of the retrocued item to reduce interference from the upcoming array. The transient nature of the effect (and the relatively modest statistical significance) may alternatively indicate that the LDAP did not play a large role in guiding attention in this task, or that this late potential peaked at different timepoints in different participants, leading to a temporal smearing (and consequent reduction) of the effect. Another possibility is that this late negativity reflects contralateral delay activity (CDA), a sustained negative potential often observed in delay periods of WM tasks (Vogel and Machizawa, 2004). In the current study, this negativity emerged only late after the cue, and was not sustained. This

makes it unlikely to reflect persistent maintenance of the cued information in WM, but this interpretation cannot be ruled out entirely.

The visually evoked responses to the second memory array were also differentially modulated by precues and retrocues. Precues again led to the expected enhancement of processing, as indicated by an increase in P1 amplitude contralateral to the cued stimulus (Hillyard et al., 1998). In contrast, retrocues led to an overall reduction in P1 amplitude, and to a reduced P1 contralateral to the cued side. Similarly, the N2 also showed a main effect of cue type (although there was no strong evidence for lateralization in this interval, possibly because selection had already begun earlier). The slightly later P2 response was also smallest after retrocues, with increasing amplitudes on precue and neutral trials. This pattern points to a reduction in processing of the second array after retrocues, possibly to protect WM contents by reducing interference. The increasing amplitude of the P2 may also have reflected increased encoding demands after the second array on precue and neutral trials.

Alpha-band Lateralization May Support Different Operations After Prospective and Retrospective Attention Shifts

The main focus of the study lay on how alpha-band (8-14 Hz) oscillations respond to retrocues and precues. Importantly, I expected the alpha-band responses to provide information about the neural mechanism of prioritization during WM maintenance. After both precues and retrocues, there was a contralateral, quadrant-specific

desynchronization of posterior alpha-band power. Under both precues and retrocues, the quadrant-specific biasing of alpha power appears to reflect desynchronization over areas representing the cued quadrant, rather than increased synchronization in irrelevant areas. This conclusion can be drawn from the cross-comparison of the quadrant-specific, multivariate pattern evoked by the probe stimulus (which most likely led to a desynchronization in the alpha band). After retrocues, strong cross-generalization was limited to the interval between the cue and array 2.

On precue trials, the alpha-band lateralization coincided with other markers of increased excitability (the LDAP and an increased P1 amplitude to the subsequent array). Therefore, the pre-cue alpha-band effects likely supported sustained attention towards the anticipated location of the cued stimulus. On retrocue trials, the alpha-band effect was not compatible with sustained anticipation. The ERP response to the second, irrelevant array was attenuated (especially contralateral to the cue). Also, lateralization was not sustained throughout the trial: the appearance of array 2 eliminated the lateralization (for both cue types), and it only re-emerged immediately before probe onset. However, longer delays are needed to evaluate more rigorously this tentative conclusion.

The current set of results could help arbitrate among several theories of how prioritization acts in WM. First, the results do not support a pure inhibition mechanism: under an inhibition account, alpha-band power over irrelevant cortex would increase so that prioritized WM information is protected from interference (Sauseng et al., 2009). Several

studies have argued that alpha power over currently task-irrelevant cortex increases in order to inhibit processing of distracting stimuli that could otherwise interfere with the contents of WM (Klimesch et al., 1999; Jensen et al., 2002; Haegens et al., 2010; Bonnefond and Jensen, 2012). Retrocues could have resulted in a similar effect: once the cued item has been selected for protection, alpha power could have increased, especially contralateral to the cued side, in order to reduce interference from the second array. Although there was some evidence for inhibition of the second array, this was contradicted by the pattern of alpha lateralisation: a relative contralateral desychronization.

A second possible mechanism of WM prioritization is sustained attention (Awh et al., 2006; Chun, 2011; Chun et al., 2011). A sustained attention account would predict that retrocues improve memory by persistently increasing excitability at the anticipated probe location. This should have led to a sustained desynchronization contralateral to the cued (and eventually probed) location. However, desynchronization was only transient, such that lateralization did not immediately re-emerge after processing of the second array. Moreover, alpha lateralisation did not lead to an overall increase in excitability (as measured by the response to array 2). A recent study showing contralateral alpha-band desynchronization after a retrocue (Poch et al., 2014) could be compatible with sustained attention. However, their design and analysis did not allow a distinction between the sustained attention account and a more transient mechanism. The authors did not examine the time course of

desynchronization, and so were not able to test whether lateralization was sustained or transient. Perhaps more importantly, there were no intervening events between cue and probe onset, making it difficult to dissociate sustained attention to the cued item from anticipatory attention towards the probe stimulus. Another recent study using a similar design found that alpha lateralization was transiently and returned to baseline well before the onset of the probe (Wallis et al., 2015).

Putative Role for Spatiotopic Alpha Shifts in Item-Specific Access

A parsimonious explanation for the results in this study is that retrocues elicit short-term increases in activity in spatiotopic visual cortex. This could give control areas (such as prefrontal cortex) access to the population storing the cued visual features in order to move them into a more robust representational state (Sligte et al., 2008; Olivers et al., 2011; Lewis-Peacock and Postle, 2012; Lewis-Peacock et al., 2012; Oberauer, 2013; Chatham et al., 2014). Alpha-band desynchronization could either reflect increased local activity or communication between visual cortex and top-down control areas.

Recent studies into the mechanistic effects of alpha-band desynchronization have demonstrated that a decrease in alpha power primarily coincides with a gain in neural excitability. Alpha desynchronization has been shown to increase baseline (Bollimunta et al., 2011) and stimulus-evoked (Haegens et al., 2011) firing rates, the likelihood of perceiving two stimuli instead of one (Lange et al., 2013),

and, to some extent, the amplitude of the stimulus-evoked EEG response (P1, Rajagovindan and Ding, 2011). I speculate that, after a retrocue, the selective desynchronization in the alpha band results in a temporary spatiotopic increase in excitability for the cued item. The increased excitability presumably translates into enhanced retention for the cued feature information. This could occur by making the activity in the sensory area more resilient to decay (reducing the chances of sudden loss of feature information, Murray et al., 2013), or by involving control regions themselves in the representation of the cued feature. Alternatively, the observed alpha-band effects could be a consequence of increased local neural activity, which would lead to decreased alpha power specifically at the activated site. In either case, once the cued information has been accessed, there is no longer a need for sustained alpha lateralization, and it can return to baseline levels.

Methods capable of assessing local activity (either invasively or using gamma oscillations as a proxy) may be necessary to arbitrate between the two putative consequences of alpha lateralization outlined here. In addition, future studies will need to use longer delay periods to identify conclusively the temporal dynamics of alpha-mediated memory access when there is less time pressure from an imminent distractor array. Finally, neurodisruptive techniques (such as transcranial magnetic stimulation, TMS) might also provide evidence of the functional role of alpha oscillations in the current task. For instance, TMS may affect only a cued item, and only after alpha synchronization has shifted to prioritize it.

Zokaei and colleagues (2014) recently demonstrated that TMS to visual cortex disrupted the precision of WM items only after they had been cued. This specific disruption for preferred items might depend on a prior reduction in alpha power that sufficiently increases excitability, allowing TMS to have a larger effect. Examining the temporal dynamics of WM disruption could help arbitrate between the mechanisms outlined above.

Beyond the implications for WM, the results show that retrospective attention tasks may be used to illuminate the role of alpha oscillations in top-down prioritization separately from their role in anticipation. Even though the alpha-band lateralization and the LDAP occur in the same time window, they appear to give different results after retrocues. The discrepancy might have occurred because the two markers reflect different operations that are decoupled in retrocueing, but not in precueing. Specifically, alpha lateralization after retrocues may not reflect anticipatory spatial attention shifts. This is particularly surprising in light of findings that, in anticipatory cueing tasks, alpha lateralization is correlated with ERP markers of anticipation (such as the LDAP, see Kelly et al., 2009) and with contralateral delay activity (see Mazaheri and Jensen, 2008; van Dijk et al., 2010). Additionally, alpha lateralization usually predicts subsequent visual processing (Rajagovindan and Ding, 2011) and visual excitability (Romei et al., 2008a; 2008b).

After retrocues, alpha-band lateralization and LDAP may reflect different neural operations, for example, access to the cued part of sensory cortex (alpha-band desynchronization) vs. facilitation of sensory

processing of upcoming stimuli (LDAP). In sum, studies using retrospective attention shifts may prove fruitful in distinguishing the role of alpha oscillations in top-down attention from its role in anticipation.

Conclusion

I conclude that retrospective shifts of attention to items in working memory trigger spatiotopic access to information, reflected in early ERP markers of spatial attention shifts and lateralization of alpha-band power over visual cortex. Even though alpha-band lateralization was similar after precues and retrocues, the size and lateralization of visual responses evoked by a WM array after the cue depended on the cue type. The task design and the difference in ERPs to the second array make alternative mechanisms of retrocueing less probable: neither sustained protection via distractor inhibition, nor sustained spatial attention, could easily account for the pattern of results. Nonetheless, the proposed transient spatiotopic access may occur on its own, or in combination with a prospective shift of attention in anticipation of a probe stimulus. The access operation may help shift prioritized information into a more robust representational format (Cowan, 2000), thus increasing the likelihood of correct recall.

Chapter 4: Oscillatory Brain State Predicts Variability in Working Memory

Chapter summary[2]

Our capacity to remember and manipulate objects in working memory (WM) is severely limited. However, this capacity limitation is unlikely to be fixed, as behavioral models indicate variability from trial to trial. I investigated whether fluctuations in neural excitability at stimulus encoding, as indexed by low frequency oscillations (in the alpha band, 8-14 Hz), contribute to this variability. Specifically, I hypothesized that the spontaneous state of alpha-band activity would correlate with trial-by-trial fluctuations in visual WM. Electroencephalography (EEG) recorded from human observers during a visual WM task revealed that the prestimulus desynchronization of alpha oscillations predicts the accuracy of memory recall on a trial-by-trial basis. A model-based analysis indicated that this effect arises from a modulation in the precision of memorized items, but not the likelihood of remembering them (the recall rate). The *phase* of posterior alpha oscillations preceding the memorized item also predicted memory accuracy. Based on correlations between prestimulus alpha levels and stimulus-related visual evoked responses, I speculate that the prestimulus state of the visual system prefigures a cascade of state-dependent processes, ultimately affecting WM-guided behavior. Overall,

[2] The experiment in this chapter has been published as: Myers NE, Stokes MG, Walther L, Nobre AC (2014) Oscillatory brain state predicts variability in working memory. Journal of Neuroscience 34:7735–7743.

the results indicate that spontaneous changes in cortical excitability can have profound consequences for higher visual cognition.

Introduction

Visual working memory (WM) allows us to maintain and manipulate visual information over brief periods of time, and is critical for guiding on-going behavior. However, the storage capacity of WM is extremely limited (Cowan, 2000), either in quality (Bays et al., 2009; Keshvari et al., 2013) or quantity (i.e. 3-4 items: Zhang and Luck, 2008; Luck and Vogel, 2013) of information that can be maintained. Such capacity limits have profound implications. Individual differences in WM performance strongly predict higher cognitive functions such as IQ (Engle et al., 1999; Fukuda et al., 2010) and real-world performance (Gathercole et al., 2003).

Interestingly, WM capacity may not be a static trait even within an individual. Recent studies measuring the fidelity of representations stored in WM reveal significant trial-by-trial variability *within* observers, both in the number of encoded items (Sims et al., 2012) and in how precisely they are represented (Fougnie et al., 2012; van den Berg et al., 2012; Orhan and Jacobs, 2013). Previous modeling studies have been limited to the behavioral phenomenon of variability in visual WM, and only speculated on its neural causes.

I hypothesized that variation in the state of excitability of sensory neural systems could contribute to the variability of encoding into WM. Slow neural oscillations have been proposed to reflect dynamic changes in the state of cortical excitation (Klimesch et al., 2007; Wyart and Sergent, 2009; Jensen and Mazaheri, 2010). In particular, alpha-band oscillations (between 8 and 14 Hz) correlate with visual excitability, such

that firing rates in response to visual stimulation are reduced by neural synchronization in the alpha band (Haegens et al., 2011; Spaak et al., 2012) and at specific phases of the alpha cycle (Bollimunta et al., 2011; Saalmann et al., 2012). Correspondingly, the top-down reduction of alpha band power using attention-guiding cues has been shown to improve the detection of liminally presented stimuli (van Dijk et al., 2008; Mathewson et al., 2011), and to speed reaction times in spatial cueing tasks (Gould et al., 2011). Studies of visual perception have also shown that, in the absence of task-relevant cues, the detectability of stimuli co-varies with *spontaneous* fluctuations in the prestimulus power (Hanslmayr et al., 2007; Romei et al., 2008; VanRullen and Macdonald, 2012) and phase (Busch and VanRullen, 2010; Chakravarthi and VanRullen, 2012) of alpha oscillations. Whether such spontaneous changes in the state of visual excitability can have longer-lasting consequences for WM has not been established.

Here this idea was tested by investigating whether prestimulus alpha oscillations measured with EEG influence the encoding of items into WM. The task (Fig. 1a) was optimised for detecting oscillatory states that relate to individual items by presenting stimuli in sequence (Gorgoraptis et al., 2011). Further, a free-recall paradigm made it possible to measure subtle trial-by-trial fluctuations in the precision of memory (Zhang and Luck, 2008). This experimental approach made it possible to test whether any influences of alpha might independently modulate the precision of a WM representation, or its likelihood of being remembered.

Materials and Methods

Task Design

Seventeen healthy right-handed volunteers (9 female, mean age 22.9 years, range 20-30 years) participated in a sequential visual WM task (Fig. 1a). Each participant finished two sessions, conducted on separate days, consisting of 500 trials each. On each trial, participants saw a stream of 3-4 oriented bars (each displayed for 150 ms) appearing at random locations on an imaginary circle (6° of visual angle from a fixation dot, which subtended 0.30°), with a stimulus-onset asynchrony (SOA) uniformly jittered between 950 and1200 ms. After a memory delay following the last shape (950-1200 ms for 4-item trials, 1900-2400 ms for 3-item trials), a probe appeared at a previous stimulus location. The probe consisted of a black circle with a diameter of the same length as the bars. After participants started moving the mouse, two small tracking balls appeared, 180° apart, at a random location on the circle. Participants rotated the tracking balls to the remembered orientation at the probed location.

Stimuli

Stimuli were oriented bars with a length of 1.40° and a width of 0.32°. To emphasize the point of rotation of the stimuli, the center of the bar contained a circular disc (with a diameter of 0.68°). Some trials (30%) contained a distractor stimulus at the end of the sequence. Distractors were never probed, and participants were encouraged to ignore them.

Target shapes were distinguished from distractors by their color (isoluminant blue and green, counterbalanced across sessions).

Behavioral Analysis

For each trial, the deviation (in degrees) of the response from the actual orientation of the probed stimulus was calculated, generating a response distribution across trials. Initially, memory precision was calculated as the inverse of the circular standard deviation of the response distribution (corrected for the precision expected from randomly distributed responses, see 3), separately for each condition (i.e. the 10 combinations of memory load (3 or 4), distractor presence, and position of the probed item in the sequence (1st, 2nd, 3rd, or 4th)). To get a more revealing measure of memory, observers' responses were modelled with a mixture of two distributions:

$$p(x \mid \theta) = \frac{P_{MEM}}{2\pi I_0(\kappa)} e^{\kappa \cos(x-\theta)} + \frac{1 - P_{MEM}}{2\pi},$$

where x is the response angle, θ is the presented angle, and I_0 is the modified Bessel function (of order 0). The two variable parameters, P_{MEM} and κ, define the two distributions, respectively: the probability of recalling an item (P_{MEM}), and the concentration parameter, or precision, of remembered items (κ). This model was fit using an expectation-maximization algorithm (Bays et al., 2009), separately for each condition. In an alternative analysis, the probability of misreporting an item was also estimated, but it had no effect on the results reported here. Model

parameters for different conditions were compared using repeated-measures ANOVA. A two-way 3x3 ANOVA with the factors item position (first, second, or third item, ignoring the fourth item on load 4 trials) and load (3 items, 3 items with distractor, or 4 items) was used to test for the effect of item position across load conditions.

EEG Acquisition

The EEG was recorded continuously using NuAmp amplifiers (Neuroscan Inc., Albany, USA) from 40 Ag/AgCl electrodes positioned according to the 10–20 international system (Sharbrough et al., 1991). Recordings were taken from electrodes: Fz, FCz, Cz, CPz, Pz, POz, Oz, FP1/2, F3/4, F7/8, FT7/8, T7/8, TP7/8, FC3/4, C3/4, CP3/4,P3/4, P7/8, PO3/4, PO7/8, and O1/2. Blinks and eye movements were monitored by deriving bipolar recordings from an electrode placed below the right eye and FP2 (VEOG) and from electrodes placed to the left and right of the right eye (HEOG).

The electrode placed at AFz served as the ground, and the right mastoid served as the active reference. Electrical impedances were kept below 5 kΩ and activity was filtered with a low-pass filter of 300 Hz. The analogue-to-digital sampling rate of brain activity was set at 1000 Hz and data were recorded continuously for the entire experiment.

EEG Processing

All further processing and analyses were done offline. Firstly, the

EEG was re-referenced to the average of the left and right mastoids. The two bipolar electrooculogram (EOG) signals were derived by calculating the difference between the upper and lower VEOG electrodes and between the left and right HEOG electrodes. Data were down-sampled to 250 Hz and bandpass filtered between 0.10 and 80 Hz. The filtered data were cut into trial-length epochs encompassing the stimulus presentation, memory delay, and probe presentation periods (-1000 ms to 5500 ms relative to trial onset). Slow linear, quadratic, and cubic drifts in the epochs were removed using a 3rd-order polynomial fit, performed separately for each channel and epoch. Control analyses on data that were neither filtered nor detrended yielded essentially the same results.

After rejecting epochs and channels with non-stereotyped artifacts (such as saturated channels), data were decomposed by independent-component analysis (ICA), as implemented in the EEGLAB toolbox (Delorme and Makeig, 2004) to identify components related to eye blinks and muscular artefacts, which were not included in the remixing of the data. After the ICA stage, trials with eye blinks (as identified in the EOG channels, which were not included in the ICA) and artifacts in a ±100-ms window around any stimulus presentation were also discarded after visual inspection. The manual artifact removal step was added to remove trials in which eyeblinks may have precluded observers from seeing the stimulus (which is not addressed in the ICA step).

ERP Analyses

To test for differences in the evoked responses to items as a function of task relevance (targets vs. distractors), the average evoked response potentials were calculated for each stimulus category. Data were temporally smoothed with a Gaussian kernel (24 ms standard deviation) and the average potential in the prestimulus baseline period (the 200 ms before a stimulus onset) was subtracted before calculating subject-specific averages. Paired samples t-tests (for the comparison between targets and distractors in the fourth position) were then used to calculate group-level statistics. These results were corrected for multiple comparisons across time and sensors using 2D cluster-based permutation testing, with a cluster-forming threshold (and a cluster mass significance threshold) of $p = 0.05$.

Correlation between Alpha Power and Accuracy

To estimate trial-by-trial prestimulus power, power from 5 to 20 Hz (in 0.5 Hz steps) was estimated at each electrode using Morlet wavelets (with a length of 5 cycles) using Matlab and the FieldTrip Analysis Toolbox (Oostenveld et al., 2011) and log-transformed ($10*\log_{10}$). Power timeseries were then averaged across contralateral sensors (relative to the side of the stimulus, averaged across electrodes O1/2, PO3/4/7/8, and P3/4/7/8) at each time point around the onset time of the stimulus that would be probed on that trial (-500 ms to +100 ms).

For each frequency and each timepoint around the stimulus, a general linear model (GLM) was fit to the z-transformed contralateral

alpha power, using the z-transformed response accuracy as a regressor. Specifically, the measure of response accuracy was the negative absolute distance, in degrees, between the response angle and the tested stimulus angle, $-|x-\theta|$, where x is the response angle, and θ is the stimulus angle. Larger values (those closer to 0) therefore indicated more accurate performance. The potentially confounding effects of sequence length (3 or 4 items), distractor presence, and sequence position of the tested item were accounted for by fitting models for each of the 10 possible conditions separately, and then averaging regression weights across conditions. To account for potential session effects on power (i.e. linear increases or decreases in power over the course of the experiment), trial number and session number were added as nuisance regressors in the GLM. Although trial number correlated positively with alpha power throughout the peristimulus interval, this had no effect on the correlation between contralateral alpha power and accuracy. Moreover, there was no significant decrease in memory accuracy over the course of an experimental session, indicating that it had no or little additional impact on behavior.

At the group level, t-tests were used across the weights estimated in the GLM to test for a significant relationship between alpha lateralization and response accuracy at each frequency and each timepoint in the trial. Cluster-based permutation testing was applied to correct for multiple comparisons (across frequencies and time points,

with 5000 permutations, a cluster-forming threshold of 0.05, and a significance threshold of 0.05, see Maris and Oostenveld, 2007).

The topography of the correlation in the prestimulus period (500 to 100 ms before stimulus onset) was derived by calculating trial-by-trial alpha power (as above), separately for each sensor. After averaging across frequencies (8-14 Hz in 0.5 Hz bins) and flipping topographies along the midline on trials where the stimulus was presented in the right hemifield (to visualize contralateral vs. ipsilateral correlations), the regression analysis was repeated separately at each sensor. T-tests at each sensor and timepoint were entered into group-level analysis.

Correlation between Alpha Phase and Accuracy

The same time-frequency decomposition (Morlet wavelets from 5 to 20 Hz in 0.5 Hz steps) was used to estimate instantaneous phase at contralateral parieto-occipital sensors (PO3/4/7/8, P3/4). The circular average phase (across sensors) was then entered into a circular-linear correlation with memory accuracy (i.e., $-|x-\theta|$), resulting in a timecourse of correlation values for each frequency. The circular-linear correlation between alpha phase and 100 random permutations of memory accuracy (permuting across trials) was then calculated, and the z-transformed rank of the real correlation coefficient in the permutation distribution was determined (via the inverse normal cumulative distribution function). One-sample t-tests were used to compare this z-score timecourse at the group level. Again, cluster-based permutation tests corrected for multiple

comparisons. In addition, the same approach was applied at each sensor to calculate the topography of the peak of the phase-accuracy correlation in the prestimulus time window (-350 to -150 ms). A standard topographical difference analysis (TANOVA, Murray et al., 2008) was used to compare the resulting topography with the topography of the alpha power effect. In brief, the difference between two topographies is the square root of the sum of squared differences at each sensor (after scaling each sensor by the variance across all sensors, making it equivalent to the normalized Euclidean distance between topographies). This difference is compared to a permutation distribution to assess significance.

Modeling the Effect of Alpha Power on Precision and Guess Rate

Mathematical models of fluctuations in WM accuracy predict that fluctuations in alpha power should primarily affect the precision of memory, rather than the probability of recall (van den Berg et al., 2012). In order to test this hypothesis, alpha power was correlated separately with trial-by-trial fluctuations in precision or in guess rate. First, the probability of recalling an item (P_{MEM}), and the precision of remembered items (κ), was estimated (with maximum-likelihood estimation, MLE) using the mixture model described above:

$$p(x \mid \theta) = \frac{P_{MEM}}{2\pi I_0(\kappa)} e^{\kappa \cos(x-\theta)} + \frac{1 - P_{MEM}}{2\pi}.$$

After estimating these two parameters, I tested extended models that take trial-to-trial variations in alpha power into account. In the first model, alpha power fluctuations were fit to changes in precision

$$p(x \mid \theta, \alpha_{POWER}) = \frac{P_{MEM}}{2\pi I_0(\kappa(1+\beta^* \alpha_{POWER}))} e^{\kappa(1+\beta^* \alpha_{POWER})\cos(x-\theta)} + \frac{1-P_{MEM}}{2\pi}$$

where α_{POWER} is trial-by-trial and timepoint-by-timepoint posterior alpha power (averaged over O1/2, PO3/4/7/8, and P3/4/7/8), and β is the influence of alpha power on precision (estimated using MLE).

In an alternative model, alpha power was fit to changes in recall rate:

$$p(x \mid \theta, \alpha_{POWER}) = \frac{P_{MEM}(1+\beta^* \alpha_{POWER})}{2\pi I_0(\kappa)} e^{\kappa \cos(x-\theta)} + \frac{1-P_{MEM}(1+\beta^* \alpha_{POWER})}{2\pi}.$$

The resulting parameter-estimate timecourses were then tested at the group level using one-sample t-tests (under the null hypothesis that the average β is 0 if alpha power has no effect). Additionally, this model was fit to prestimulus alpha power at each sensor to obtain topographies of the effect.

This novel method was also validated using an alternative, simpler approach. Here, prestimulus alpha power was sorted into two bins (separately for each behavioral condition), and the mixture model (see above) was fit separately to high-alpha power and low-alpha power trials. Precision and recall rates between high- and low-alpha power were then compared using paired t-tests.

Link between Visual Evoked Responses, Prestimulus Alpha Power, and Memory

If, as I argue, prestimulus alpha power is an index of cortical excitability, it should predict the magnitude of visual responses to the memory stimulus. To test this hypothesis, the broadband EEG signal (bandpass-filtered between 1 and 40 Hz and baseline-corrected relative to 500 to 100 ms before stimulus onset, the time range of the prestimulus alpha power) was used as the dependent variable, and was correlated with the trial-by-trial average of prestimulus contralateral posterior alpha power (using the same sensors and time window as above, i.e. 500-100 ms before stimulus onset). In addition, memory accuracy, $-|x-\theta|$, was added as a second regressor (and, as above, trial number was added as a nuisance regressor). The rationale for this analysis was the following: if alpha power modulates stimulus processing in order to improve performance, then both alpha power and accuracy should correlate with comparable epochs and sensors in the evoked signal.

In order to eliminate any possibly confounding effects of alpha power on prestimulus baseline potential, I conducted a control analysis in which I added each sensor's average prestimulus voltage as a nuisance regressor to the model. This did not noticeably affect the regression weights of alpha power and accuracy. To test the similarity of the correlation topographies with the ERP topography, the difference (TANOVA) between the mean ERP and the ERP-to-behavior and ERP-to-alpha power correlations was calculated.

Results

Behavioral Results

Observers completed a sequential working-memory precision task. After encoding 3 or 4 uniquely positioned peripheral items (see *Materials and Methods* for details) and maintaining them for a variable delay, a peripheral probe matching the location of one of the items prompted them to recall that item's orientation with a manual response. As expected, observers recalled more recently presented items with higher accuracy (main effect of item position in the trial sequence, $F_{2,32}$=37.4, p<0.001; see also Gorgoraptis et al., 2011). A mixture model analysis quantified separately the concentration of the Gaussian component (kappa, or κ), and recall rate (P_{MEM}, the average likelihood of recalling any information about a probed stimulus; see *Materials and Methods* and Fig. 1d-e). More recently presented items were remembered with higher precision ($F_{2,32}$=7.08, p=0.003) and with a higher recall rate ($F_{2,32}$=7.89, p=0.002). Therefore, in the subsequent regression analyses, correlations between alpha oscillations and memory accuracy were calculated separately for each item position and then regression weights were averaged across conditions. There was also a main effect of load on accuracy ($F_{2,32}$=19.8, p<0.001), indicating that performance decreased when a distractor or a fourth target was presented. This effect was strongest when probing the third item in the sequence ($F_{2,32}$=20.16, p<0.001), leading to a significant interaction between load and item position ($F_{2,32}$=3.75, p<0.008). The mixture-model analysis showed that,

while the presentation of a fourth target decreased both the precision (t_{16}=4.45, p<0.001) and the recall rate for the third item (t_{16}=3.76, p=0.002), presentation of a *distractor* only had a significant effect on precision (t_{16}=3.01, p=0.008), but not on recall rate (t_{16}=1.37, p=0.19).

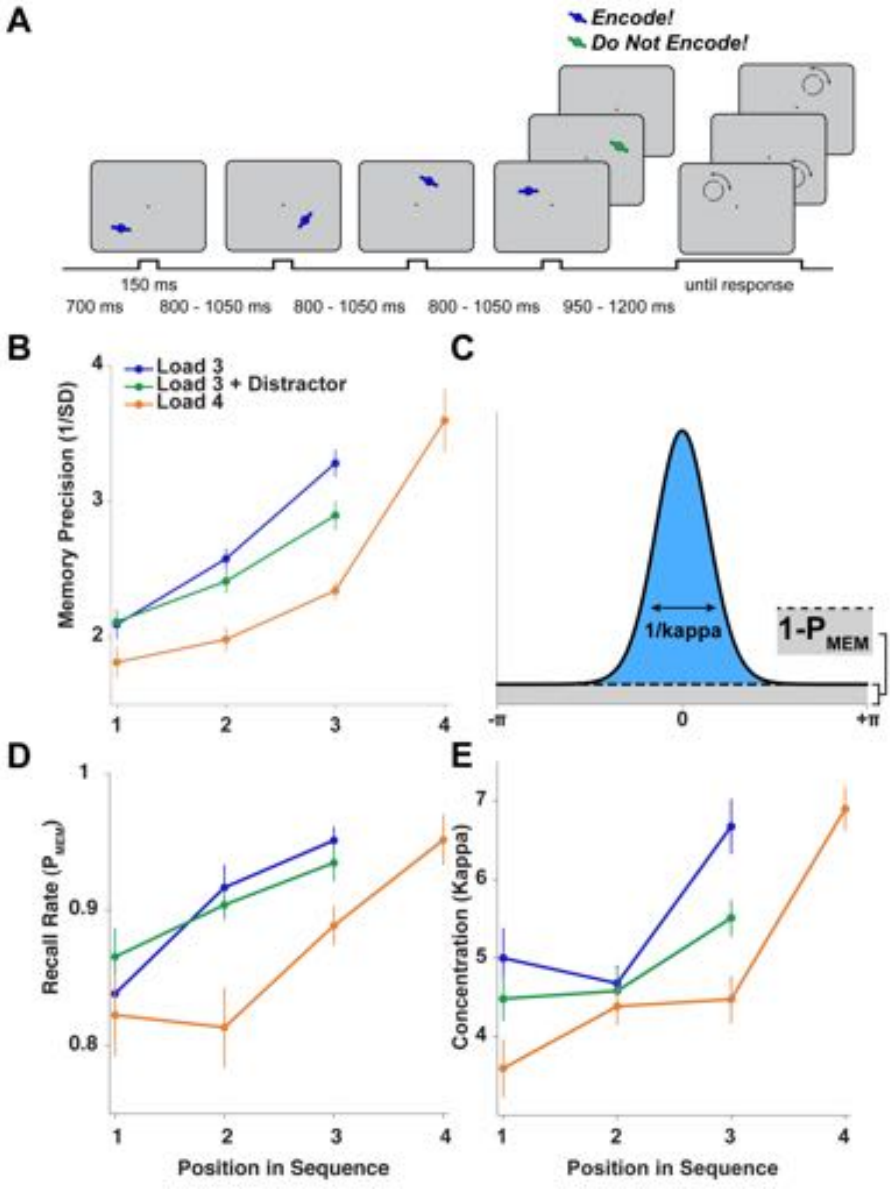


Figure 1. Design and behavioral results of the sequential WM task. A. Observers saw a temporally jittered sequence of at least 3 oriented bars presented at different locations on the screen. On 70% of trials, a fourth

bar appeared and could be either task-relevant (blue, bottom row, probed in 25% of 4-target trials) or task-irrelevant (green, middle row, never probed). Colors were counter-balanced across sessions. After a variable delay, a probe circle appeared at the location of one memory item. Observers indicated that item's orientation by moving indicator dots. After the response, the stimulus reappeared over the indicator dots as feedback (not shown here). B. The precision of the response distribution (i.e., the inverse of the circular standard deviation) showed a robust recency effect for each condition. In addition, participants performed better on load 3 trials than on trials with a distractor or with a fourth relevant item. C. Schematic of response distribution consisting of a mixture of two underlying components: a uniform distribution driven by random guesses, with height 1-P_{MEM} (gray shading), and a von Mises distribution driven by noisy recall, with concentration kappa (blue shading). D and E. Results of mixture-model fitting. Response distributions were fit separately for each participant and condition, estimating the concentration (kappa) of the von Mises distribution and the recall rate (P_{MEM}) for each. D. Recall rates increased for items presented later in the trial, and were higher on Load 3 and Load 3+D trials relative to Load 4. E. Kappa also increased as a function of item position in the trial sequence, and decreased as a function of load, with Load 3+D and Load 4 showing lower precision than Load 3. Error bars denote the within-subject standard error.

Prestimulus Alpha Power Predicts Precision of WM

Spontaneous 8-14 Hz power before stimulus onset (at contralateral occipital-parietal sensors) predicted memory accuracy for the upcoming stimulus (Fig. 2a, two-tailed t-test, cluster-corrected p=0.008). The negative correlation indicates that memory was better when alpha power was lower, i.e., when alpha was more desynchronized contralateral to the target. The correlation with accuracy was restricted to the alpha band and to the 500 ms preceding stimulus onset (Fig. 2a).

To evaluate the precise consequence of spontaneous alpha fluctuations on encoding, a model-based approach was used to test whether decreased alpha power predicted increases in the precision (kappa) or the recall rate (P_{MEM}) of WM. Alpha power fluctuations affected only the precision (cluster $p < 0.05$, again restricted to parietal and occipital sensors in the prestimulus epoch, see Fig. 3a-b), but not the recall rate. A median split on prestimulus alpha power (8-14 Hz, 500-100 ms before stimulus onset) confirmed significant differences in precision (t_{16}=2.15, p=0.024) but not in recall rate (t_{16}=-0.34, p>0.7, Fig. 3c-d).

Further, if higher excitability improves encoding, spontaneous *decreases* in excitability (as indexed by an increase in alpha power) might facilitate the inhibition of irrelevant distractors, thereby improving memory for previously encoded items. To examine the effect on irrelevant input, a differently colored task-irrelevant distractor appeared in the fourth position on a minority of trials (30%). On these distractor trials, pre-distractor alpha power was correlated with accuracy for one of the preceding items. Initially, the same analysis approach was used as before

targets, i.e. correlating accuracy with contralateral power. This did not result in a significant effect (cluster $p > 0.25$). Given the relatively small number of trials for this analysis and the indirect behavioral measure, the similar (but possibly more sensitive) approach of correlating the prestimulus *lateralization* (contralateral minus ipsilateral power) was also tested. The alpha power lateralization before onset of the distractor predicted accuracy for the preceding items (Fig. 4, cluster-corrected $p=0.023$, left panel). Here, there was a *positive* correlation, indicating that alpha lateralization (*higher* power in contralateral relative to ipsilateral sensors) before a to-be-ignored distractor improves memory for the items already encoded into WM. There was no such effect for task-relevant unprobed items in the fourth position (Fig. 4, right panel). On average, the correlation before distractors (500 to 100 ms before stimulus onset) was higher than that before unprobed relevant items ($t_{16}=2.567$, $p=0.021$). Because high alpha before relevant items appeared to have no strong impact on memory of preceding items, active inhibition may be needed at the point of encoding for a retrospective effect of alpha power to occur.

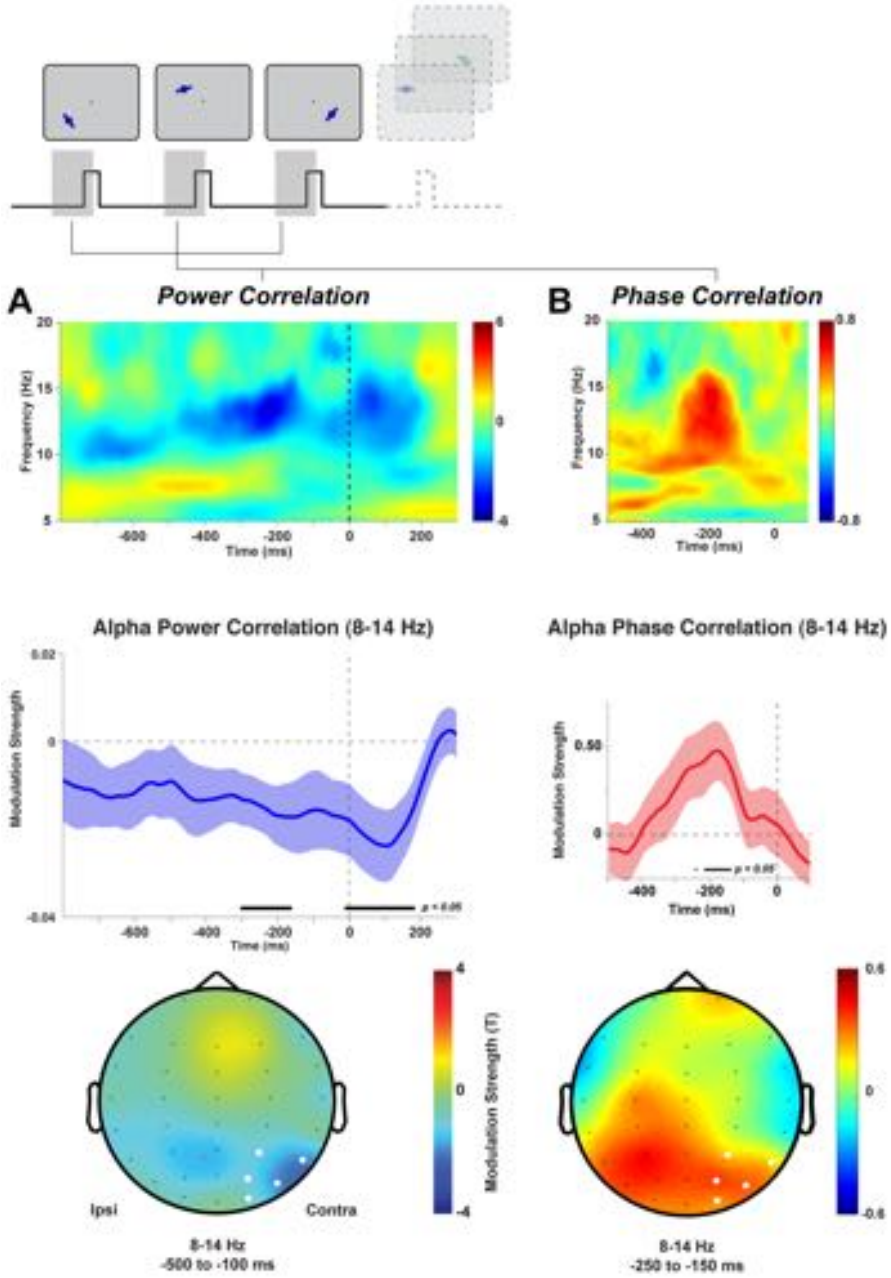


Figure 2. Posterior alpha oscillations predict WM accuracy. A. Lateralized posterior alpha power correlates with recall accuracy. Central left: Alpha power lateralization prior to presentation of the probed stimulus correlates negatively with memory accuracy. Black bar indicates significant differences from 0 (t-tests on regression weights from GLM, p<0.05). Blue shading indicates s.e.m. across participants. Top Left: Time-frequency representation of the strength of correlation (t-test on regression weights),

showing a significant cluster in the alpha band preceding stimulus onset. Bottom left: topographical distribution of regression weights preceding stimulus onset. White sensors were used for analyses in the two panels above. B. Posterior alpha phase preceding stimulus onset predicts recall accuracy. The circular-linear correlation between the instantaneous alpha phase (8-14 Hz, top right) on parietal sensors and response error increased 250-150 ms before the onset of the probed stimulus (red line). Shading indicates s.e.m. across participants. Black bar indicates a significant correlation ($p<0.05$, permutation test). This prestimulus cluster is limited to the alpha band (top right). The plots show the z-scored circular-linear correlation coefficient (normalized relative to permutation distribution at each time point and frequency). Bottom right: Prestimulus increases in alpha phase correlations are strongest at parietal sensors (color indicates z-scored circular-linear correlation coefficient).

Prestimulus Alpha Phase Correlates with Behavior

In addition, occipital-parietal alpha phase preceding the probed stimulus onset correlated significantly with memory performance (Fig. 2b, two-tailed t-test, cluster-corrected $p=0.035$). Again, the effect peaked approximately 200 ms prior to stimulus onset. Similar to the alpha-power effects, the topography of alpha-phase effects was strongest at posterior sensors (Fig. 2b), but more broadly distributed. That said, there was no significant difference between the average topographies (500 to 100 ms prestimulus: $D=0.519$, $p=0.965$; 250 to 150 ms prestimulus: $D=0.569$,

p=0.935). Nevertheless, the apparently broader distribution compared to alpha power may result from alpha phase having a stronger spatial consistency across posterior sensors, leading to a spurious extension of the effect across the topography. To ensure that alpha phase was spontaneous, and not entrained by the preceding stimulation, phase-locking values in the prestimulus period were calculated. There was no significant phase locking in the alpha-band in the tested epochs (cluster $p>0.84$), confirming that the prestimulus alpha phase was randomly distributed. In contrast to the alpha power effect before the onset of a task-irrelevant distractor, there was no relationship between alpha *phase* and accuracy before distractor onset ($p>0.40$ on parietal-occipital electrodes).

It is possible that targets presented in the fourth position were processed differently from items presented earlier (since observers may have actively suppressed processing of the fourth item in anticipation of a possible distractor). When the trials probing the fourth item in the sequence were omitted from the analyses, essentially the same pattern of correlations between memory accuracy and contralateral alpha power (cluster p=0.018) and alpha phase (cluster p=0.061) could be observed.

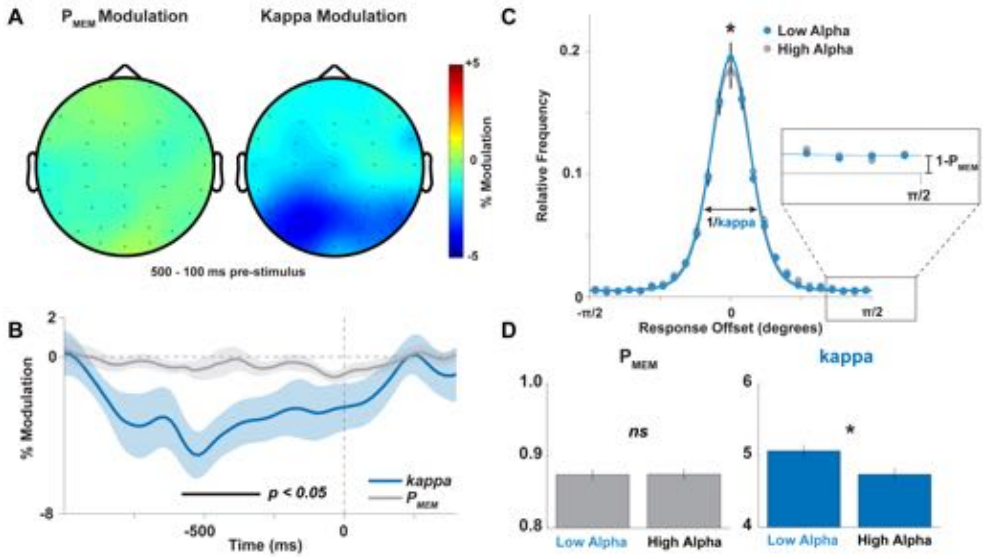


Figure 3. Prestimulus alpha power predicts changes in memory precision, but not in guess rate. Using a model-based analysis, the parametric influence of alpha power on precision and guess rate in the memory mixture model was estimated. While there was no significant effect on guess rate, there were no significant effects (cluster-corrected $p < 0.05$) of alpha power on precision. This effect peaked over posterior sensors A, approximately 250 to 500 ms before stimulus onset (B, significant differences between precision modulation and guess rate modulation indicated by the black bar). C. Histogram of responses, split into high- and low-alpha trials (binned separately for each condition and then merged). The only significant difference between high- and low-alpha trials was found in the height of the histogram peak, with low alpha trials producing more high-precision responses ($p<0.05$). D. This effect was reflected in a significant difference between precision on high- and low-alpha trials ($p=0.024$). Guess rate did not differ between high and low alpha ($p>0.70$).

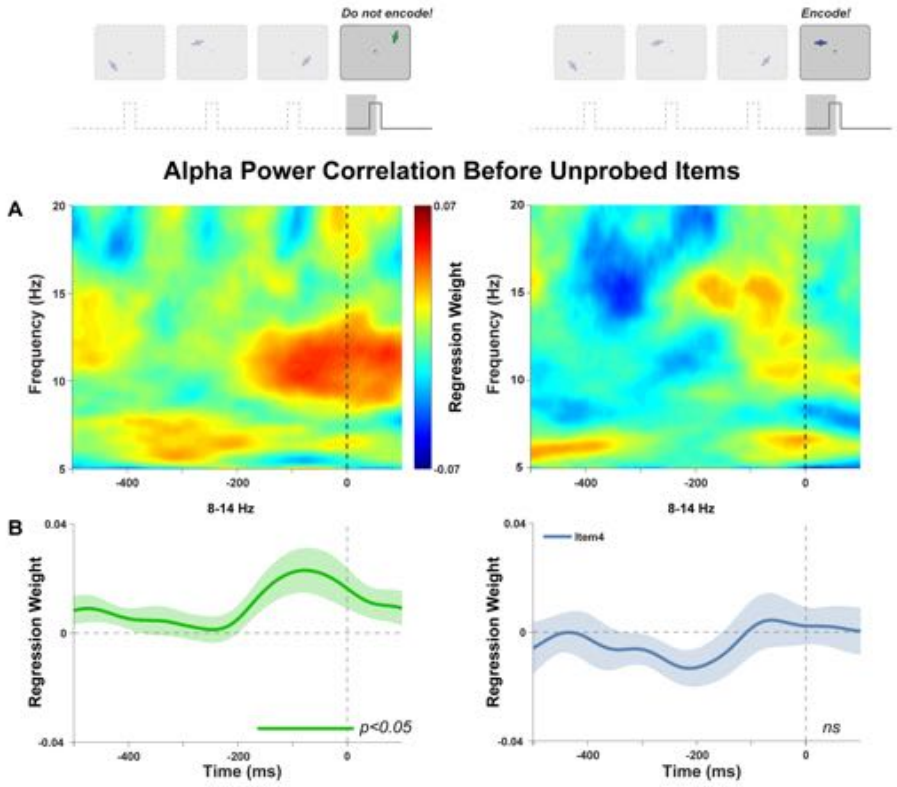


Figure 4. Alpha power lateralization before an irrelevant distractor predicts memory accuracy for preceding items. A. Posterior alpha lateralization before onset of a distractor at position 4 positively correlates with response errors to items 1, 2, or 3. This effect was again restricted to the alpha range (cluster corrected p=0.02). There was no correlation with response errors to items 1, 2, or 3 when a task-relevant item was presented at position 4 (right panels, cluster p > 0.20). Note that this is distinct from the correlation before the probed item shown in figure 2, which was specific to the upcoming, rather than preceding items. B. Significant correlations with accuracy in the alpha band (8-14 Hz) began approximately 200 ms before stimulus onset.

Prestimulus Alpha Power and Memory Accuracy Correlate with Parietal Evoked Potential Amplitude

A final analysis examined how visual evoked responses might form a mechanistic link between changes in cortical state (i.e., alpha power) and memory. Multiple regression was used to correlate the trial-by-trial stimulus-evoked EEG with prestimulus alpha power and memory accuracy. Both alpha power and accuracy showed a peak in correlation over posterior sensors, approximately 200 ms after stimulus onset (see Fig. 5, middle panel). This peak corresponds in time with the lateral posterior N1 potential (see Fig. 5, top panel). However, in contrast to the lateral-posterior topography of the N1, the correlation with accuracy was strongest over central parietal sensors. In line with apparent topographical discrepancy, statistical comparisons (via TANOVA) confirmed that the ERP topography (200-250 ms post-stimulus) differed from the topographies of the ERP-accuracy correlation ($D=0.725$, $p=0.053$) and the ERP-alpha power correlation ($D=0.910$, $p=0.004$). Effects of alpha power and accuracy on the ERP, therefore, may have shared only a subset of the sources active in this time window. The correlation with alpha power was positive, indicating that decreases in alpha power increased the amplitude of the (negative-going) ERP. Conversely, the correlation with accuracy was *negative*, indicating that for high-accuracy trials, the ERP was more negative. Both correlations peaked over parieto-occipital sensors (see Fig. 5, bottom panel), although the peak of the accuracy correlation was more central, whereas alpha

power also correlated with postero-lateral sensors.

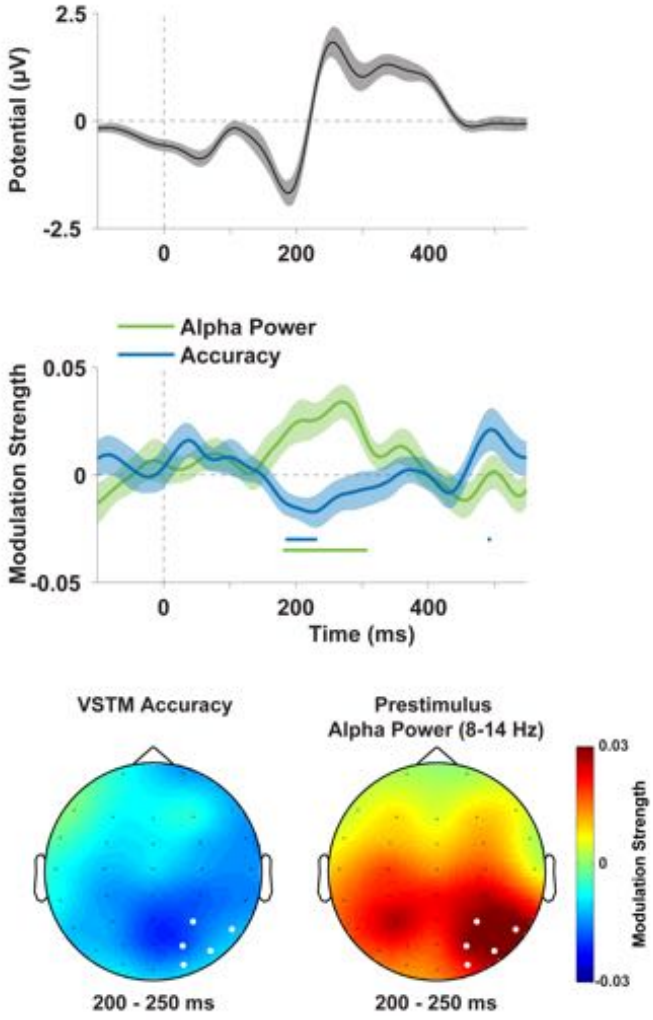


Figure 5. Prestimulus alpha power and WM accuracy correlate with posterior visual ERPs. Potentials on sensors contralateral to the stimulus (highlighted in white on the topographies, bottom panel) correlated significantly with posterior contralateral 8-14 Hz power (calculated at the same sensors, 500 to 100 ms before stimulus onset) and with memory accuracy, $-|x-\theta|$, 200 to 250 ms after stimulus onset (see regression timecourses in middle panel). The time range (shaded blue and green

boxes in top panel) when both variables correlated significantly with broadband EEG corresponds approximately to the time of the negative-going N1 potential. This indicates that higher alpha power, which showed a positive correlation (green line in middle panel), tended to decrease the (negative) amplitude of the EEG, whereas accuracy (blue line in middle panel) was higher when the ERP was more negative (as indicated by a negative correlation). Both correlations peaked over posterior parietal and occipital sensors (see bottom panel). Shaded errors indicate s.e.m.

Broadband Effects of Distractor Processing

Since prestimulus alpha power preceding a distractor predicted memory accuracy for previous items, while alpha preceding a relevant but unprobed fourth target item did not, it seemed reasonable to assume that these two types of stimuli were indeed processed differently by the visual system (see *Materials and Methods* above). There were early frontal differences between targets and distractors (cluster-corrected $p<0.05$ for two clusters), followed closely by a relatively early dissociation in the posterior visual evoked response (Fig. 6), with significant differences in the topographies after 200 ms. These differences may account for the different behavioral effects of targets and distractors (see Fig. 1b).

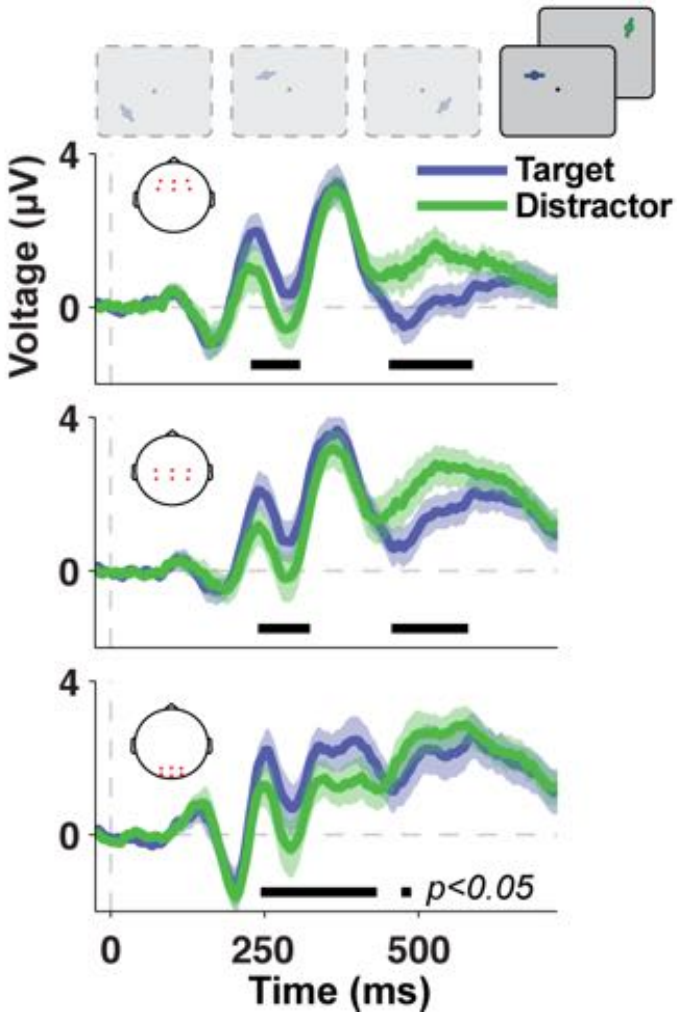


Figure 6. Broadband responses to task-relevant and irrelevant stimulus onsets. Shown here are average ERPs for frontal, central, and posterior electrodes (electrode locations shown in insets), separately for targets presented in position 4 (blue), and for distractors in position 4 (green). A comparison of targets with distractors showed significant differences in three clusters (p<0.05). Targets led to an enhanced P2 (200-275 ms) and P3 (300-425 ms) response, and an attenuation of the sustained response around 500 ms (450-600 ms). Shaded areas around the ERP timecourses represent within-subject s.e.m. Black bars indicate timepoints when target and distractor-evoked potentials are significantly different (p<0.05, uncorrected).

Discussion

The current results indicate that spontaneous changes in visual excitability, measured by oscillations in the alpha band, contribute to the representational precision of WM encoding. This effect was present even though stimuli were presented in isolation, and well above contrast threshold, suggesting that alpha-band synchronization can bias encoding even when stimuli are easily detectable and not directly competing for attention.

A model-based analysis approach indicated that posterior alpha power predominantly influences the precision of memory for an item, rather than the likelihood that it will be robustly encoded in WM. This finding fits predictions of recent mathematical models of WM, which have argued that stochastic variability in neural gain influences the neural precision of encoded information (Ma et al., 2006; van den Berg et al., 2012). One alternative model (Sims et al., 2012) proposed that the variability in WM can be explained by variability in how many items are encoded from a multi-item display (while encoding precision is constant). Research into the separate neural contributions to limits in quality and quantity of WM is only beginning to emerge (e.g., Anderson et al., 2011; Machizawa et al., 2012; Emrich et al., 2013). While this study favors the interpretation that spontaneous alpha fluctuations influence precision, it does not contradict models emphasizing variability in the quantity of encoded items. Other state variables, such as excitability in prefrontal cortex or connectivity between prefrontal and visual areas (Curtis and

D'Esposito, 2003; Gazzaley et al., 2004; Zanto et al., 2011), could influence *how many* items are encoded into WM on a trial-to-trial basis. Similarly, an influence on quantity might depend on perceptual competition between items. Here, low-level inter-item competition was explicitly avoided by presenting stimuli in sequence. This design made it possible to show that the prestimulus state has long-lasting consequences for memory even when perceptual constraints are minimized. It is possible that prestimulus alpha power could also influence the probability of encoding specific items in a design using simultaneous presentation of multiple items.

Prior to a distractor, lateralization of alpha power in the opposite direction (i.e. a relative *increase* in contralateral, compared to ipsilateral, alpha power) improved memory for a previously presented item. There was no such effect before an unprobed task-relevant item, possibly indicating that the correlation is specific to items that are actively inhibited during visual processing (as may be borne out in processing differences between targets and distractors, see Fig. 6). In addition, the item-specificity of the correlation may indicate that different items are encoded (and possibly maintained) independently (see also Chapter 2). However, since the difference between distractors and target items was modest, future studies could employ a more powerful design to test for differential effects preceding task-relevant or irrelevant items. This result complements a recent finding that alpha power prior to an expected, centrally presented distractor predicts the speed of recall from WM

(Bonnefond and Jensen, 2012). In that study, distractors appeared on every trial with predictable timings, and distractor strength was manipulated between blocks. The predictability of the distractor item presumably allowed for precise top-down control over alpha synchronization by temporal expectations (Rohenkohl and Nobre, 2011). Moreover, because distractors were centrally presented, the effects on alpha were distributed across central posterior sensors. By contrast, in the present study overall alpha power did not significantly predict memory performance, whereas the more spatially specific *lateralization* did significantly covary with accuracy of the task-relevant items. Although the results are broadly consistent with Bonnefond and Jensen, differences in the predictability of the distractor item presumably account for differences in the precise pattern of prestimulus effects.

For task-relevant items that were probed, there was also a correlation between prestimulus alpha phase and WM performance. This finding may seem counterintuitive at first glance: since the presentation time of each stimulus (150 ms) was longer than one alpha cycle (~100 ms), there should have been sufficient opportunity for stimulus processing to occur at the optimal alpha phase. However, alpha phase preceding stimulus onset may affect the number of alpha duty cycles occurring during stimulus presentation. More phases of high excitability should therefore affect the quality of encoding (see Hanslmayr et al., 2013, for a similar argument). Alternatively (or additionally), in a dynamical system such as the brain, small shifts in initial conditions (caused by

variability in alpha phase) may have long-lasting effects on the neural trajectory of stimulus processing. Because stimulus onsets were temporally jittered, observers were unable to predict the exact onset time of the next target. It is therefore unlikely that they were able to use anticipatory top-down control to entrain alpha oscillations to a behaviorally optimal phase. This strengthens the conclusion that stochastic fluctuations in alpha (phase and power) were directly responsible for WM improvements. The lack of strategic control over alpha phase may also explain why phase before a distractor had no relationship with accuracy (in contrast with Bonnefond and Jensen, 2012), where the temporal predictability of the distractor onset was thought to have played an important role.

Amplification of visual evoked potentials may be a mechanism by which prestimulus alpha influences memory. In accordance with previous studies (Brandt and Jansen, 1991; Brandt et al., 1991), prestimulus alpha power correlated with the amplitude of evoked responses approximately 200 ms after stimulus onset (for effects on earlier components, see Rajagovindan and Ding, 2011). Around the same time, ERP amplitude also correlated with memory accuracy, indicating that the modulation of visual ERPs could mediate between prestimulus state and ultimate memory performance. It should be noted that the topography of the correlation with accuracy had a more central-parietal peak than the postero-lateral correlations with alpha power. Although the coarse spatial resolution of EEG precludes any strong conclusions from this shift, it does

point to the possibility that the link between alpha fluctuations and behavior may depend on facilitating a cascade of multiple processing stages occurring in different areas and at different timepoints during encoding. An intriguing possibility is that this initial state influences the latency of certain processing steps during encoding, leading to a more efficient trajectory through neural state space (Harvey et al., 2012) to reach a stable memory state (Stokes et al., 2013). The correlation between memory accuracy and the visual ERP spanned both the peak of the N1 and the onset of the P2, making it impossible to disambiguate the exact point of influence. Subsequent studies are needed to identify the specific influence of the prestimulus state on the processing cascade of WM encoding.

The current study differs from many related experiments in that it examined the effect of spontaneous electrophysiological dynamics on behavior, rather than the effect of top-down attentional cues on behavior and electrophysiology. Such cues can have enduring consequences for WM (Gazzaley and Nobre, 2011; Nobre and Stokes, 2011): Attention during encoding improves memory (Rutman et al., 2010; Gazzaley, 2011), and has been linked to electrophysiological markers of endogenous attention shifts (Murray et al., 2011). Possibly, top-down attention constrains the variability in firing rate (Churchland et al., 2010; Cohen and Maunsell, 2010; Churchland et al., 2011), leading to increased encoding precision. Future studies could test this proposal through a direct comparison of spontaneous and cue-evoked variability in the alpha state

and how this relates to stimulus-evoked response variability. The ability to control spontaneous fluctuations in neural excitability could be linked to inter-individual variability in WM capacity (Vogel and Machizawa, 2004; Vogel et al., 2005; Sauseng et al., 2009). Finally, single-unit electrophysiology or magnetoencephalography studies will need to test whether alpha oscillations indeed act on WM by increasing firing rates or high-frequency (gamma-band) evoked power in visual cortex. This could help formulate a stronger mechanistic link between spontaneous state changes and task performance.

Chapter 5: Testing Sensory Evidence Against Mnemonic Templates

Chapter summary

Most perceptual decisions require comparisons between current input and an internal template. Classic studies propose that templates are encoded in sustained activity of sensory neurons. However, stimulus encoding is dynamic, tracing a complex trajectory through activity space. Which part of this trajectory is pre-activated to reflect the template? Here magneto- and electroencephalography were recorded during a visual match-to-template orientation task, and used for pattern analyses of stimulus representation, template representation, and decision formation. The findings ran counter to the dominant model of sustained pre-activation. Instead, template information transiently emerged around stimulus onset and quickly subsided. Cross-generalization between stimulus and template coding, indicating a shared neural representation, occurred only briefly. The results are compatible with the proposal that template representation relies on a matched filter to transform input into task-appropriate output. This proposal was consistent with a signed difference response at the perceptual decision stage, which can be explained by a simple neural model.

Introduction

Human perception is flexible. The dimensions guiding our perceptual decisions can be updated rapidly as a function of the current task. When decisions are based on perceptual analysis, task goals could influence behaviour by creating an internal template: incoming sensory information is then matched against it. The representation of templates therefore plays a fundamental role in guiding perception and decision-making. Biased competition (Desimone and Duncan, 1995) provides a broad framework for how the brain interprets new sensory information in light of the current search template. A central tenet of this framework is that attention amplifies responses to template-matching stimuli in visual cortical neurons (Reynolds and Chelazzi, 2004). Single-cell neurophysiology (Chelazzi et al., 1993; Luck et al., 1997; Chelazzi et al., 1998) and human functional magnetic resonance imaging (fMRI; Chawla et al., 1999; Kastner and Ungerleider, 2000; Silver et al., 2007; Kastner et al., 2009; Reddy et al., 2009) have demonstrated that template representation and stimulus processing can occur in overlapping neural populations in visual cortex. Moreover, stimulus and template activity patterns cross-generalize (when measured with fMRI, Stokes et al., 2009), implying that the two share a common neural code. In the simplest case, increasing baseline activity of a stimulus-specific representation could boost target processing (Sylvester et al., 2009). This boost could facilitate target selection and reduce distractor competition for downstream processing resources (Bundesen et al., 2005).

However, recent findings complicate this simple model. Population-level analyses of time-resolved neural recordings show that stimulus decoding is highly time-specific (King and Dehaene, 2014), with discriminative activity patterns changing at the millisecond scale. Such dynamic coding has been observed at the level of population spiking patterns within individual brain areas (Meyers et al., 2008; Crowe et al., 2010; Stokes et al., 2013), and at the level of distributed activation patterns across the cortex (King et al., 2013; Cichy et al., 2014; Wolff et al., in press), suggesting that this temporal dimension is an inherent aspect of neural coding (Buonomano and Maass, 2009). Importantly, neural populations in visual (Meyers et al., 2008; Sreenivasan et al., 2014) and prefrontal cortex (Hussar and Pasternak, 2012; 2013; Stokes et al., 2013; Astrand et al., 2015) appear to represent a memorized stimulus with an independent pattern from that used during initial encoding.

The highly dynamic trajectory that stimulus processing traces through activation state-space challenges the classic model of template representation. The classic model proposes tonic activation of a static neural pattern, begging the question: Which of the many points along the processing trajectory should be pre-activated? The relationship between template representation and stimulus encoding must then be reconsidered in the light of two new questions: *When* (along the trajectory) is stimulus coding related to template coding? *How* (or, in what coding space) are the two related to form a perceptual decision?

One possibility is that the stimulus and template representations rely on different population codes in the same area. From the view of the experimenter, this would indicate that the codes may be comparable only at the level of representational similarity between different stimulus (or template) values (Haxby et al., 2014). From the perspective of the downstream neural circuit that needs to read out the similarity between stimulus and template in order to identify matches, this pattern could indicate the presence of a matched filter: the stimulus is initially encoded in a population that is insensitive to the template (encoding the stimulus value), which in turn activates a population that *is* tuned to the template (thereby encoding the template, c.f. Sugase-Miyamoto et al., 2008; Nikolic et al., 2009; Stokes, 2015), leading to an output that indicates the degree of overlap between template and stimulus.

While visual templates for target detection have been central to attention research, their role has been somewhat neglected in the study of perceptual decision-making. Perceptual decision-making tasks usually require the judgment of a visual stimulus feature against a fixed decision boundary or template (Gold and Shadlen, 2007). These tasks typically require judgments to be made at varying levels of perceptual difficulty (Vogels and Orban, 1990; Ghose et al., 2002; Purushothaman and Bradley, 2005; Summerfield and Koechlin, 2008; Scolari and Serences, 2010). However, the majority of perceptual decision-making studies to date have kept the decision boundary (or template) constant over the entire experiment, prohibiting a clear evaluation of the representation of

templates as distinct from stimulus representation and from the sensory-to-template comparison. In the present study, template and stimulus values were varied independently, permitting an examination of the extent to which their coding and their temporal profiles overlap. I used pattern analysis of simultaneously recorded magneto- (MEG) and electroencephalography (EEG) to track visual template matching with high temporal resolution as human observers performed a parametric match-to-template orientation task (Fig. 1a). Neural responses rapidly traversed a cascade of discriminative patterns, transforming the initial task-invariant stimulus code, in conjunction with the template code, into a decision-relevant code. Template patterns and stimulus patterns cross-generalized only in a short time window during initial processing, suggesting that there is some independence between the two neural codes. By contrast, the representational similarity structure between stimulus and template was similar for a more sustained period, suggesting that these distinct neural patterns nonetheless reflect comparable representations. This would be expected if templates are encoded as a matched filter in the connections between stimulus-sensitive and decision-relevant populations.

Interestingly, after the stimulus information was already reliably present and the response-relevant information had begun to emerge, neural signals additionally encoded the (task-irrelevant) *signed* difference between the current stimulus and the search template. This processing stage lends further support to the presence of a matched filter that

permits the flexible calculation of deviations from a search template. I argue, on the basis of a simple neural model, that this effect is consistent with the use of a population code for perceptual decision-making (Ma et al., 2006; Zemel et al., 2006; Beck et al., 2008a).

Materials and Methods

Participants

Ten healthy, right-handed volunteers (age range: 21-27yrs, 6 females) took part in the study, and were paid £10/hour for their time. Visual acuity was normal or corrected to normal. Ethical approval for methods and procedures was obtained from the Central University Research Ethics Committee of the University of Oxford. Each participant completed two experimental sessions, approximately one week apart, with each lasting approximately two hours (of which approximately one hour was spent on performing the task). Since the pattern analysis approach demanded high within-participant power, this design was geared towards a large number of trials per participant, rather than a large number of independent participants. While analyses were generally consistent across experimental sessions (i.e. independent replications), an ideal experiment would test a large number of trials in a large number of participants.

Experimental Setup

Participants completed the M/EEG scan inside a sound-attenuated, dimly lit, and magnetically shielded room. Stimuli were displayed onto a rear-projection screen (placed at a viewing distance of 90 cm) via a projector (Panasonic DLP Projector, PT-D7700E) with a spatial resolution of 1024 by 768 pixels and a refresh rate of 60 Hz. Stimuli were presented using Psychophysics Toolbox (Brainard, 1997),

running on MATLAB (Mathworks, Natick, WA). Participants responded using an optic-fibre response box by lifting their right index finger to indicate whenever they had seen a target. Participants were instructed to respond as quickly and accurately as possible.

Task

The task required the detection of visual targets within a stream of sequentially presented stimuli. The stream consisted of oriented Gabor patches (diameter: 4° visual angle, spatial frequency: 2 cycles/°), presented foveally for 100 ms, at an average rate of 650 ms (inter-stimulus interval, ranging from 516 to 783 ms). Orientations were drawn without replacement from a set of 16 possible angles. Stimuli were equally spaced from 5.625° to 174.375°, in steps of 11.25°. The task consisted of eight brief (approximately six-minute) blocks, in which 480 stimuli were presented (resulting in a total of 3,840 stimulus presentations per session). Each block began with the presentation of a target orientation (drawn at random, without replacement, from the 16 stimulus orientations), displayed centrally as a green line (4° length). Participants were instructed to respond whenever a Gabor patch with a matching orientation appeared. Since stimuli were drawn equiprobably from the 16 possible orientations, 1/16 of all stimuli were targets. Each block was cut into three shorter segments, giving participants brief rest periods. During the rest periods, the target orientation was presented again as a reminder.

Behavioral Data Analysis

Because of the rapid succession of stimuli, it is difficult to attribute unequivocally each response to a single stimulus. Therefore a stimulus-response assignment procedure was designed in order to attribute, in a probabilistic fashion, each response to a single stimulus.

First, response time (RT) distributions to stimuli were computed on the basis of their absolute angular distance (tilt) from the target orientation (from 0 to ±90º). When RTs were averaged relative to the orientation of the stimuli, it was clear that the responses fell within a certain time window (from approximately 200 to 1,000 ms), consistent with the approximately periodic presentation of stimuli. Tilt-dependent RT distributions were used to estimate the tuning of responses to the target. At each RT, the response tuning profile - the probability of a response given the tilt of the stimulus, from 0 to ±90º - was fitted with a von Mises distribution having two free parameters: the peak of the distribution P_{MAX}, and the concentration parameter kappa κ. The von Mises distribution was constrained to be centred at the target orientation (tilt = 0), and the definition of κ was modified such that $\kappa = 0$ indicates no tuning, $\kappa > 0$ indicates a preferred tuning for the target orientation, and $\kappa < 0$ indicates a preferred tuning for the orientation perpendicular/opposite to the target. For each subject, the tuning concentration showed a clear positive response following stimulus onset (approximately 200 to 1,000 ms post-stimulus).

This tuning information was then used to assign probabilistically each response to an individual stimulus. First, for each response, all stimuli that fell into the time window during which the tuning concentration was positive were preselected. Next, among these candidate stimuli (which had different tilts with respect to the target), the stimulus that maximised the probability of a response at the observed RT was selected. The resultant RT distributions truncated the low and high RT values, leaving the central part of the original RT distributions.

MEG and EEG Data Acquisition

Each participant completed two sessions: one MEG-only session, and one session in which EEG data were recorded concurrently. Participants were seated in the MEG scanner in a magnetically shielded room. Their legs were placed on leg rests and arms on their lap to avoid movements. Both experimental sessions lasted approximately one hour. Participants were instructed to maintain fixation on the centre of the screen during the stimulus blocks, and to minimize blinking.

Neuromagnetic data were acquired using a whole-head VectorView system (204 planar gradiometers, 102 magnetometers, Elekta Neuromag Oy, Helsinki, Finland). Magnetoencephalographic signals were sampled at a rate of 1,000 Hz and on-line band-pass filtered between 0.03 and 300 Hz. The participant's head position inside the scanner was localised and tracked continuously using head-position index coils placed at four distributed points on the head. Electrodes were placed above and

below the right eye for the vertical electrooculogram (EOG) and to the side of each eye for the horizontal EOG. In addition, eye movements were monitored using a remote infrared eye-tracker (SR research, EyeLink 1000, sampling one eye at 1,000 Hz, controlled via Psychophysics Toolbox, Cornelissen et al., 2002).

Electroencephalographic (EEG) data were collected in half of the sessions (for each participant), using 60 channels distributed across the scalp via the international 10-10 positioning system (AEEGS, 1991). Filtering, downsampling, epoching, and rejection of artefactual trials were performed on EEG data in the same way as on the MEG data. EEG data were added to all decoding analyses for the MEG+EEG sessions (except for the topographies in Fig. 3). There were no substantial differences in decoding between MEG-only and MEG+EEG sessions, apart from a small increase in decoding sensitivity in the latter. Therefore, all within-session analyses were averaged to arrive at participant-level results.

MEG Data Preprocessing

Data were preprocessed using the in-house OHBA software library (OSL), drawing on SPM8 (http://www.fil.ion.ucl.ac.uk/spm), Fieldtrip (Oostenveld et al., 2011), and Elekta software. The raw MEG data were visually inspected to remove and interpolate any channels with excessive noise, and were further de-noised and motion-corrected using Maxfilter Signal Space Separation (Taulu et al., 2004). Next, data were downsampled to 500 Hz. Remaining epochs with unsystematic noise

corruption were then excluded via visual inspection. Systematic artefacts, arising from eye blinks and heart beats, were identified via independent component analysis, and regressed out of the raw data. The cleaned data were then epoched with respect to each stimulus onset (from -1 s to +1s). In a final step, data were imported into Fieldtrip and inspected using the semi-automatic rejection tool to eliminate any remaining trials with excessive variance. All data were then baseline-corrected by subtracting the mean signal between -150 and -50 ms relative to stimulus onset (for analyses relating to the template, an earlier baseline, from -200 to -150 ms relative to stimulus onset, was used to explore the possibility that template information might be 'pre-activated' around the expected onset time. Using the standard baseline from -150 to -50 ms, however, did not change the results presented here). In addition, the data were smoothed with a 32-ms Gaussian kernel to reduce noise.

Orientation Decoding

A population tuning curve model was used to recover information about the stimulus orientation from the full M/EEG signal. Instead of looking to relate imaging data to different stimulus orientations directly, each stimulus orientation is represented using weights from a linear basis set of population tuning curves. Tuning curve models are well suited to recovering information about parametric features like orientations (Saproo and Serences, 2010; Brouwer and Heeger, 2011; Serences and Saproo, 2012; Garcia et al., 2013) or colors (Brouwer and Heeger, 2009).

To recover stimulus orientations, data were separated into a training set (all trials from 7 of 8 blocks) and a test set (the left-out block). For all trials in the training set, a matrix of 16 regressors was created, with the height of each regressor on any trial reflecting that trial's stimulus orientation (i.e., a regressor was set to 1 when the corresponding orientation was presented on that trial, and to 0 otherwise). The regressor matrix was then convolved with a half-cosine basis set (raised to the 15th power, see Brouwer and Heeger, 2009), in order to pool information across similar orientations. Orientation sensitivity at each M/EEG sensor was then calculated by regressing the design matrix against the signal (across all 306 sensors, or all 366 sensors in MEG+EEG sessions), separately for all time points in the epoch (in 4 ms steps, using a sliding window of 20 ms). The linear regression equation was fit to the data:

$$\mathbf{B_1 = WC_1;} \quad \text{(Equation 1)}$$

where C_1 is the design matrix (16 regressors x no. of training trials), B_1 is the training data set (306/366 sensors x no. of training trials), and W is the weight matrix (306/366 sensors x 16 orientation values) that needs to be estimated. This was done using ordinary least squares:

$$\mathbf{W = B_1C_1^T(C_1C_1^T)^{-1};} \quad \text{(Equation 2)}$$

Overall differences in signal magnitude between sensors were modeled out using a constant regressor in C_1. W was used to estimate the population orientation response (or tuning curve) in the test set, B_2 (306/366 sensors x no. of test trials):

$$\mathbf{C_2 = (W^TW)^{-1}W^TB_2;} \text{ (Equation 3)}$$

where C_2 is the recovered tuning curve, W is the weight matrix, W^T is its transpose, and W^{-1} is its inverse. Since both the design matrix and the estimated weight matrix were of full rank, this approach was equivalent to using the pseudo-inverse for estimation. For each trial, this curve was then zero-centered relative to the presented orientation. This procedure was repeated for each time point in the epoch before moving to the next iteration in the leave-one-out procedure. Zero-centered orientation curves were then averaged across trials.

The time course of the tuning curve was then converted into a stimulus information time course by calculating the linear slope of the tuning curve from -90° to 0°. Stimulus channels that were equidistant from 0° (i.e. +11.25° and -11.25°, +33.75° and -33.75°, etc.) were first averaged and then each resulting (sign-invariant) orientation channel time course was smoothed (with a 16-ms Gaussian kernel). A linear slope was fit across the orientation channels (from -90° to 0°), separately for each time point, session, and participant. Decoding accuracy was then evaluated using one-sample t-tests (against 0), under the assumption that

slopes are randomly distributed around 0 if there is no stimulus information in the signal. Correction for multiple comparisons across time was done using cluster-based permutation testing (10,000 permutations, Maris and Oostenveld, 2007).

A similar approach was used to test for encoding of the current template orientation, with the exception that here a 32-ms sliding window was used to increase sensitivity to a more slowly evolving effect. Since there were only 8 template orientations per session, and these were randomly selected from the 16 possible stimulus orientations, they were not always equally spaced across the circle. Orientation tuning curves were estimated across the eight irregularly spaced angles (using eight equally spaced regressors), and then the estimated tuning values at the 8 intermediate values were linearly interpolated. After interpolation, the template orientation tuning curves were treated as above to derive decoding time courses. Finally, this approach was also applied to calculating tuning profiles for information about the angular distance between the two (i.e., the orientation of the current stimulus with respect to the template, ranging from 0° for template matches, in steps of 11.25°, to 90°, for stimuli that were orthogonal to the template).

Onset latencies between stimulus, template, and angular distance were compared using a jack-knife approach (Miller, Patterson & Ulrich, 1998). We compared the onset times of significant coding ($p<0.05$, corrected) using t-tests. To estimate the variance of each onset time, we used an iterative procedure that left out one participant in turn and

calculated the onset time of significant coding across all remaining participants. The standard error of the latency difference was calculated using a revised measure that takes into account the reduced variability caused by the jack-knife procedure (Miller et al., 1998). The latency difference calculated across the entire set of participants was divided by this standard error estimate to provide t-statistics that were then evaluated using the conventional t-distribution.

Univariate Orientation Sensitivity Analysis

In addition to the pattern analyses that averaged signals over all sensors, the orientation sensitivity of individual M/EEG sensors was also evaluated to generate a topographical distribution of the sensitivity to the three task variables (stimulus orientation, template orientation, and decision-relevant angular distances). The baseline-corrected signal at each sensor and time point in the epoch was fit (across all trials) using a GLM consisting of pairs of regressors containing the sine and cosine of the three task orientations, along with a constant regressor. The pair of regression coefficients for the sine (β_{SIN}) and cosine (β_{COS}) of an orientation allowed calculation of orientation sensitivity A:

$$\mathbf{A = \sqrt{(\beta_{COS}^2 + \beta_{SIN}^2)};} \text{ (Equation 4)}$$

The amplitude A expected by chance alone was calculated by permuting the design matrix and repeating the amplitude analysis 1,000

times. The observed (unpermuted) amplitude was ranked within the permutation distribution of amplitudes to calculate a p-value, which was transformed into a z-score using the inverse of the cumulative Gaussian distribution (with center 0 and standard deviation 1). Sensitivity at the group level was then estimated by averaging z-scored amplitudes across session, participant, and the magnetometer and two gradiometers at each sensor location. These values were then plotted as topographies to illustrate the distribution of orientation sensitivity for the three task variables.

Cross-temporal Orientation Decoding

To assess the temporal stability of stimulus-specific topographies, the population tuning curve model was trained on one time point in the epoch, and the estimated weights were applied to all time points in the test data (using a sliding window of width 20 ms, applied every 12 ms). This was then repeated for all time points, creating a two-dimensional matrix of cross-temporal tuning curve slopes (with no additional smoothing). Dynamic coding can be inferred by comparing the decoding slopes on within-time training (i.e., training and testing on time t_1, or time t_2) with the decoding slopes on between-time training (i.e., training on t_1 and testing on t_2). The criterion for a dynamic epoch was: for each pair of time points $t_{i,j}$, coding is dynamic if the tuning curve slope is significantly higher (as measured by a paired t-test across 10 participants) within time than across time ($t_{i,i} > t_{i,j}$ AND $t_{j,j} > t_{i,j}$). Time windows of significant

decoding ($t_{i,j} > 0$) and windows of significant *dynamic coding* were identified using 2-dimensional cluster-based permutation testing (i.e., across both time axes).

Cross-Generalization between Stimulus and Template Patterns

To test whether stimulus-specific patterns cross-generalize to template-specific patterns, the cross-temporal tuning curve analysis was repeated, but calculating weights based on the presented template orientations in the training set, and then zero-centering the tuning curves of the test set with respect to the *stimulus* orientations. Here, a significantly positive tuning curve slope at time pair $t_{i,j}$ indicates that stimulus coding around time point i shares orientation-specific topographic patterns with template coding around time point j. For consistency with the other analyses, the training data were treated as in the analyses evaluating template coding, and the test data were treated as in the stimulus decoding analyses. Therefore the training data were preprocessed with a baseline of -200 to -150 ms, and smoothed with a Gaussian kernel of width 32 ms. The test data were preprocessed using a baseline of -150 to -50 ms, and were not smoothed. Weights were calculated using a sliding window of 32 ms, moving in 12-ms steps. The results were smoothed with a 20-ms Gaussian kernel. Again, permutation testing was applied to correct for multiple comparisons.

Representational Similarity Analysis

The rapid dynamics of the population tuning curve data indicated that, while the exact neural pattern might differ between time points and task variables, the information represented (as measured by their representational geometry) might be more constant over time. This possibility was tested with representational similarity analysis (RSA). Specifically, the approach involved calculating neural dissimilarity matrices between the M/EEG topographies evoked by different stimulus orientations. For each session, all trials were sorted by the presented stimulus orientation (into 16 bins), and then split in half (by separately averaging odd and even trials). The odd-even split allowed a comparison of the dissimilarity structures in two independent data sets, and verification of the reliability of the RSA. For simplicity, averages from both independent splits of the data were concatenated, resulting in 32 bins (i.e., 16 orientations x 2 data splits). For each of the 32 bins, the baseline-corrected average evoked response (across trials) was calculated at all sensors and time points. Next, for each time point (moving in 4-ms steps), the neural dissimilarity matrix was calculated by computing all pairwise Mahalanobis distances between orientations (using the within-condition covariance, pooled over all conditions). These dissimilarity matrices can be interpreted as reflections of the representational structure at each time point in the epoch.

In the first instance, I was simply interested in whether the neural dissimilarity structure was more stable over time than the underlying neural patterns (as calculated in the tuning curve analyses). To test for

this, the dissimilarity matrix from one half of trials at one time point was (Pearson-) correlated with the dissimilarity matrix from the other half of trials at all time points, generating a cross-temporal matrix of correlations between dissimilarity structures. If the dissimilarity structure is stable over time, this should result in significant correlations (measured via one-sample t-tests at the group level on the Fisher-transformed correlation coefficients) between time points (e.g., off-diagonal coding). This analysis was repeated for the other two task variables – the template orientation and the decision-relevant angular distance.

Next, I asked whether the neural dissimilarity structure, or geometry, was related to the parametric dissimilarity structure of the stimuli: since a 45° angle is more similar to a 60° angle than a 90° angle, the corresponding M/EEG topographies might be more similar as well. The stimulus dissimilarity matrix based on the pairwise angular distances between all presented orientations was regressed against the M/EEG dissimilarity matrix using a general linear model, fitting the model separately for each time point, session, and participant. Significant fits were assessed via one-sample t-tests. As an illustration of the presence of circular structure in the representational geometry, the 32x32 dissimilarity matrix was projected into two dimensions using multi-dimensional scaling (MDS).

Geometric analyses were repeated on the dissimilarity structure with respect to the template orientation, and the decision-relevant angular distance. For the latter (angular distances), the MDS results indicated that

the circular geometry of stimulus relationships was distorted by the decision likelihood. Therefore multiple regression was used to account for its possible influence (pattern component modeling, Diedrichsen et al., 2011). A first nuisance regressor was based on the similarity in response likelihood, which was estimated by calculating the participant-wise differences in response frequency between all decision-relevant angular distances. This regressor reflected differences in response likelihood between orientations, accounting for any effect of motor preparation. The second nuisance regressor captured the differences in the absolute decision value, i.e. the distance between the *unsigned* angular distances (the distance between 0° and -22.5° was 22.5°, but the distance between -22.5° and +22.5° was 0°, rather than 45°). This regressor accounted for the effect of linear decision value on the MEG pattern. This pattern was regressed out because it could reflect two possible decision mechanisms: population-based coding, as proposed here, or linear evidence accumulation (Gold and Shadlen, 2007). While the latter may still be at play in this task, the primary purpose of the analysis was to dissociate the two coding mechanisms.

Angular Distance Analysis

A final analysis examined how the entire population of M/EEG sensors dynamically encodes different task variables relating to angular distance representation. This was done by projecting population responses from the high-dimensional space (306 dimensions) of the MEG

signal into a low-dimensional task space (with each dimension representing a unique task variable). The approach emulated here has recently been described for populations of neural spike trains (Mante et al., 2013). First, in order to de-noise the data, the data were smoothed with a 20-ms Gaussian kernel, and the dimensionality of the MEG data was reduced from 306 sensors (or 306+60 sensors for MEG+EEG sessions) to 30 principal components (PCs) by calculating coefficients over the average time series at each sensor. The task variables were then fit to the reduced-dimensionality data using a GLM. The regressors were derived from the three main task variables: stimulus orientation, template orientation, and angular distance. Since all three are circular variables, pairs of regressors were used, consisting of the sine and cosine of each task angle, yielding a design matrix consisting of six regressors in total.

The fitting was done in a leave-one-block-out procedure: in turn, all trials from one task block were held out as a test set, and the GLM was fit on the trials in the remaining seven blocks (the training set). The GLM was solved on normalized data (by subtracting the mean and dividing by the variance across all trials in the training set). This yielded a set of six regression coefficients ('betas') for each time point in the trial and for each of the 30 PCs, which were then symmetrically orthogonalized (following Mante et al., 2013; Colclough et al., 2015). After normalizing the data from the test set (using the mean and variance from the training set), mean responses were calculated for all 16 angular distances in the test set (yielding a 16 angles x 30 PCs matrix). The means (16x30) were then

projected onto the task axes by multiplying them, time point by time point, with the betas (30 PCs x 6 regressors) from the training set, creating a 16x6 matrix at each timepoint for each left-out block. Projections were then averaged across the eight leave-one-out folds. The resulting projections estimate the sensitivity of each condition (i.e., each of the 16 angular distances) to each task variable (i.e., the six regressors), separately for each time point in the trial.

In line with Mante and colleagues, consistent deviations from 0 (as measured by one-sample t-tests across observers), in either direction along an axis, were interpreted as task variable sensitivity. In particular, the two regressors for the angular distance partialled out task-relevant and task-irrelevant aspects of the angular distance between stimulus and template: the cosine axis, with a maximum of 1 at 0º (targets), a minimum of -1 at the farthest non-targets (±90º, in 180º orientation space), and equal magnitudes for equivalent non-targets (e.g., 0.92 for both +11º and -11º), measured only the task-relevant aspect of the angle (i.e., the decision value, as shown in Figure S3). By contrast, the sine axis is insensitive to decision value (since, in 180º orientation space, sin(0º) = sin(±90º) = 0), but distinguishes the sign of non-targets (e.g., sin(+11º) = 0.38 = -sin(-11º)).

Neural Population Model

To summarize the MEG results, and to illustrate how they could arise from a very simple decision circuit, I used a population-based neural coding

model capable of performing the template-matching task used in this experiment. The model consisted of a three-layer architecture, with each layer consisting of neurons coding for different task variables (a stimulus layer, a template layer, and a decision layer). The stimulus layer consisted of a set of 100 units, each tuned to a different veridical stimulus orientation, with tuning determined by a von Mises distribution:

$$\mathbf{R_i(\theta) = exp(\kappa * cos(\theta - \theta_i))/A;}\ \text{(Equation 5)}$$

$R_i(\theta)$ indicates the response R of model unit i (tuned to θ_i) to orientation θ, with concentration parameter κ determining the tuning width of the response, and A reflecting a normalizing constant. Activation in the layer was then normalized to a range between 0 and 1. In the stimulus layer, the concentration parameter was set to 5 (Beck et al., 2008b, for the same parameter choice).

The template layer was identical to the stimulus layer, with the exception that template tuning was broader (κ=2), under the assumption that remembered stimuli would be encoded with lower precision than currently visible stimuli. Finally, the decision layer was identical to the template layer, with the conceptual difference that here, units were not tuned to veridical stimulus orientations, but to decision-relative orientations. In other words, activation of units near 0° in the decision layer reflected choice-relevant signals, irrespective of the current template orientation.

The stimulus and template layers were connected via a one-to-one mapping between identically tuned units (weight matrix $\mathbf{W}^{ST}$, with all connections between non-identical orientations set to 0). On each trial, the stimulus layer was initialized by setting the population response vector $\mathbf{R}^{S}$ in accordance with the stimulus orientation, and the template layer response vector $\mathbf{R}^{T}$ in accordance with the template orientation. In a second step, corresponding to a later processing stage, activation in the template layer was updated as a function of activation in the stimulus layer, by computing the element-wise product between $\mathbf{R}^{S}$ and $\mathbf{R}^{T}$. This step is similar to a Bayesian update, in which the prior distribution (the template layer response) is multiplied with the current evidence (the stimulus layer response) to produce a posterior distribution.

The crucial mapping for the task was between the template and decision layers, which consisted of an all-to-all reciprocal weight matrix $\mathbf{W}^{TD}$. The template layer unit tuned to the current target orientation had the strongest connection with the 0° unit in the decision layer (and neighboring template layer units were connected to correspondingly shifted units in the decision layer). All other connection weights fell off according to a von Mises distribution with κ=5 (although the exact tuning width did not substantially alter model behavior). This weight matrix shifted the response profile $\mathbf{R}^{T}$ in the template layer (which was still in veridical orientation space) to a response $\mathbf{R}^{D}$ in decision space. The decision layer response therefore permitted a direct mapping to decision- or motor-related output regions (which are omitted here). Importantly,

only the weight matrix $\mathbf{W}^{TD}$ needed to change in response to a change in the current template orientation.

Since I was mainly interested in the effects of reading out population activity in the decision layer, this model contained the simplification that codes in the three layers did not change over time. However, the dynamics of coding were not of interest for the question of whether population activity, *in principle*, could account for the neural results.

Results

Behavior

On average, observers detected approximately 70% of targets (Fig. 1b). They also made a large proportion of false alarms to near targets (approximately 50% for offsets from the target angle of ±11.25º), with false alarms rapidly dropping for more distant non-targets. Reaction times were distributed around 550 ms (Fig. 1c), with no effect of target proximity on reaction time ($p > 0.35$, Fig. 1d).

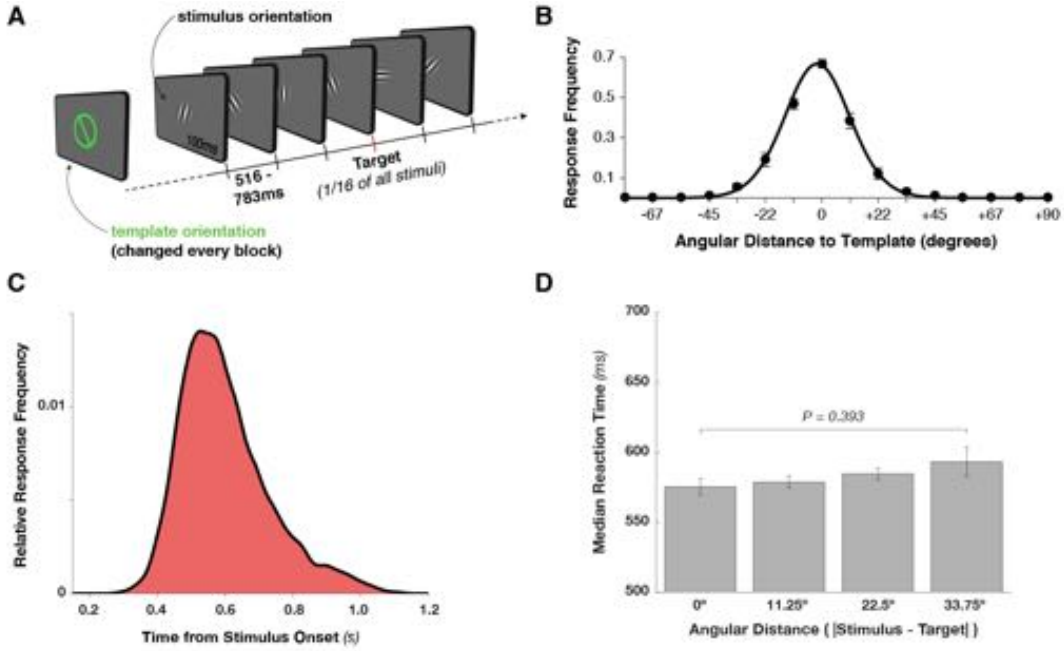


Figure 1. Task design and behavior. A. Each block began with the presentation of a target orientation, which observers maintained for the duration of a task block. Template presentation was followed by a serial stream of randomly oriented stimuli. Observers were asked to respond with a button press whenever the stimulus matched the template orientation. B. Response frequency was highest for target trials, and dropped off sharply for non-targets with increasing angular distance

between template and stimulus orientation. Error bars indicate standard error of the mean across observers. The black line denotes a von Mises distribution fit to the responses. C. Distribution of reaction time frequencies (as a proportion of all responses), from stimulus onset (collapsing over hits and false alarms). Beginning around 400 ms after stimulus onset, response frequency rises rapidly up to approximately 550 ms, after which responses slowly taper off. D. Target proximity (absolute angular distance between the stimulus and template angles) does not affect reaction time ($F_{3,27}$ = 1.036, p = 0.393*).*

M/EEG Signals Reflect Population Tuning Curves For Stimulus and Template Orientations.

Time-resolved population tuning curves (Fig. 2) showed that M/EEG signals reflected stimulus orientation shortly after its onset (Fig. 3b,c, Fig. 3a, 52-500 ms relative to stimulus onset, cluster-corrected p=0.0019).

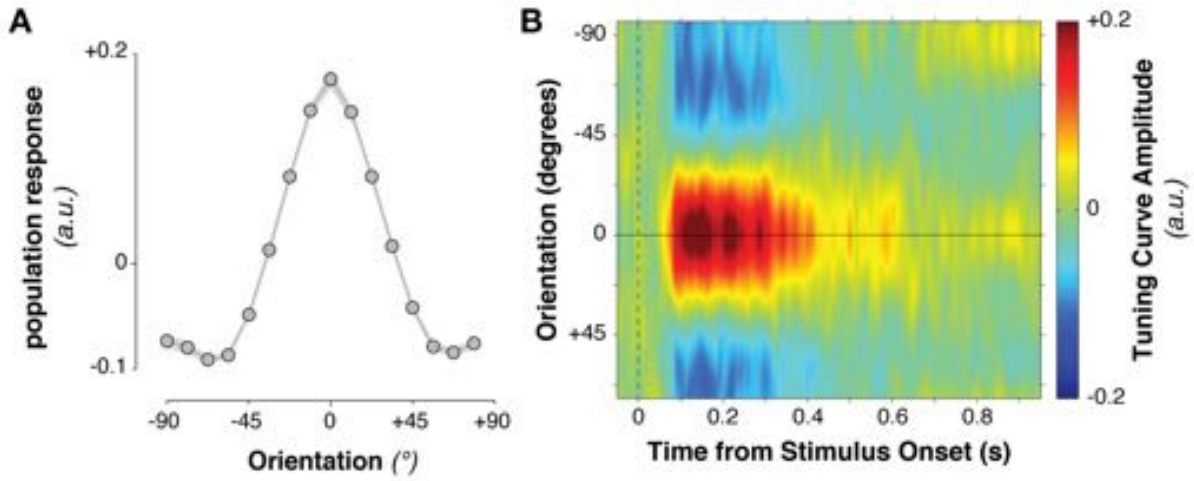


Figure 2. Stimulus-evoked population tuning curves. A. Average population tuning curve, 50-300 ms after stimulus onset. B. Time-resolved

population tuning curve, showing a sharp increase in the tuning curve slope shortly after stimulus onset, tapering off approx. 300-500 ms later.

Template orientation information was also present in the M/EEG signal during stimulus processing (Figure 3c, -72-324 ms, cluster p=0.0045). A jack-knife analysis comparing onset latencies between template and stimulus coding showed that template coding began significantly earlier (template: -72±13 ms, stimulus: 52±6 ms, t_9=-8.61, p=3*10^{-6}).

To track the temporal evolution of task-relevant coding (i.e., the decision-value), the distance of the current stimulus to the current template was also decoded (i.e. the angular distance between stimulus and template angles, from here on simply 'angular distance'). A strong effect of angular distance coding emerged around 200 ms (Fig. 3e, 164-596 ms, cluster p=0.0024, with an onset that was significantly later than for the stimulus orientation decoding, t_9=3.25, p=0.002). This effect was also present when training only on trials without a response (172-596 ms, cluster p = 0.0014), which discounts the possibility that this analysis simply reflected the difference in neural signals between responded trials (which were most frequent for small angular distances) and non-responded trials (most frequent for large angular distances).

While the three main task variables (stimulus, template, and angular distance) were all present in the signal, it is unclear whether different brain regions are involved. The sensor-level univariate sensitivity

to each variable was calculated (at MEG sensors only, averaging across the magnetometer and two gradiometers at each location) to determine the topographies associated with different task variables. To a first approximation, all three variables were encoded in signals in visuo-parietal sensors (Fig. 3b,d,f). While sensitivity was again strongest to stimulus orientation, template and angular distance responses nonetheless showed very similar topographies, indicating that all three variables might be computed in overlapping or nearby populations.

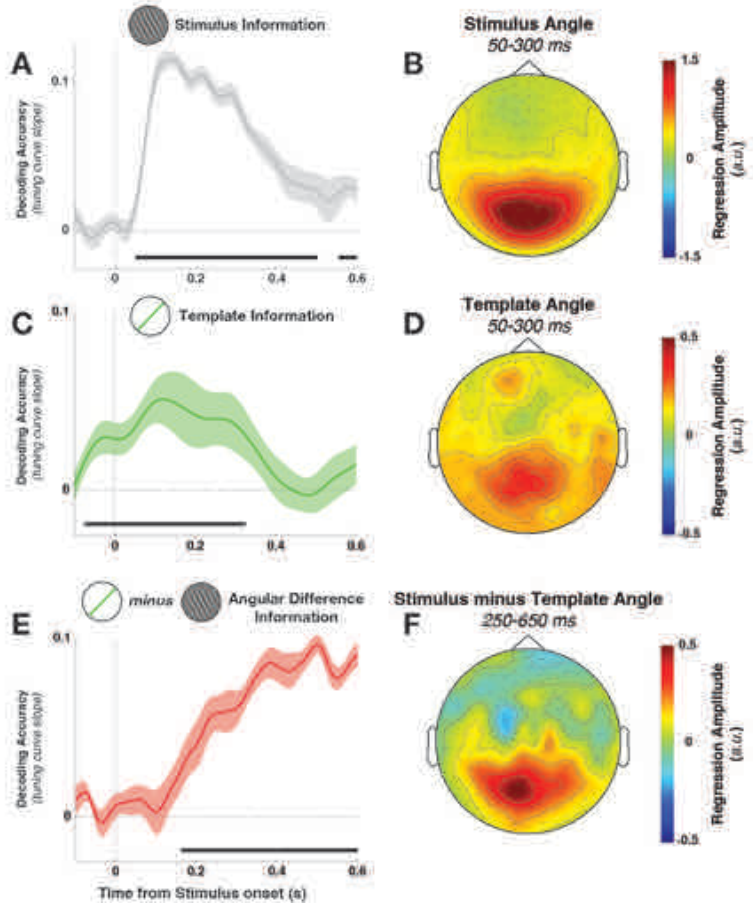


Figure 3. Task variable representation using population tuning curves (see Fig. 2). A. Stimulus orientation was represented in the early visual response. Weights were fit via linear regression, using all trials in all training blocks and estimated virtual channel responses in the test block.

Orientation-specific coding was estimated by calculating the linear slope of the tuning curve (between 0° and 90°). Consistent positive slopes indicate orientation selectivity at a given time point. Shading indicates between-subject standard error of the mean. Black bars denote significant time points (cluster-corrected). B. Univariate sensitivity for stimulus orientation, calculated at each sensor and time point. Topography shows the shuffle-corrected orientation sensitivity (z-scored against a distribution generated from permuting stimulus orientations 1,000 times), averaged across sensor triplets (two orthogonal planar gradiometers and one magnetometer) and the stimulus-decoding window. Color coding denotes the z-score, averaged across observers. C. Tuning curve slope and topography (D) for template orientation sensitivity. E. and F. show the same analyses, sorting trials by the angular distance between template and stimulus (i.e., the decision value).

Stimulus and Task Activity Patterns Vary Dynamically Throughout The Epoch.

To examine whether patterns of stimulus activity changed dynamically throughout the epoch, I tested for cross-temporal generalization of decoding (as elaborated in King and Dehaene, 2014). The population tuning curve approach was extended across time by calculating weights on one time point in the trial (on a training data set), and applying those weights to all time points in the trial (on the left-out test data set).

Decoding was significant in one main cluster along the diagonal (64-544 ms in training time, 52-436 ms in test time, cluster p = 0.0024, significant cluster extent indicated by color saturation in Fig. 4a). More importantly, stimulus decoding was time-specific, with decoding dropping off at off-diagonal train-test time points: Tuning curve slopes were significantly higher along the within-time diagonal than in their cross-temporal counterparts. This was evaluated using a conjunction t-test: each off-diagonal combination of timepoints ($t_{1,2}$) was compared against both on-diagonal within-time pairs ($t_{1,1}$ and $t_{2,2}$). Evidence for dynamic decoding was inferred if decoding for $t_{1,2}$ was significantly lower than both $t_{1,1}$ and $t_{2,2}$.

This drop-off is characteristic of dynamic coding (King and Dehaene, 2015; Stokes, 2015): despite significant decoding at two respective time points, the discriminative patterns do not generalize from one time point to the other. Off-diagonal generalization was significantly lower in a cluster (black outline in Fig. 4a) stretching from 52-304 ms (training time) and from 88-436 ms (generalization time, cluster p=0.0031, based on cluster extent). Since cross-generalization across time was significantly worse than within-time decoding, multiple stimulus-specific activity patterns seem to have been triggered in sequence. This suggests a dynamic code: the patterns of neural activity that map to specific information content change over time (Stokes et al., 2013). This can be contrasted with the relative stability of the information content (see representational similarity analyses below). It is perhaps interesting to

note that there was no evidence for periodic reactivation of orientation-specific patterns, which would be expected if the discriminating signal was oscillatory and phase-locked to the stimulus presentation (King & Dehaene, 2014).

Template decoding exhibited similar dynamics (Fig. 4b). Template information was present in an early cluster (training time: -140-340 ms, test time: -140-316 ms, relative to stimulus onset, cluster p=0.0191). In contrast to the stimulus decoding, template decoding showed no significant attenuation of decoding on the off-diagonal.

Decision-relevant angular distance decoding showed a similarly dynamic pattern, although off-diagonal decoding appeared to be more pronounced (Fig. 4c). Nevertheless, the strongest decoding was along the diagonal (training time: 4-592 ms, test time: 64-592 ms, cluster p = 0.0009), with significantly reduced off-diagonal generalization in this time window (training time 316-424 ms, generalization time 472-592 ms, cluster p=0.008). Since off-diagonal decoding was nonetheless significant in a large time window, it is possible that the angular distance exhibits both a dynamic and a static element. This could happen for two reasons. Firstly, it could occur if one population follows a dynamic processing cascade, while a separate population is tonically active in response to a given angular distance. Additionally, significant off-diagonal generalization could occur if there is temporal variability in the underlying processes, smoothing out the dynamic time-specificity across trials.

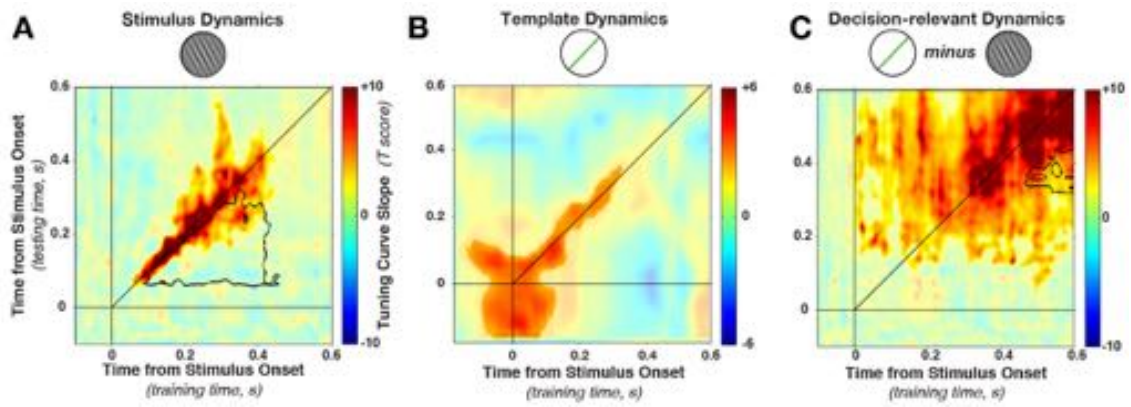


Figure 4. Cross-temporal generalization of orientation decoding. A. Tuning curve amplitude for stimulus orientation, estimated by calculating weights at one time point and applying them to test data at all time points in a trial. While the slope is consistently high along the diagonal (in the time window that contains significant stimulus information, between 60 and 500 ms, saturated colors), the slope drops sharply at off-diagonal train-test time coordinates. This indicates that the discriminative patterns are not consistent across time – rather they change rapidly, even while the stimulus can be readily decoded (i.e., off-diagonal decoding is significantly lower than on-diagonal decoding, black outline). B and C show the same analyses as in A, but sorting all trials by the template angle and the decision-relevant angular distance, respectively.

Cross-Generalization Between Stimulus and Template Neural Patterns

Training the population tuning-curve weights on template orientations around the time of stimulus onset (-150 to +300 ms) generalized to *stimulus* decoding shortly *after* onset (Fig. 5a, 52-124 ms, corr. p = 0.063). Using only the pre-stimulus time window (-150 to 0 ms) to train the template pattern, stimulus information could still be extracted

in this immediate post-stimulus period (Fig. 5b, average over 50-150 ms after stimulus onset, $t_9 = 2.45$, $p = 0.037$), indicating that template-specific neural patterns may be pre-activated immediately before stimulus onset. This result demonstrates that template activity patterns and stimulus activity patterns do cross-generalize (e.g., Stokes et al, 2009), but that the template-discriminative patterns around the time of stimulus onset only corresponds to stimulus-discriminative patterns in the earliest encoding period, and not later (even though stimulus decoding itself persisted up to 500 ms).

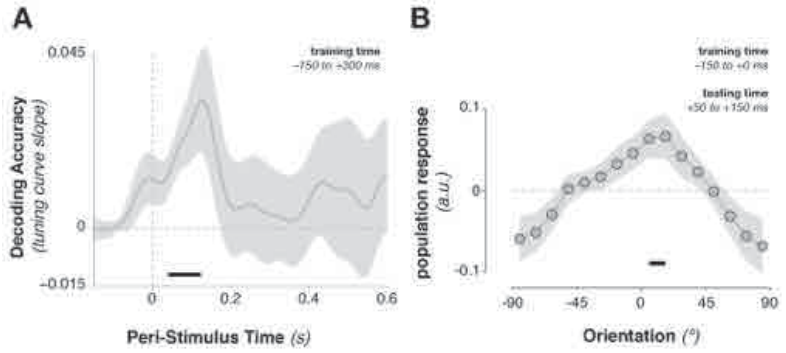


Figure 5. Cross-generalization from template-discriminative patterns to stimulus-discriminative patterns. A. Calculating tuning-curve weights relative to the template orientations in a training data set (in window from -150 to +300 ms around stimulus onset), applying these weights on test data, and sorting them relative to the stimulus orientation, showed decoding early after stimulus onset that quickly returned to baseline. B. Calculating population weights only on the pre-stimulus period (with respect to the template orientations) yielded a population tuning curve with a significant peak around the presented stimulus orientation (e.g. a

significant peak above the average response at 0°, and a significant positive tuning curve slope between ±90° and 0°). Shading indicates s.e.m. Black bars indicate significant time points or orientations ($p < 0.05$).

Stable Representational Structure for Stimulus and Task Variables.

While the underlying patterns separating different stimulus orientations change dynamically after stimulus onset, the information content that is represented might be more stable. The basic decoding analysis already implies that dynamic neural patterns contain stable information: the same basis set is used for decoding throughout the epoch. Therefore, significant decoding along the time-specific diagonal axis in the cross-temporal analysis suggests that stimulus orientation (Figure 4A), or angular difference (Figure 4C), are represented throughout significant changes in the underlying neural patterns. However, a more formal test of the representational structure of multivariate activity is provided by representational similarity analysis (RSA; Kriegeskorte, Mur et al. 2008). Rather than testing for discrimination per se, this approach focuses on the second-order pattern of condition-specific differences. This allows us to characterise the representational structure of the population code independently of the specific neural patterns associated with different stimulus orientations (Kriegeskorte and Kievit 2013). As an example, a 40°-tilted stimulus might elicit a topography shortly after onset that is more similar to the topography from a 30°-stimulus than that of a 90° stimulus. While the MEG patterns separating these stimuli would

change throughout the trial, the relative difference between them might stay the same. This would indicate that the same kind of information about the stimuli is represented throughout the trial. This approach has been developed specifically to characterize implementation-independent representational geometry (Kriegeskorte and Kievit 2013), and therefore is well-suited here to test whether dynamic neural patterns essentially code stable information.

This was tested by repeating the cross-temporal analyses on the dissimilarity relationships between MEG responses evoked by different orientations. Dissimilarity was quantified by the Mahalanobis distance matrices between all pairs of stimulus orientations (on one half of trials), separately for each time point in the trial. At each time point, this yielded a 16x16 distance matrix. Next, the same distance matrix was calculated for the remaining half of the data, and the two were compared via Spearman correlation coefficients between distance matrices, for each combination of time points in the trial.

In contrast to the dynamically varying stimulus-discriminative patterns, the representational similarity remained much more stationary (Fig. 6a), with a stable plateau of high correlations (Fisher-transformed Pearson's rho) from the earliest time of stimulus decoding. There was a significant (on- and off-diagonal) cluster early in the epoch (28-596 ms, cluster $p = 0.005$). The temporal stability of the representation was, as above, tested by examining whether on-diagonal similarity was higher than off-diagonal similarity. While one short-lived dynamic cluster

emerged (training time 82-146 ms, generalization time 130-226 ms, cluster p=0.0075, see black outline in Fig. 6a), the majority of the epoch was dominated by time-stable correlations.

Similarly, the representational similarity of angular distance was stable throughout the trial (Fig. 6b). Dissimilarity matrices correlated significantly in a later window in the trial (172-588 ms, cluster p = 0.0026), with no time points where within-time correlations were significantly higher than between-time correlations (all p > 0.20). These complementary analyses highlight the cardinal feature of dynamic coding: discriminative dimensions vary with time even though the information content remains constant (Stokes 2011).

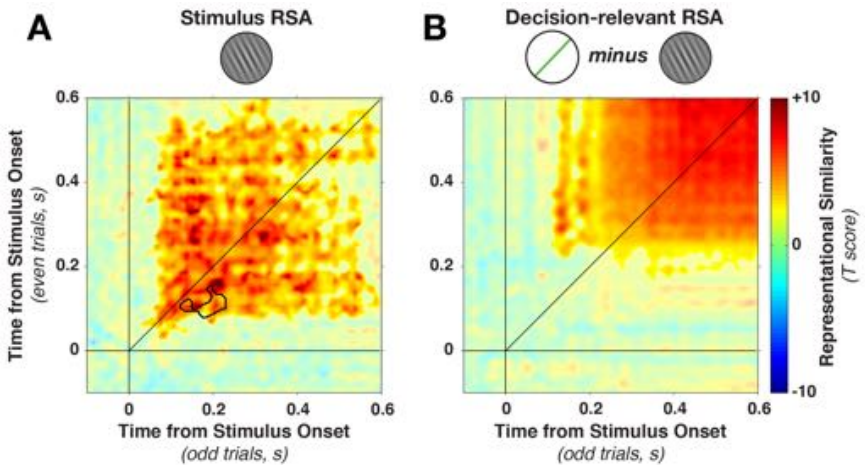


Figure 6. Cross-temporal generalization of representational similarity. A. Pearson correlations between distance matrices (between all possible stimulus orientations), calculated at different time points and on independent data sets. Black outline shows significant clusters at the group level (permutation test). The cluster extends off the diagonal,

indicating cross-temporal generalization. B shows the same analysis as in A, but sorting all trials by the decision-relevant angular distance.

Circular Representational Structure Suggests A Common Coding Scheme Between Stimuli and Templates

As the previous section indicates, the representational similarity of different stimulus orientations is more temporally stable than the underlying discriminative pattern. If RSA can reveal stable representations over time, it could also uncover representational similarity between the template and the stimulus. In other words, even though the MEG patterns did not persistently cross-generalize between stimulus and template decoding, the dissimilarity matrices calculated for template orientations and for stimulus orientations might correlate. This would indicate that stimuli and templates are represented similarly, even though the precise neural implementation might differ.

This relationship was quantified by testing for cross-temporal generalization of the dissimilarity matrix. However, here the dissimilarity matrix calculated between template orientations was correlated with the matrix calculated between stimulus orientations. Specifically, the Mahalanobis distance matrix between all eight template orientations, calculated at each time point, was correlated with the distance matrix between stimulus orientations (limiting our analyses to the eight stimulus orientations which were targets at some point in the session). The template-sorted dissimilarity matrix during a large proportion of the trial

correlated significantly with the stimulus-sorted dissimilarity around the time of visual processing (cluster in Fig. 7a, template structure from -48-196 ms, stimulus structure from 88-208 ms, cluster p = 0.038). The within-time comparison between templates and stimuli (i.e., the values along the diagonal) also showed a significant correlation (Fig. 7b, 104-176 ms, p = 0.036, with trends toward significance between 412-464 ms, p=0.079, and 552-596 ms, p=0.086).

What is the basis of this representational similarity between templates and stimuli? Given the simplicity of the stimulus set, a straightforward representational structure comes to mind: more similar stimulus (or template) orientations evoke more similar MEG topographies. However, this cannot be deduced from the population tuning curve analysis alone. To evaluate the possibility, neural dissimilarity matrices were calculated between the mean responses evoked by each of the 16 stimulus orientations, and this 16x16 neural dissimilarity matrix was projected into two dimensions for visualization (using multi-dimensional scaling, Fig. 7d). During the stimulus-encoding period (50-250 ms after stimulus onset), conditions fell onto a well-ordered circle: topographies were more similar (had a smaller Mahalanobis distance) if they were evoked by more similar stimulus orientations. This geometry was not present in the data before the onset of stimulus processing (-50 to +50 ms relative to stimulus onset, Fig. 7c). The temporal stability of this representational structure was evaluated by correlating the data-derived (Mahalanobis) distance matrix at each timepoint with an idealized

distance matrix, derived from the angular distances of the respective stimulus orientations (i.e., a 16-point simplex, corresponding to the pairwise angular distances between all 16 presented orientations, Fig. 7f). The idealized distance matrix of the stimuli was fit (via linear regression) to the neural distance matrix at each time point. The stimulus similarity matrix significantly fit the neural data (Fig. 7e, 44-432 ms, cluster p=0.002).

Likewise, the neural dissimilarity matrix between different *template* orientations was well described by the same simplex structure (Fig. 7g,h, 48-300 ms, cluster p = 0.0012). Therefore, while the discriminative patterns for stimuli and templates cross-generalized only briefly, their representation appears to be both stable over time and similar between task variables.

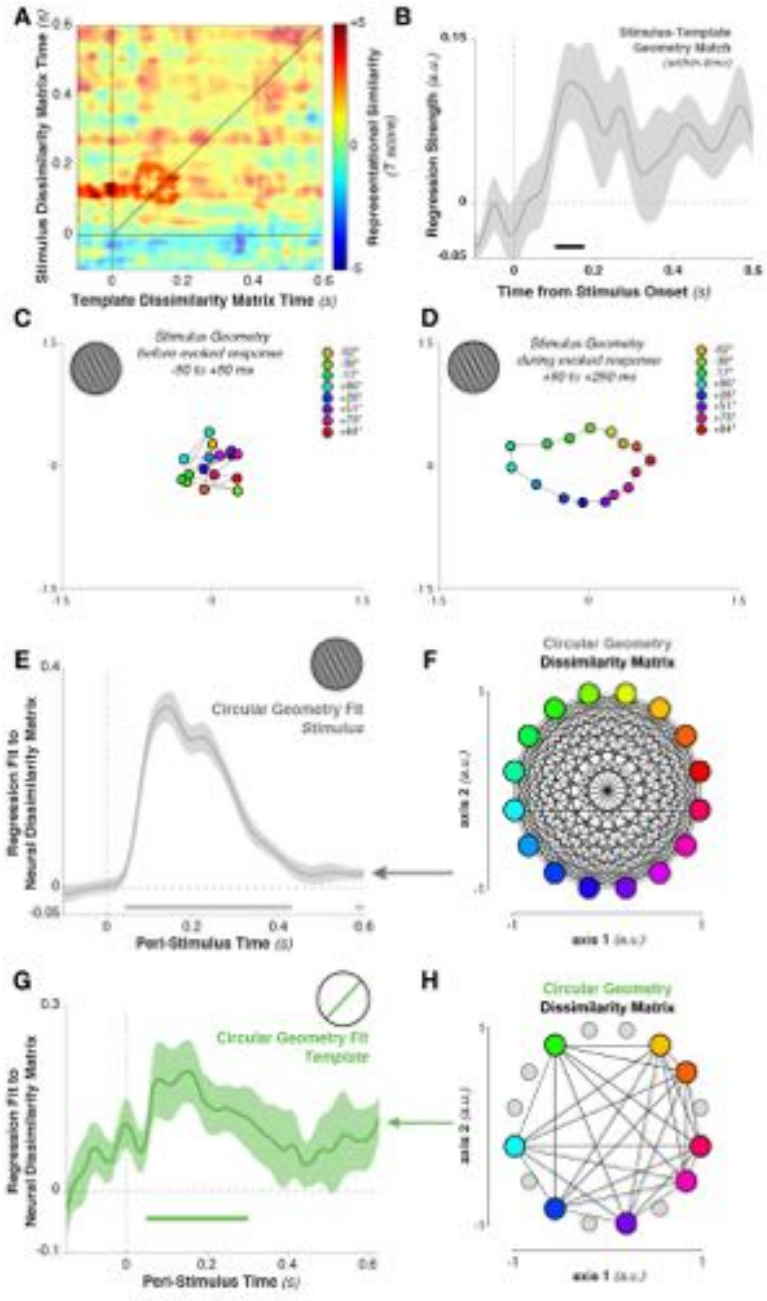


Figure 7. Geometry of stimulus and template coding. A. The representational similarity structures between template-ordered responses and stimulus-ordered responses were significantly correlated in the early stimulus-processing window (saturated colors indicate significant cluster). B. The within-time comparison also showed a significant correlation in the representational similarity structure up to 300 ms. Values correspond to the mean regression coefficient across all observers. Shading is between-

subjects s.e.m. C. Multi-dimensional scaling of the distances between stimulus orientations was not visible before stimulus onset. D. Circular structure indicated that responses fell onto a circular geometry shortly after stimulus onset. E. To quantify the representational structure over time, the distance matrix between all angles (16 different angles, split randomly into two sets of trials, resulting in a 32x32 distance matrix of Mahalanobis distances) was fit (via GLM) to the distance matrix of a 16-point circular simplex, shown in F. G. Similarly, relationships between the 8 template orientations fit a circular structure, particularly around stimulus onset times. H. An example simplex from one session, with the 8 chosen template angles highlighted in color, and the 8 stimulus orientations which were never targets shown in grey.

MEG Responses Encode the Task-Irrelevant Sign of the Stimulus-Template Difference

The next question was whether the decision-relevant angular distances (between the current stimulus and the template), which were necessary for guiding behavior, showed a comparable similarity structure. However, for angular distances, unlike for stimulus or template orientations, different angles had specific meaning for the task. In particular, the angular distance on each trial could be decomposed into two components: its magnitude and its sign. The magnitude of the angular distance (i.e., the absolute difference between stimulus and template orientation) solely determined the task-relevance of a stimulus:

the closer the magnitude is to 0º, the more likely the stimulus led to a target response. By contrast, the sign of the angular distance (i.e., whether a stimulus was oriented clockwise or counter-clockwise with respect to the template) had no relevance at all to the task, because it did not influence how close that stimulus was to the current template. The next analysis, therefore, attempted to isolate the effects of magnitude and sign on the neural response. Neural responses (from all 16 possible absolute distances, and from 306 MEG sensors) were projected (using independently calculated weights) onto two axes measuring the influence of magnitude and sign of the angular distance at each point in the trial (Mante et al., 2013). This allowed an analysis of the MEG signal's sensitivity to the decision-relevant magnitude (measured by the amplitude of the response along the magnitude axis), independently of its sensitivity to the decision-irrelevant sign of the absolute distance (measured along the sign axis).

The mean responses to the 16 angular distances fell roughly onto a circle that stretched out along the magnitude axis, with targets and near-targets separating from the extreme non-targets (Fig. 8a). Near non-targets that were either clockwise or counter-clockwise to the target separated along the decision-irrelevant axis, indicating an unexpected result: angular distances with equal magnitude but different sign (i.e., at an identical distance to the template orientation, such as -23º and +23°) evoked distinct and separable neural responses. Mean projections along the task-irrelevant axis for conditions with equal magnitude but different

sign diverged around the time of decision formation (348-588 ms, corrected p = 0.004, Fig. 8b). This was the case even without relying on the task projection approach: Mahalanobis distances between pairs of angular distance trials that had equal magnitude (i.e., -11° vs. +11°, -22° vs. +22°, -34° vs. +34°) were significantly greater than chance, corroborating the previous analysis (see Appendix, Fig. A1). In addition, a control analysis confirmed that different angular distances were not separable merely because of differences in response bias (Appendix, Fig. A2).

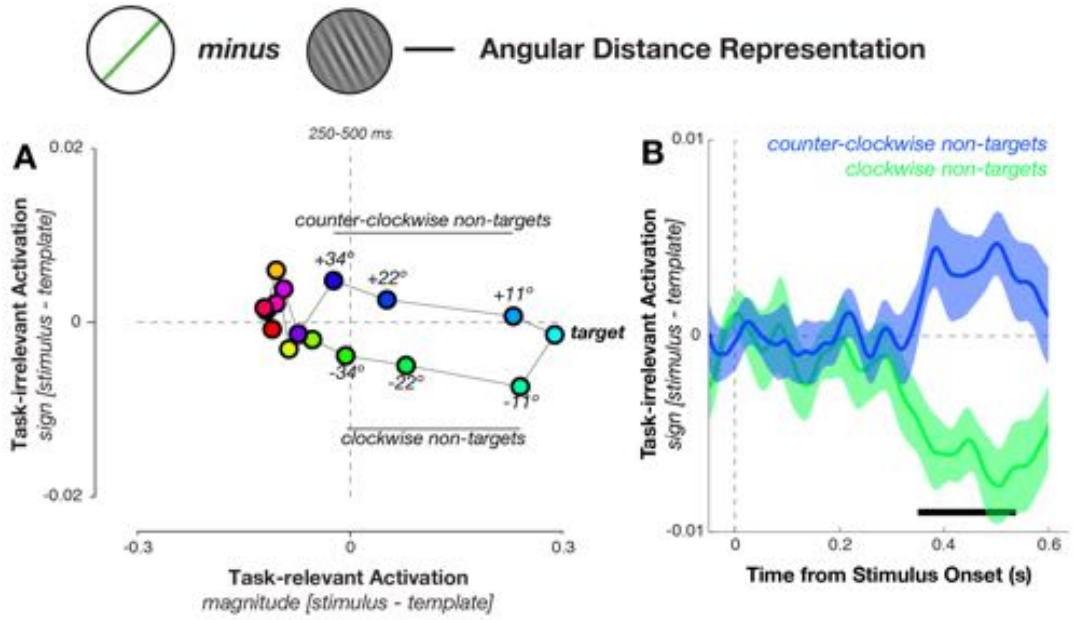


Figure 8. Geometry of response-related coding. A. Dissimilarity structure of angular distances. Data dimensionality was reduced using PCA, and weights calculated between sensor activity and different task variables using independent training data. Mean responses for each angular distance, calculated using the left-out data, were then projected via the calculated weights onto the task axes (the magnitude and sign of the angular distance). Since the task relevance of a

particular angular distance was defined solely by its magnitude, projections onto the sign axis measured sensitivity to task-irrelevant signed differences between conditions. Prior to decision onset (250-500 ms after stimulus onset), geometry is elliptical: in addition to conditions separating along the target-relevant magnitude axis (horizontal), near non-targets separate along the task-irrelevant sign axis (vertical). B. Task-irrelevant coding emerges approximately 300 ms after stimulus onset. Time courses for the three nearest non-targets (11°, 22°, 34°) separate along the task-irrelevant axis, depending on whether they are clockwise or counterclockwise to the target.

A Neural Model Based on Probabilistic Population Codes Accounts for MEG Responses

The neural encoding and sustained representation of the *signed* difference between the current stimulus and the template was unexpected because representing the sign of the angular distance was not necessary for solving the task at hand. However, the result yields some insight into the particular neural implementation of the decision process in this task. Specifically, a representation of both magnitude and sign is consistent with the use of a probabilistic population code. Probabilistic population codes assume that the brain uses the activation pattern across a whole population of neurons, each tuned to a different stimulus value (angular distance, in this case), to encode a probability distribution across the entire stimulus space (Ma et al., 2006; Zemel et al., 2006). The peak of activity of this distribution lies in or near neurons tuned to the presented angular distance, so that the location of the peak in the

population can drive behavior. Therefore, angular distances with equal magnitude but different sign will be separable in the population response, even if that separability is task-irrelevant. Importantly, this would not be the case if the entire population simply encoded the magnitude of the absolute distance (i.e., the overall match between stimulus and template, as is the case with accumulator models of decision-making). Therefore, the presence of signed difference signals in the MEG response strongly suggests that the brain uses a (probabilistic) population code to represent the decision variable in this task.

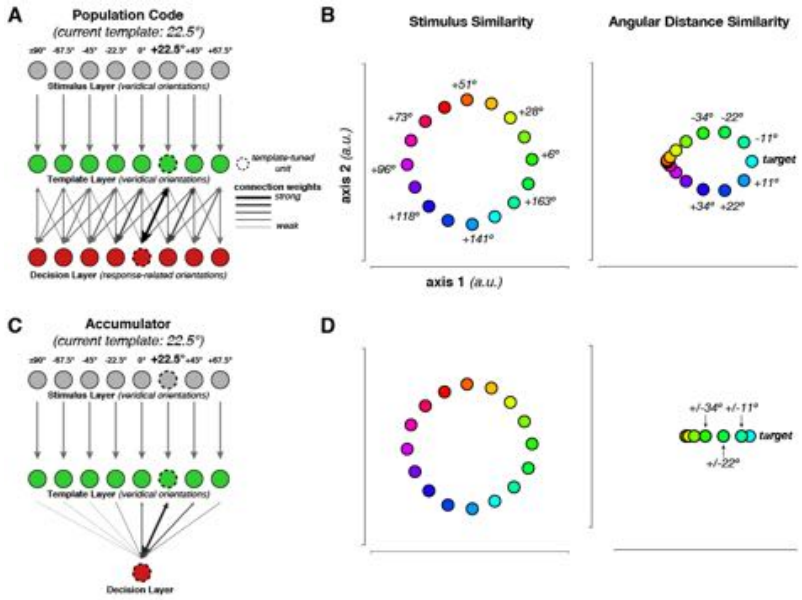


Figure 9. Neural Population Model. A. Population Code Model Architecture. B. Dissimilarity structure of responses in the stimulus layer (left panel) and in the decision layer (right panel). C. Accumulator Model Architecture. In contrast to the population code, decision value here is represented only in a single node (red unit in the decision layer). Otherwise, the architectures are identical. D. Dissimilarity structure of responses in the accumulator model. While responses in the stimulus

layer are identical in both cases, the decision layer differs from the population code model, in that the magnitude, but not the direction, of angular differences between stimulus and template, is represented.

I created a simple neural architecture (Fig. 9) to elaborate this argument for a population code and to link it to the neural data. The model consisted of three interconnected layers (see Appendix and Fig. A4 for details of model behavior). Each layer encoded information about one of the three task variables (stimulus, template, and angular distance between them). Each unit in a layer was tuned to a different orientation. Tuning in the decision layer represented *decision-relevant* angular distances, meaning that angles closer to 0° represented stimuli closer to the current template. The same dissimilarity matrices used in the M/EEG analyses were re-created from synthetic responses generated by each layer in the model. Identically to Fig. 8a, conditions with equal magnitude but different sign led to separable population responses in the template and angular distance layers. In contrast, a simpler model with only a single accumulator unit in the decision layer showed only a differentiation of conditions if they differed in magnitude, reflecting that here the signed angular difference was not encoded.

The population coding model used here is almost identical in architecture and behavior to a more elaborate biophysical model that was recently developed to predict the learning of new categories in a population of lateral intraparietal neurons in monkeys performing an

orientation discrimination task (Engel et al., 2015, Freedman and Assad 2006). As such, the results provide a bridge between animal models (and the access to single-cell activity they provide) and human decision-making.

Discussion

I used pattern analysis and representational similarity analysis of human M/EEG data, collected as observers performed a visual match-to-template-orientation task. Distinct, dynamically evolving temporal profiles reflected the representation of stimulus and template features, as well as the angular distance between the two (the behaviorally relevant variable). Unexpectedly, the task-irrelevant, signed angular distance between template and stimulus was also encoded.

Contrary to standard models of top-down attention, there was no tonic activation of the template neural pattern. Instead, the template representation emerged transiently around the time of the stimulus onset, and then quickly returned to baseline. In addition, stimulus-specific neural patterns changed rapidly during encoding, consistent with recent observations in visual cortical populations in monkeys (Meyers et al., 2008).

Although M/EEG measures neural activity at a larger scale relative to intracranial micro-electrode recordings, recent modeling demonstrates that the electromagnetic signal contains rich spatiotemporal information suitable for multivariate decoding (Cichy et al., 2015). Even very subtle differences in dipole position and/or angle elicit measurable, and statistically separable, patterns at the scalp surface (see Stokes, Wolff & Spaak, 2015). For orientation decoding, these differences presumably depend on idiosyncratic differences in the distribution of orientation columns along geometrically irregular cortical surfaces (Cichy et al.,

2015). Such subtleties average out at the group level, but can be characterized within participants using multivariate analyses. This logic extends in time: even very subtle differences in spatial distribution of activity patterns at different time points would result in idiosyncratic changes in dipole position and/or angle, resulting in a time-specific signal at the scalp surface. Indeed, the cross-temporal analyses suggest that orientation-specific patterns are also time-specific (see also Wolff et al., in press). In animal models, similar spatiotemporal pattern have been attributed to a cascade of neural engagement within the same brain area (Harvey, Coen and Tank, 2012), or time-specific changes in cell preferences (Sigala et al., 2008; Stokes et al., 2013). Although it is important to appreciate that decodability does not necessarily imply encoding in the brain (Tong and Pratte, 2012), the spatiotemporal structure observed in neural activity could plausibly provide a potentially rich source of information for guiding flexible (Fusi and Miller, 2013) and context-dependent behavior (Buonomano & Maass, 2009).

There was evidence that template-specific neural patterns measured around stimulus onset (and even slightly before) matched stimulus-specific patterns in the earliest encoding phase, but not later, when stimulus-specific differences had moved on to a different part of activity state space. The transient, rather than sustained, activation of template-specific patterns in this task may reflect the reactivation of a latent code (Mongillo et al., 2008; Buonomano and Maass, 2009) that was laid down in altered synaptic weights, but reactivated with increased (top-

down or stimulus-driven) input. The time course of template decoding could additionally indicate temporal reactivation in a quasi-rhythmic environment (as in 'rhythmic sampling mode', c.f. Schroeder and Lakatos, 2009), so that template information is maximal around the anticipated time of stimulus onset. These explanations are not mutually exclusive (both might be at play), but future studies manipulating bottom-up input strength (or their rhythmicity) could tease them apart (for example, by manipulating the contrast of random noise stimuli to probe the current state of the system; i.e., Wolff et al., in press).

The rapid dynamics of stimulus decoding further raised the question of whether the brain might use an alternative means of comparing dynamically evolving population codes (Laurent, 2002; Meyers et al., 2008; Stokes et al., 2013; King and Dehaene, 2014). Representational similarity analysis (Kriegeskorte and Kievit, 2013; Haxby et al., 2014; Nili et al., 2014) provides an approach that permits higher-order comparisons between different task variables, even if their underlying neural patterns are different. This is achieved by comparing the pattern of relationships between different instances of a task variable. Since here both stimulus and template variables were in orientation space, I was able to compare the dissimilarity structure across pairs of orientations between stimulus and template, even at time points when their underlying codes did not cross-generalize. In the space of pair-wise similarities, templates and stimuli were comparable for a longer duration, and especially around the putative time of decision formation (i.e., around

the time that decision-relevant decoding emerges). The fact that both variables rely on the same representational format indicates that neural activity coding for the stimulus is easy to compare to the neural representation of the template, even if the two are encoded in separate populations (and therefore do not cross-generalize at the decoding level). One simple possibility, for example, is that a stimulus-selective population is connected via a one-to-one mapping to a population encoding the template (see Fig. 9a).

A task-projection analysis that examined the representation of the decision-relevant angular distance showed that neural signals reflected both a task-relevant and, intriguingly, a task-irrelevant aspect of the angular distance. As expected from any decision-making model, the M/EEG signal reflected the task-relevant absolute distance between template and stimulus (i.e. the absolute degree of template match). However, the task-irrelevant sign of the difference between template and stimulus was also present in the M/EEG signal. This response could provide an important clue to the potential mechanism underlying perceptual decision-making in the current task. Population codes may underlie the representation of the angular distance, and encode the sign of the angular distance as a by-product of decision-making. This might be expected if population coding is the default solution to perceptual decision-making and target detection. Population coding could be useful because it can solve a variety of decision-making tasks, independent of the exact decision rule (i.e., it could switch rapidly from a target detection

task to a stimulus categorization task, as for example in Freedman and Assad, 2006; Engel et al., 2015). A similar argument has been put forth recently as an explanation for the coding of task-irrelevant stimulus variables in the population spiking activity of neurons in the lateral PFC (Mante et al., 2013; Donahue and Lee, 2015).

This idea was illustrated in a toy model that relied on population codes, and which exhibited the same behavior seen in the M/EEG data. A recent study comparing a biophysical model of category learning found, after learning, a similar circular geometry that was distorted by the location of the decision boundary in the population activity of LIP neurons (but not of MT neurons, where the circular geometry predominated, as in our data). Further, increased gain of target-tuned units in the template layer is consistent with behavioral evidence that feature-based attention modulates contrast gain (Liu et al., 2009) and that template-based perceptual learning improves template gain, which may be caused by reweighting at an intermediate decision layer (Dosher and Lu, 1999). It is important to note that the model was introduced merely to illustrate how the M/EEG results could arise from a simple decision circuit. Some (but not all) alternative architectures could equally implement a matched filter mechanism to account for the decoding of template orientation. For example, neural population tuning could shift to an increased tuning around the target orientation, essentially shifting the task-dependent modulation of synaptic weights to the connections between the stimulus and template layers (rather than the template to decision layers, as

suggested here). This effect of feature-based attention has been observed in macaque V4 (David et al., 2008; Hayden and Gallant, 2013; see also Roe et al., 2012). Importantly, such an alternative is also compatible with the matched filter operation, as it still relies on population-based decision-making. In contrast, a simple accumulator circuit is unlikely to account for all of our results.

The concept of matched filters for the evaluation of dynamically evolving stimulus codes could help bring together disparate findings on target representation in the attention, working memory, and predictive coding literatures. In attention research, classic studies on biased competition (Desimone and Duncan, 1995; Reynolds and Heeger, 2009; Chelazzi et al., 2011) have proposed that target features are either represented through persistent activation (Chelazzi et al., 1993), or through increases in synaptic gain of feature-tuned sensory neurons (Treue and Trujillo, 1999; Maunsell and Treue, 2006). In the present task, the transient temporal profile of template decoding (as opposed to persistent decoding throughout the trial) is more consistent with a latent coding mechanism (that could rely on temporary synaptic weight changes, Buonomano & Maass, 2009). Furthermore, recent computational and empirical work on the maintenance of items in working memory has argued that WM information (such as a visual template) can be stored through a reconfiguration of synaptic weights (Mongillo et al., 2008; Lewis-Peacock et al., 2012; Stokes et al., 2013), without requiring strong persistent activity (Watanabe and Funahashi, 2014). The encoding

of decision-relevant mnemonic templates in the weights of a network has the crucial advantage that processing of any new information can immediately pass through the modified weights and produce a matched filter response (Mongillo et al., 2008). Conceived in this way, contents in WM are decision rules that enforce the current stimulus-response mapping. Read-out at the time of the probe then consists in a perceptual decision (Martínez-García et al., 2011; Pearson et al., 2014). These hidden states, reflecting the current task context, could be in operation in the task, and would map onto the representation of the target orientation in the template layer of the toy model.

Since template orientation changed from block to block, decoding of the veridical stimulus could be dissociated from decoding of the decision-relevant angular distance in these analyses. How the brain passes on the *difference* between anticipated and perceived input has been formalized in the predictive-coding framework (Rao and Ballard, 1999; Friston, 2009; Bastos et al., 2012), but this framework does not address how task-specific templates are matched with current visual input. A recent computational study (Feldman and Friston, 2010), in an extension of the predictive-coding framework, proposed that feature-based attention acts by amplifying the gain of prediction error units coding for attended features, compared to unattended features. This attentional gain is captured in the toy model via the weighted response in the decision layer to near versus far non-targets: while near non-targets are clearly dissociable, non-targets that are far away from the current

template orientation all have nearly identical responses (both in the M/EEG data and in the model). This behavior could arise from synaptic gain increases of near-target units in the template layer or the decision layer.

Chapter 6: Neural Mechanisms of Updating vs. Protecting the Contents of Working Memory

Chapter summary

Previous chapters dealt with the cued inhibiting or updating of items into working memory (WM), but they did not explicitly examine their neuroanatomical basis or the timecourse of reacting to new stimuli that need to be updated or inhibited. In this chapter, this issue is revisited with an examination of the brain regions that are involved in the selective inhibition of stimuli when they need not be encoded in WM, or the selective updating of stimuli when they need to replace obsolete ones. Using MRI, we show that distinct regions respond to instructions to update vs. protect the contents of WM. We used MEG to show that these differences arise relatively late (i.e., in the delay period), indicating that updating and protecting are operations that can be performed retroactively, rather than during initial perceptual encoding of visual stimuli. The use of stimulus decoding techniques confirmed that stimulus information is initially present both for items that are to be encoded and items that are to be ignored. Only the encoded items, however, lead to persistent decoding into the delay period.

Introduction

A central component of human working memory is our ability to manipulate its contents. We can flexibly add new content, forget obsolete content, or modify existing content. An influential model of WM attributes this flexibility to a central, executive controller (Baddeley et al., 1996) that is capable of effecting all possible types of transformations. Sometimes this general function is described in terms of the ability to *update* the contents of WM, as opposed to *protecting* WM from interference. However, the literature has used this umbrella term to describe many different paradigms. Updating has been used to mean the encoding of additional items into WM (Burgess et al., 2007; Gorgoraptis et al., 2011), the replacement of individual items in WM with new stimuli (Murty et al., 2011), or the changing of stimulus-response rules from LTM (D'Ardenne et al., 2012).

This wealth of terms and paradigms has led to some confusion about which brain areas are involved in updating, and whether it is even useful to consider that control of WM is exerted by a unitary mechanism. For example, much of the neuroimaging literature has interpreted the common activation of dorsolateral prefrontal cortex (dlPFC) during WM maintenance as evidence that this region is necessary for control over the maintenance of individual items (Postle et al., 1999; Curtis and D'Esposito, 2003; Sreenivasan et al., 2014; D'Esposito and Postle, 2015), their protection from interference (Sakai et al., 2002; Suzuki and Gottlieb, 2013), and the updating of WM contents (D'Ardenne et al., 2012; Nee and

Brown, 2012). By contrast, WM updating paradigms have also reported a number of networks outside the dlPFC, including the striatum (Murty et al., 2011), parts of the cingulo-opercular network (Higo et al., 2011), or the frontal-parietal attention network (Nobre et al., 2004; Lepsien and Nobre, 2006, 2007). This suggests that WM control is not exerted by a single area, such as the dlPFC, but that it draws on distinct networks depending on the exact type and neural source of information that is being used to modify WM contents.

This chapter examines the cortical networks and temporal profiles of a simple updating operation, in which a single item in WM is replaced by a new one. Previous chapters in the thesis tested similar updating. For example, Chapter 3 cued items to be either encoded (precue) or ignored (retrocue). The cue gave observers the opportunity to prepare the selective encoding of an upcoming item, as compared to preparing to ignore all upcoming stimuli. However, in most real-world circumstances, we are not given advance warning about the behavioral relevance of an upcoming stimulus. Instead, its relevance is determined by the stimulus itself. This possibility was examined briefly in Chapter 4, where a visual stimulus was either added to the contents of WM, or ignored, depending on its color. However, this manipulation also affected WM load, making it equivocal whether differences were related to updating or reflected the increased number of items maintained in WM.

The aim of the current experiments was to use a simple update operation that did not lead to a concurrent increase in WM load. To keep

WM load constant, update cues prompted observers to replace the current content of WM with a new item, presented at the same time as the update cue. Updating was compared with a condition where a stimulus was cued as irrelevant and needed to be ignored. Here, the contents of WM needed to be protected from interference. Functional MRI (fMRI) and magnetoencephalography (MEG) permitted the isolation of networks involved in updating vs. protecting, and of the timecourse of network responses. In addition, decoding of stimulus information from the MEG signal made it possible to track the maintenance of items after *update* versus *protect* instructions.

Going beyond categorical control operations such as updating and protecting WM, more subtle updates to WM might depend on the exact features of relevant stimuli. For instance, if we need to update the contents of WM, but the new WM item happens to be identical to one already being maintained, the update might be simpler because it can draw on the similarity of the previous item. As demonstrated in Chapter 2, items recalled on the previous trial can have a strong influence on the recall of items on the current trial. Here, the question was whether similar biases exist *within* a single trial, when one item is replaced before it has been recalled. To measure this, we tested whether recall was biased towards or away from either the item recalled on the last trial, or the irrelevant item on the current trial.

Materials and Methods

Experiment 1 (fMRI)

Participants

Nineteen observers (11 females, mean age 26 years, range 22-36 years) performed the working-memory control task. All had normal or corrected-to-normal vision. Participants provided written consent in accordance with the university ethics guidelines, and were compensated at a rate of £15 per hour.

Apparatus

The task was created using the Psychophysics Toolbox (Brainard, 1997). Participants practiced one block of the task before entering the MRI scanner. During the scan, stimuli were presented on a shielded LCD monitor (40-inch Samsung LTA400HF, 60 Hz refresh rate, resolution 1024*768 pixels) that was projected into the scanner bore via a mirror. Fixation was monitored online using an infrared video-based eyetracker (Eyelink 1000, SR Research) sampling at 500 Hz. Participants were instructed to maintain fixation throughout the trial, excepting the response and inter-trial periods. Responses were made using a trackball mouse.

Task

The main task was to maintain an oriented stimulus in WM and to recall the orientation after a brief delay (Fig. 1a). During the delay, a cue

indicated whether the currently remembered item should be replaced by a second grating (*update* trials), or whether the second item should be ignored (*protect* trials). In addition, there were two single-item control conditions to compare the effect of encoding items either early or late in the trial, leading to a total of four conditions: two two-item conditions (*protect or update*), and two one-item conditions (*early or late*).

Each trial began with the presentation of a central fixation dot (0.2º visual angle) for 800 ms on a gray background (RGB: [127,127,127]). The first Gabor stimulus (6º diameter, 50% contrast, 2 cycles/º, with random orientation and phase, tapered by a Gaussian envelope with 1.5º standard deviation) was then presented for 200 ms. On trials where only the late item was presented, the central fixation dot changed color from black to light gray (RGB: [191,191,191]) to denote that the first item was not being shown. After a delay (1,600 ms), a second grating appeared (again with a random orientation, drawn independently from the first orientation). At the center of the grating, a colored cue replaced the fixation dot. The color denoted whether the first item should be kept in WM (*protect*) or be replaced with the item currently on the screen (*update*). After a further delay (5 s), a probe grating appeared at a random orientation, prompting participants to reproduce the remembered orientation using an MRI-compatible trackball. The maximum response time was 3 s, and was followed 300 ms later by feedback in the form of a grating with the correct orientation (50 ms). Trials were separated by an inter-trial interval (3 s, jittered between 1 and 5 seconds).

Cue colors were equiluminant and were counterbalanced across participants. Halfway through the experiment, the cue color assignment was also switched. Participants practiced the swapped assignment in the scanner (during the acquisition of the structural image) before starting the second half of the experiment.

The task was split into eight blocks, consisting of 30 trials each. Each block began with a 12 s fixation period to allow magnetization to reach a steady state before acquisition began. All four conditions were presented in random order in each block (10 protect and update trials each, and 5 trials each of the single-item controls). This yielded a total of 240 trials for the entire experiment.

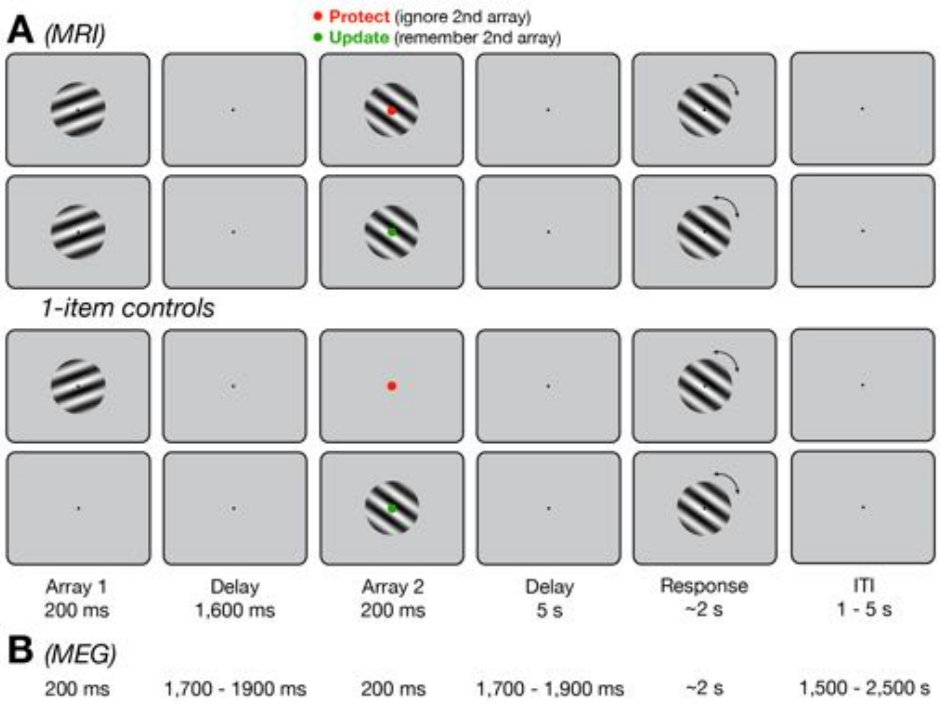


Figure 1. Task design for the MRI and MEG experiments. A. Task during the MRI scan. Two-item trials began with the presentation of a Gabor stimulus, followed after a short delay by a second Gabor stimulus,

presented together with a task cue. The cue color indicated whether the first (protect trials) or the current stimulus (update trials) would be probed at the end of a 5-s delay. After the appearance of the probe, observers had a maximum of 3 s to dial in the remembered orientation, followed by a variable delay. The one-item control trials were identical, with the exception that only one of the two Gabor stimuli was shown. B. The same task was presented in the MEG scanner, with timings tailored to the MEG environment.

Behavioral Analysis

For each response, we calculated the deviation (in degrees) of the recalled orientation from the actual orientation of the probed stimulus. We calculated recall accuracy as the inverse of the circular standard deviation of the response distribution (corrected for the precision expected from randomly distributed responses, see Bays et al., 2009), separately for each condition. We used a 2x2 repeated-measures analysis of variance with the factors cue (protect, update) and load (one, two items) to compare accuracy between conditions. We also fit a mixture model to each response distribution:

$$p(x|\theta) = P_{MEM}*VonMises(x,\theta,\kappa) + P_{NP}*VonMises(x,\theta_{NP},\kappa) + (1- P_{MEM} - P_{NP})/2\pi$$

where x is the response angle, θ is the orientation of the probed item, *VonMises* denotes the Von Mises distribution ($f(x) = e^{\kappa*\cos(x-\theta)}/2\pi*I_0(\kappa)$, where

$I_0(\kappa)$ is the modified Bessel function of order 0), i.e. the circular analogue of the normal distribution, and $(1- P_{MEM} - P_{NP})/2\pi$ is the height of the uniform distribution arising from randomly distributed guesses. The three variable parameters P_{MEM}, P_{NP}, and κ, define the relative likelihoods of drawing a response from one of the three response distributions: the probability of recalling an item (P_{MEM}) with concentration parameter κ, or the probability of erroneously misreporting a non-probed orientation θ_{NP} (with probability P_{NP}), or the probability of guessing at random ($1 - P_{MEM} - P_{NP}$). Model parameters for different conditions were compared using ANOVA.

As outlined in Chapter 1, unprobed stimuli on the same trial and on previous trials can have an influence on the recall of the probed stimulus. Here we explore this idea further by looking for biases arising from the unprobed item on the current trial, and from the probed item from the previous trial. For the within-trial analysis, trials were binned (into 64 overlapping bins, each containing 1/4 of all trials) according to the relative orientation of the unprobed item with respect to the probed item. The average response bias was calculated for each bin, yielding a bias curve as a function of relative orientation. To quantify any influence, we integrated the area under the bias curve, separately for negative and positive angular differences. If the average area for positive relative orientations was smaller than for negative orientations, this would indicate a *repulsive* effect of the unprobed stimulus, whereas a positive area difference would indicate an *attractive* bias. In order to compare the

effects reported here with those described by Fischer and colleagues (2014), we also fit a Derivate-of-Gaussian function to the binned data, described by the following equation:

$$y = awcxe^{-(wx)^2}$$

where *x* is the relative orientation of the unprobed item, *y* is the response relative to the probed item, *a* is the amplitude of the bias, *w* is the standard deviation of the bias curve, and *c* is a normalization term ($\sqrt{2}/e^{-0.5}$) to make the fit comparable to the empirical bias. Significant bias was assessed using a one-sample t-test on the bias amplitude estimate. Here, a negative value again indicated a repulsive bias, whereas a positive value indicated attractive bias.

MRI Acquisition

Functional data were acquired on a 3-Tesla MRI scanner (Siemens Trio) with a 32-channel head coil, using an accelerated multi-band echo-planar imaging sequence (voxel size 2x2x2 mm, 72 slices, 106x106 pixels per slice, 1,300 ms repetition time, 40 ms echo time, 66º flip angle, multi-band acceleration factor 6). A high-resolution structural T1 scan (MPRAGE, voxel size 1x1x1 mm, 192 slices with 256x256 pixels per slice, 2,040 ms repetition time, 4.7 ms echo time, flip angle 8º) was acquired halfway through the experiment, as participants practiced the task with the swapped cue color assignments.

MRI Preprocessing

Functional images were preprocessed using SPM8. Images were first motion-corrected and unwarped. After coregistering the structural image to the functional images (Ashburner and Friston, 1997) and warping it to the template of the Montreal Neurological Institute (Evans et al., 1992), the functional images were transformed into MNI space and resampled to a resolution of 3x3x3mm. Images were smoothed with an 8-mm isotropic Gaussian smoothing kernel and highpass-filtered (with a cutoff of 128 s).

MRI Analysis

Imaging data were analyzed using a general linear model. Each of the four conditions was modeled with a regressor locked to the onset of the cue (and second stimulus), with a separate regressor locked to the onset times of the first stimulus (pooled over conditions, although modeling both stimulus onsets in a single regressor per condition led to comparable results). Additional regressors were used to capture the effects of the probe stimulus onset (across all conditions) and of motion artifacts (translation and rotation in three orthogonal directions, six regressors in total). Separate sets of regressors of interest were created for each of the eight task blocks. After convolving regressors with the canonical hemodynamic response function (a double-gamma function peaking approximately 6 s after onset), a single model was fit for each participant. Contrasts at the first level were passed on to the group level, where statistics were computed with standard parametric tests (Friston et

al., 1995). Correction for family-wise error rate was done at the cluster level using Gaussian random field theory (Worsley and Friston, 1995), using an underlying uncorrected voxel-level threshold of $p<0.0005$, with a minimum cluster extent of 25 voxels.

Experiment 2 (MEG)

The second experiment measured magnetoencephalographic (MEG) responses in a similar WM control task (Fig. 1b). In order to optimize the task for MEG data acquisition, intervals between items were jittered between 1,800 and 2,000 ms. This shortened the average trial length, so that the total number of trials could be increased from 240 to 400 without increasing the overall duration of the experiment beyond 90 minutes. Trials were equally distributed across all four conditions (100 trials each for the protect and update trials, for both 1-item and 2-item conditions).

Participants

Twenty observers (eleven females, mean age 25.4 years, range 20-36 years) participated in the MEG experiment. Eighteen had also participated in the MRI study. In order to improve the sensitivity of the model fits, behavioral analyses were performed by concatenating the data sets from both MRI and MEG sessions, where available. All had normal or corrected-to-normal vision. Participants provided written consent in

accordance with the university ethics guidelines, and were compensated at a rate of £15 per hour.

Experimental Setup

Participants completed the MEG scan inside a sound-attenuated, dimly lit, and magnetically shielded room. Stimuli were displayed on a rear-projection screen (placed at a viewing distance of 90 cm) via a projector (Panasonic DLP Projector, PT-D7700E) with a spatial resolution of 1,024 by 768 pixels and a refresh rate of 60 Hz. Stimuli were presented using Psychophysics Toolbox (Brainard, 1997), running on MATLAB (Mathworks, Natick, WA). Participants recalled probed orientations using an optic-fibre response box by pressing their right middle finger to rotate the orientation of the probe stimulus clockwise, and their right index finger to dial it counterclockwise. Observers confirmed their response by pressing a button with their left index finger.

MEG Acquisition

Each participant completed one session. Participants were instructed about the task requirements and performed one practice block inside the MEG scanner. Participants were instructed to maintain fixation on the centre of the screen throughout the trial, and to minimize blinking. The task was stopped for rest breaks approximately every four minutes.

Neuromagnetic data were acquired using a whole-head VectorView system (204 planar gradiometers, 102 magnetometers, Elekta

Neuromag Oy, Helsinki, Finland). Magnetoencephalographic signals were sampled at a rate of 1,000 Hz and on-line band-pass filtered between 0.03 and 300 Hz. The participant's head position inside the scanner was localised and tracked continuously using head-position index coils placed at four distributed points on the head. Electrodes were placed above and below the right eye for the vertical electrooculogram (EOG) and to the side of each eye for the horizontal EOG. In addition, eye movements were monitored using a remote infrared eye-tracker (SR research, EyeLink 1000, sampling one eye at 1,000 Hz, controlled via Psychophysics Toolbox (Cornelissen et al., 2002).

MEG Data Preprocessing

Data were preprocessed using the in-house OHBA software library (OSL), drawing on SPM8 (http://www.fil.ion.ucl.ac.uk/spm), Fieldtrip (Oostenveld et al., 2011), and Elekta software. The raw MEG data were visually inspected to remove and interpolate any channels with excessive noise, and were further de-noised and motion-corrected using Maxfilter Signal Space Separation (Taulu et al., 2004). Next, data were downsampled to 500 Hz. Systematic artefacts, arising from eye blinks and heart beats, were identified via independent component analysis, and regressed out of the raw data. The cleaned data were then epoched with respect to each stimulus onset (from -1s to +2s). In a final step, data were imported into Fieldtrip and inspected using the semi-automatic rejection tool to eliminate any remaining trials with excessive variance. All data

were then baseline-corrected by subtracting the mean signal between -100 and -50 ms relative to stimulus onset (for ERF analyses). In addition, the data were smoothed with a 24-ms Gaussian kernel to reduce noise.

MEG Sensor-space ERF Analysis

In order to combine the event-related responses between cue conditions, the orthogonal gradiometer pairs were rectified and averaged, yielding a total of 102 measurements per timepoint. For the ERF analysis, the pre-stimulus baseline was subtracted after combining gradiometers. Cue conditions at each timepoint were compared using paired t-tests, and corrected for multiple comparisons with cluster permutation testing (10,000 permutations). Visual sensors (in the posterior half of the sensor array) with a significant positive response in the first 400 ms ($p<0.01$) across all cues were then averaged to create one timecourse for the visual response. This timecourse was again compared between conditions.

Sensor-space topographic distance analysis

We used a difference metric (that pooled signals across all sensors) to calculate condition differences at each timepoint. Distance metrics are calculated within participant and are therefore insensitive to imperfect alignment of different participants under the sensor helmet. At each timepoint, the average response (across all 306 sensors, used without rectification) was calculated, separately for protect and update

cues. Next, the Mahalanobis distance between the averages was calculated at each timepoint, using the noise covariance matrix estimated on the trial-wise residuals (after subtracting out the condition means). The true distance was then compared with a shuffling distribution (generated by randomly permuting trial labels and re-computing the Mahalanobis distance 5,000 times), and the z-scored percentile rank of the true distance within the shuffling distribution was calculated at each timepoint. Z-scores were then compared against 0 at the group level using a one-sample t-test (again correcting for multiple comparisons with a cluster permutation test).

Source-space Analysis

Source-space analysis was done using beamforming. Because the accuracy of this method relies on the coregistration of accurate anatomical data from a high-resolution structural scan, only the 18 participants who had also undergone the MRI experiment could be used in this analysis. Source-space signals were derived from eight regions of interest, identified in the three networks found in the MRI experiment (Fig. 5): (1) the cingulo-opercular network (bilateral insula, anterior cingulate), (2) the updating network (bilateral extrastriate visual cortex, superior frontal gyrus), and (3) attention-related parietal cortex (bilateral inferior parietal sulcus / superior parietal lobule). For each region of interest, an 8-mm sphere was drawn around the peak MNI coordinate from the corresponding MRI analysis. A virtual sensor was created for all voxels in

each ROI (typically 20-30 voxels) using a linearly constrained minimum-variance (LCMV) beamformer, which normalises sensors (magnetometers and gradiometers) by their minimum eigenvalues, and regularizes the beamformer's covariance estimation using the top 58 principal components (Woolrich et al., 2011). The single-shell (Nolte, 2003) forward model was used for beamforming, fit to each participant's structural scan. Signals from all voxels in an ROI were then submitted to time-frequency analysis. For the time-frequency decomposition, the time-domain signal was submitted to a fast Fourier transform in a sliding window moving in steps of 20 ms, with a variable width determined by a Hanning taper that depended on the frequency (4 cycles), in a window from -600 to +1,900 ms. Power was estimated at frequencies from 4 to 40 Hz, in 31 logarithmically-spaced steps.

Power at each frequency was log-transformed ($10*\log_{10}$) and baseline-corrected (-350 to -150 ms before onset of the cue), and power at each frequency and timepoint was then compared between protect and update trials. For a more focused analysis, we averaged power in each ROI over two epochs: early (0-500 ms after cue onset) and late (500-1,500 ms), and over three frequency bands: theta (4-8 Hz), alpha (8-14 Hz), and beta (15-30 Hz).

Sensor-space Stimulus Orientation Decoding

A final analysis step was aimed at recovering single-trial estimates of the orientation of the second stimulus from the event-related MEG

responses. This was done using an iterative leave-one-trial out procedure. For each left-out trial, the 1/4 of remaining trials with the most similar stimulus orientation (‘same trials’) and the 1/4 of trials with the most different stimulus orientation (‘orthogonal trials’) were selected. For each timepoint in the trial, the Mahalanobis distance between the left-out trial and the average of the same trials and of the orthogonal trials was then computed (using the same noise covariance matrix, computed on all but the currently left-out trial). Under the assumption that, on average, the left-out trial should have a shorter distance to the same orientation trials than to the orthogonal orientation trials, the two distances were then subtracted (to create a ‘distance difference’, see also Wolff et al., in press). This was repeated for all trials, and the distance differences were averaged separately for protect and for update trials. These were compared at the second level against 0 using one-sample t-tests (to identify timepoints of significant stimulus decoding) and against each other (using paired t-tests) to identify timepoints where one condition provided significantly more stimulus information than the other. Statistics were corrected for multiple comparisons using cluster permutation tests.

Results

Behavior

Behavioral data were concatenated from both the MEG and the MRI sessions, where available, yielding a total of 21 independent data sets (18 participants with both sessions, two participants with MEG, and one participant with MRI).

Accuracy

A 2x2 repeated-measures ANOVA showed that recall accuracy (Fig. 2a) was affected both by the cue condition (*protect*/*update*, $F_{1,20}$=32.8, p=1.32*10^{-5}) and by load (1 vs. 2 items, $F_{1,20}$=27.0, p=4.37*10^{-5}). There was a significant interaction between cue and load ($F_{1,20}$=9.25, p=0.0064) because the effect of presenting two items was more detrimental on *protect* (t_{20}=-5.55, p=1.97*10^{-5}) than on *update* (t_{20}=-2.00, p=0.059) trials. These statistics confirmed that presenting a second item was in fact distracting on the *protect* trials, and, to a lesser degree, that having an item in mind on *update* trials led to a (slight) reduction in accuracy (but see the modeling result in Fig. 2b, which did not show any difference).

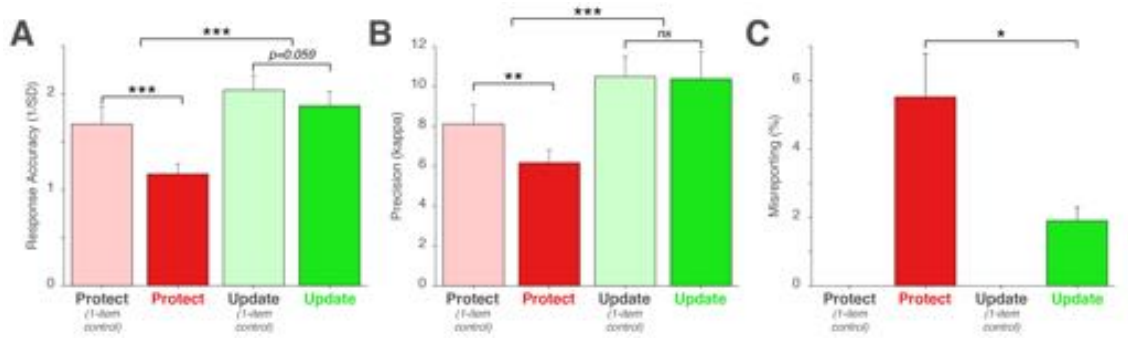


*Figure 2. Accuracy and Mixture Modeling of Behavior. A. Accuracy, as measured by the inverse of the standard deviation (1/SD) of the response distribution, was lower on protect trials. This was not merely the result of a longer maintenance period on protect trials, since the performance difference was much smaller in the 1-item control conditions (leading to a significant cue by load interaction: $F_{1,20}$=9.25, p=0.0064). B. The same pattern was largely reflected in the precision parameter of the fitted model. C. On trials where two items were presented, the model additionally calculated the misreporting rate, which was significantly higher for protect trials. Error bars denote standard error of the mean. *: p<0.05, **: p<0.01, ***:p<0.001.*

Mixture-Modeling Results

The presentation of a second item on *protect* trials could have decreased accuracy either by degrading the representation of the first item (for example, via competition or distraction) or by creating an opportunity for misreporting the wrong item (sometimes termed 'misbinding', see Bays et al., 2009). The application of a mixture model to the response distribution revealed separate effects on both. First, the

precision of recalled responses (as estimated by concentration parameter κ, Fig. 2b) was affected both by cue type ($F_{1,20}$=30.4, p=2.12*10^{-5}) and by load ($F_{1,20}$=7.26, p=0.014), with no significant interaction ($F_{1,20}$=2.20, p=0.153). The effect of load was most apparent in the *protect* condition (1 vs. 2 items: t_{20}=3.37, p=0.003), while precision did not suffer significantly from a preceding item on update trials (t_{20}=0.136, p=0.893). However, since the interaction was not significant, this qualitative difference between *update* and *protect* should be interpreted cautiously. Additionally, precision was lower on 2-item *protect* trials than on 2-item *update* trials (t_{20}=-4.94, p=7.89*10^{-5}). By design, erroneous recall of the incorrect item could only be estimated on 2-item trials. Since the rate of erroneous recall was higher on *protect* trials than on *update* trials (Fig. 2c, t_{20}=3.34, p=0.0032), this also contributed to the overall reduction in accuracy on protect trials. The rate of random guessing (not shown) was only marginally higher on protect trials (2-item trials: t_{20}=1.82, p=0.084, main effect of cue across both loads, $F_{1,20}$=4.43, p=0.048).

Within-Trial Bias Analysis

The next analysis asked whether, in addition to misreporting of items, there was a more subtle effect of the unprobed item on recall (this analysis excluded all 1-item trials). An attractive bias could exist even on trials where the correct item was recalled. Alternatively, the bias could be repulsive. In the presence of a proportion of misreported trials, however, the repulsive bias might be masked by the apparent attraction on

misreported trials. Therefore, the data were analyzed for attractive and repulsive biases both with and without excluding misreporting trials.

The second item had no significant biasing effect on the recall of the first (i.e., on *protect* trials, Fig. 3a, area under the curve: 0.025±0.039 [mean±s.e.m.], t_{20}=0.654, p=0.52). In agreement with this null effect, the amplitude of the fitted Gaussian derivative did not differ from 0 (Fig. 3b, 0.46±0.78°, t_{20}=0.59, p=0.56). On *update* trials, however, responses were significantly biased *away* from the orientation of the first item (area under the curve: -0.089±0.021, t_{20}=-4.16, p=0.0005, amplitude from model fit: -1.54±0.41°, t_{20}=-3.74, p=0.0013). The bias on *update* trials was also significantly different from *protect* trials (comparison of areas: t_{20}=-2.98, p=0.0075, comparison of model amplitudes: t_{20}=-2.52, p=0.020).

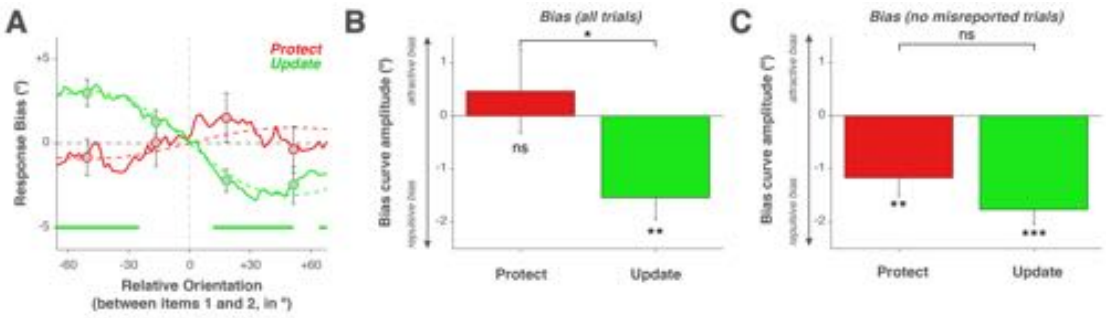


Figure 3. Within-trial biasing effects. A. Bias curve estimating the effect of the relative orientation of the unprobed item on responses, separately for protect and for update trials. Curves (solid lines, red for protect trials, green for update) were created by binning the data into 64 overlapping bins, each containing 1/4 of the data, binned according to the relative orientation between items. In each bin, the average response orientation was calculated relative to the probed item. The grey circles with error bars

(s.e.m.) denote four points along the curve that contain independent data. Points along the curve that differ significantly from 0 (one-sample t-test) are indicated by the colored bar at the bottom of the plot. Dashed lines indicate the fit of the derivative-of-Gaussian. B. Amplitude parameter of the derivative-of-Gaussian fit, separately for protect and update trials. Note that an attractive bias leads to a positive amplitude, whereas a repulsive bias leads to a negative amplitude. C. Amplitude parameters, estimated only on trials where the incorrect item was not misreported.

It is possible that the absence of bias on protect trials was the result of the higher proportion of trials on which the unprobed item was misreported. Therefore, the bias was recalculated after excluding misreporting trials (i.e., using only those trials on which the response was closer to the probed angle than to the non-probed angle). We also excluded trials where the response was closer to an angle with the same relative orientation as the non-probed angle, but in the other direction, in order to avoid any bias in the analysis. Using this method, a repulsive bias emerged for *both* cue conditions (Fig. 3c): *update* trials still showed repulsion (area: 0.104±0.016, t_{20}=-6.65, p=1.80*10^{-6}, amplitude: -1.77±0.28°, t_{20}=-6.41, p=3.00*10^{-6}), but now *protect* trials also showed repulsion (area: -0.070±0.020, t_{20}=-3.59, p=0.0018, amplitude: -1.17±0.35°, t_{20}=-3.30, p=0.004). There was no longer a significant difference in bias magnitude between conditions (comparison of areas: t_{20}=-1.39, p=0.181, comparison of model amplitudes: t_{20}=-1.36, p=0.190).

Therefore, once misreporting of the incorrect item has been taken into account, both conditions show a repulsive bias.

Serial-Bias Analysis

While the two items encoded on the same trial had a repulsive effect on each other, previous studies led to the hypothesis that items probed on *previous* trials should have an attractive effect. This was confirmed in the analyses here, for both conditions (Figure 4a). The bias was significantly positive for both *protect* (Fig. 4b, area: 0.330±0.095, t_{20}=3.47, p=0.002, amplitude: 3.24±0.791°, t_{20}=4.10, p=5.54*10^{-4}) and for *update* trials (area: 0.148±0.025, t_{20}=6.04, p=6.59*10^{-6}, amplitude: 1.609±0.226°, t_{20}=7.12, p=6.76*10^{-7}). *Protect* trials showed a stronger attractive bias than *update* trials (t_{20}=2.34, p=0.030). Once again, the attractive bias might have been caused (in part) by misreporting of the previous trial's orientation. To account for this, the analysis was re-run after excluding misreporting trials. The attractive effect in both conditions was still significant even after excluding misreported trials on which the response was closer to the previous trial's orientation than to the probed orientation (Fig. 4c*, protect trials*: area: 0.087±0.034, t_{20}=2.55, p=0.019, amplitude: 0.806±0.322°, t_{20}=2.50, p=0.021; *update trials*: area: 0.053±0.023, t_{20}=2.30, p=0.032, amplitude: 0.700±0.224°, t_{20}=3.12, p=0.0054). Here, *protect* and *update* trials no longer differed significantly (t_{20}=0.30, p=0.767). Therefore, the initially observed increased bias on

protect trials was likely caused by an increased rate of misreporting the previous trial's orientation.

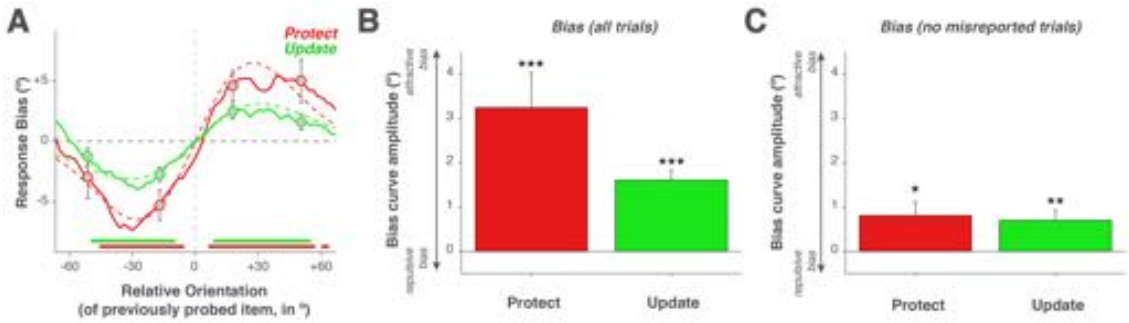


Figure 4. Serial bias across trials. A. Bias curve estimating the effect of the relative orientation of the item probed on the preceding trial, separately for protect and for update trials. Conventions are the same as in Figure 3. Points along the curve that differ significantly from 0 (one-sample t-test) are indicated by the colored bar at the bottom of the plot. Dashed lines indicate the fit of the derivative-of-Gaussian. B. Amplitude parameter of the derivative-of-Gaussian fit. C. Amplitude parameters, estimated only on trials where the incorrect item was not misreported.

MRI Results

The primary aim of the MRI experiment was the localization of areas that are activated more in response to one cue type than the other. In order to test for this, we contrasted BOLD responses on protect trials with those on update trials (using only load-2 conditions).

A cingulo-opercular set of areas responded more strongly to protect cues. This network consisted of bilateral anterior insula and frontal operculum, as well as the dorsal anterior cingulate, stretching into the pre-supplementary motor area (see Fig. 5a and Table 1). By contrast, update trials led to greater activation primarily in bilateral extrastriate cortex (inferior occipital gyrus, stretching into inferior temporal gyrus, consistent with the lateral occipital cortex, LOC), along with bilateral increases along the border between the superior and the medial frontal gyrus (Fig. 5b and Table 1). Responses in the superior frontal region were negative (see bar plot in Fig. 5b), so the difference between conditions is perhaps best interpreted as a relatively reduced deactivation on update trials.

In addition to cue-specific responses, we were interested in more general responses when any kind of WM control was called for. To get an estimate of this, both protect and update trials were compared with the 1-item control trials (Load 2 > Load 1), where neither updating nor protection was required. A conjunction analysis was used to identify voxels showing greater activity for load 2 in both conditions (Load 2 > Load 1 for both *protect* and *update*). Among other areas (see Table 1),

this yielded activation in a bilateral attention-related network spanning the intraparietal sulcus, putative FEF, and ventral visual areas (Fig. 5c).

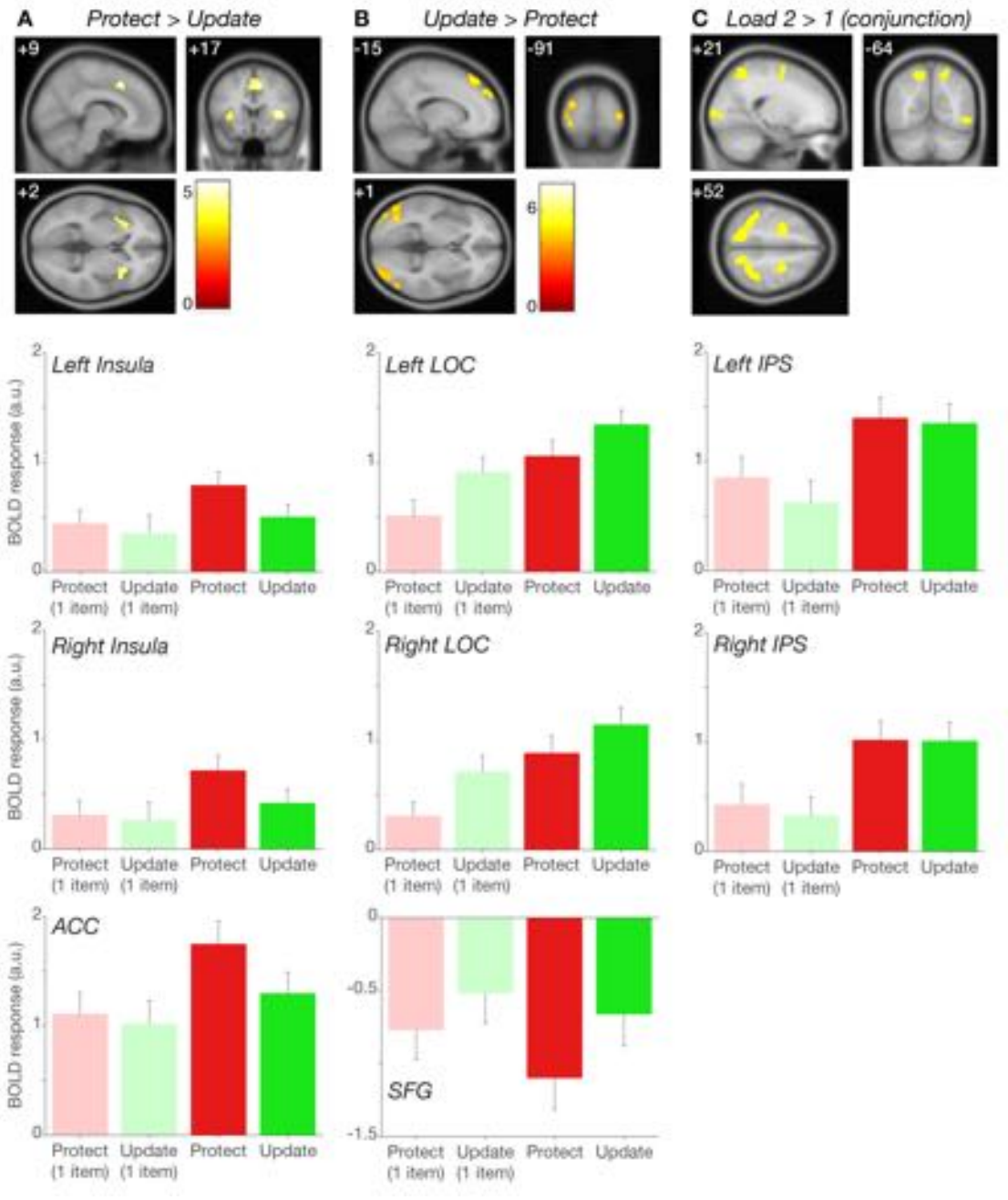


Figure 5. MRI results. T-statistic maps for group-level contrasts of interest are shown in sagittal, coronal, and axial slices (inset: MNI coordinates in the x, y, and z directions, respectively). A. Top row: Clusters showing greater activation after protect than update cues (cluster-corrected

p<0.05). Color scale indicates t values. Bottom rows show bar plots illustrating the regression strengths of each of the four conditions, for the three significant clusters. Regression weights (betas) were extracted from and averaged over all significant voxels in a cluster (p<0.001, uncorrected). The primary purpose of the bar plots is to indicate whether responses were positive or negative for each condition. Error bars indicate s.e.m. B. Regions showing greater activation for update than protect cues. Same conventions as in A. C. Regions showing greater activation for either update or protect trials (compared to 1-item control conditions, conjunction across protect and update conditions). All clusters were significant at a cluster-level family-wise error rate of 0.05, with a voxel-level cluster-forming threshold of p<0.0005 and a minimum cluster size of 25 voxels.

Protect > Update						
P_{corr}	cluster size	peak Z	X	Y	Z	label
0.018	64	3.93	-36	11	7	L insula / frontal operculum
0.015	67	4.19	33	17	10	R insula / frontal operculum
0.004	90	4.08	9	17	43	anterior cingulate / paracingulate cortex
Update > Protect						
<0.001	241	5.02	-48	-76	-8	L lateral occipital cortex
<0.001	216	4.57	42	-79	-8	R inferior occipital gyrus / lateral occipital cortex
0.040	51	4.71	-12	62	31	L superior frontal gyrus / frontal pole
0.006	84	4.17	-15	35	52	L superior frontal gyrus / frontal pole
0.010	74	4.97	12	62	28	R superior frontal gyrus / frontal pole
Load 2 > Load 1						

<0.001	342	4.70	-36	-49	52	L superior parietal lobule (intraparietal sulcus)
<0.001	317	5.22	36	-43	49	R superior parietal lobule (intraparietal sulcus)
0.002	141	5.70	-24	-4	52	L superior frontal gyrus (FEF)
0.007	109	5.34	27	-4	55	R superior / middle frontal gyrus (FEF)
<0.001	327	6.16	-27	-94	4	L middle occipital gyrus
0.028	79	4.27	-48	-79	-2	L middle occipital gyrus
<0.001	313	5.08	21	-97	1	R middle occipital gyrus

Table 1. Peak activation coordinates for contrasts of interest in the MRI experiment. P-values indicate family-wise corrected values (cluster level). Peak coordinates are reported in MNI space. Anatomical labels are taken from the Harvard-Oxford Anatomical atlas, as implemented in FSL.

The MRI comparisons yielded three primary networks of interest: (1) a cingulo-opercular network that was more activated on protect trials, (2) visual and superior frontal areas that were more active on update trials, and (3) attention-related areas that were active for both. We hypothesized that source-space estimates of time-frequency responses from these three networks would yield similar condition differences in the MEG experiment. To test this, the peak coordinates from the MRI study were used as regions-of-interest in the MEG source-space analysis.

MEG Event-related and Time-Frequency responses to update and protect cues

After finding a network of regions involved in WM control in the MRI experiment, the primary purpose of the MEG experiment was to investigate the timing of WM control. In a first step, this involved investigating the event-related responses to protect and update cues in sensor space. Event-related fields locked to the onset of the cue (and second stimulus) were computed at each gradiometer pair and were compared between protect and update trials. Early visual responses were remarkably similar (Fig. 6a&b), and did not yield any significant differences (one non-significant cluster ($p=0.476$) from 266-296 ms, using paired t-tests over visual sensors and permutation testing to correct for multiple time-points). Given that topographies were imperfectly aligned between participants (and there may have been individual differences in topographies), the same sensors may have picked up a different signal in different participants, reducing the sensitivity of comparisons.

To account for this, in a next step we computed a difference metric at the within-participant level that calculated the overall dissimilarity in the ERF (at all sensors) between protect and update cues, and pooled this difference metric across participants. Because conditions were compared at the within-participant level, misalignment of the topographies between participants should have no influence on the results. This approach also reduced the number of comparisons (because it was not done separately at each sensor), increasing its sensitivity. Here we found that strong

differences between update and protect cues appeared, but only relatively late in the delay (Fig. 6c, 872-1,494 ms after cue onset, corrected p=0.002, with an additional later cluster from 1,593-1,838 ms, p=0.041, and a non-significant cluster at a slightly earlier time point, 672-766 ms, p=0.293).

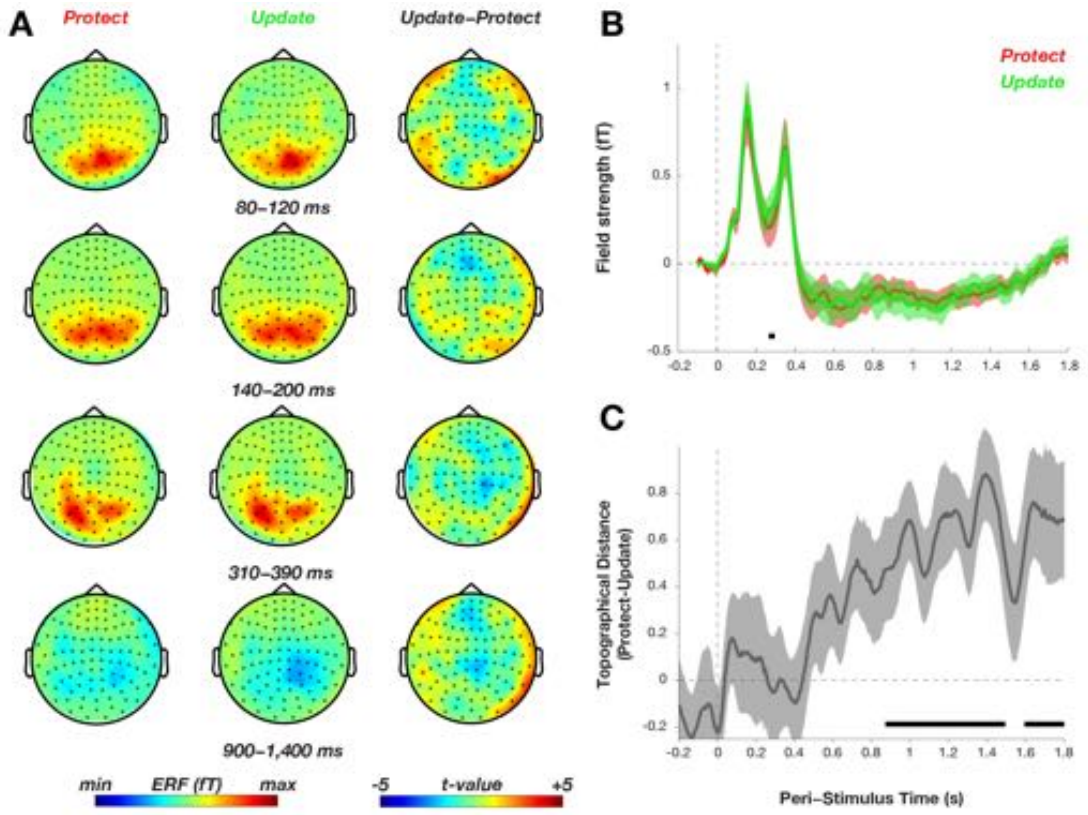


Figure 6. Sensor-space responses indicate late differences between Protect and Update cues. A. Topographies after Protect and Update cues, and difference maps between conditions, for peaks in the visual event-related response (80-120 ms, 140-200 ms, 310-390 ms, and one late time window from 900-1,400 ms, when the topographic distance analysis indicated differences in the ERF between cues). Colors indicate field strength in the left and middle column, and t-statistics in the right

column. B. The event-related response at visual sensors did not differ significantly between conditions (one short window of significant difference between 200 and 300 ms is indicated by the black bar, but this is not significant when correcting for multiple comparisons ($p>0.476$). Shading indicates standard error of the mean. C. By contrast, a topographic distance analysis (shuffle-corrected, z-scored Mahalanobis distance between condition means) indicates significant differences emerging relatively late (from 800 ms), after the visual response has returned to below baseline. Black bars indicate time windows of significant separation between protect and update cues (cluster-corrected $p<0.05$).

Considering the MRI results, the late differences between cue types should be visible in the source-space signals from the regions of interest (see Fig. 7a for location of ROIs). Because of the inherent ambiguity in dipole orientation when inferring source-localized activity, here we used time-frequency representations of the timecourses in each ROI, and compared update and protect trials early (0-500 ms) and late (500-1,500 ms) after cue onset, and at different frequencies (theta: 4-8 Hz, alpha: 8-14 Hz, and beta: 15-30 Hz). As can be seen from the reconstructed responses for each ROI, responses were again remarkably similar (Fig. 7b). In each region and for each frequency band, timecourses in response to update and protect cues were nearly identical (Fig. 7c), yielding no significant differences in either early or late time windows for

any of the frequency bands (all p>0.05, uncorrected). One notable exception appeared in the early (0-500 ms) theta-band response localized to the ACC region of interest: here, power was significantly lower on protect trials than on update trials (t_{17}=-3.18, p=0.006). However, this effect did not hold up to multiple comparisons correction ($p_{Bonferroni}$=0.252).

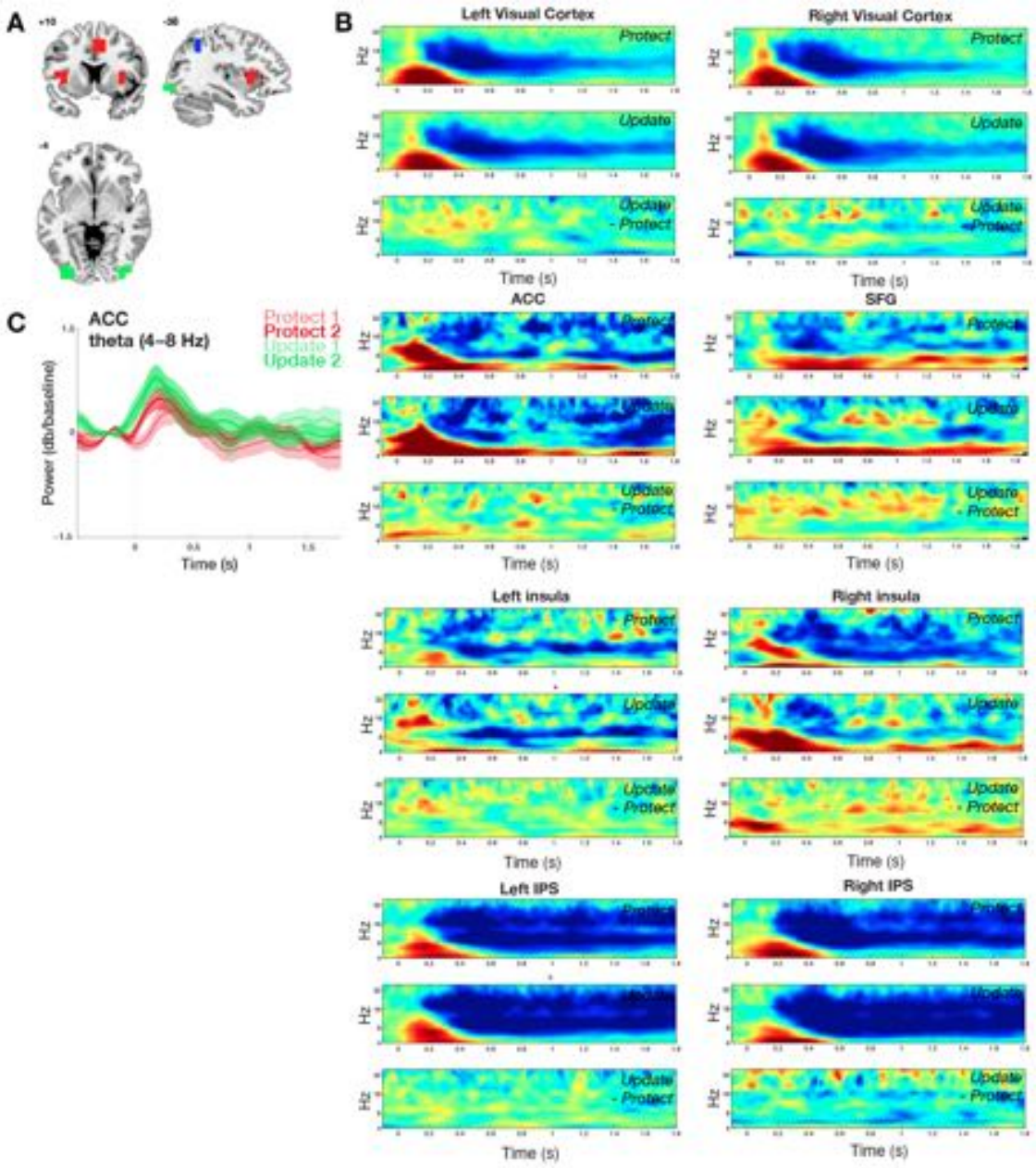


Figure 7. Source-space analysis does not differentiate between cue conditions. A. Location of ROIs used for source-space analysis. Colors denote ROIs derived from different contrasts in the MRI experiment, with red ROIS in the anterior cingulate and bilateral insula derived from the Protect>Update contrast, green ROIs in visual areas (and right superior frontal gyrus, not shown) from the Update>Protect contrast, and blue ROIs in bilateral inferior parietal lobule from the Load 2> Load 1 contrast. B. Timecourses from the ROIs were used for time-frequency analysis (4-40 Hz), log-transformed, and baseline-corrected, separately for Protect and Update trials. TFR maps show conditions separately, relative to baseline, in units of decibels, and the difference between conditions, in t-values (color range from -5 to +5). None of the contrasts showed significant differences between cue types. C. There was a trend towards an increased theta-band (4-8 Hz) response in the anterior cingulate ROI for Update trials, compared to Protect trials. However, this difference did not survive correction for multiple comparisons. Shading indicates standard error of the mean.

Stimulus decoding and decoding the contents of working memory

In a second step, the MEG data were used to investigate whether WM control had an effect on the stimulus information available in the neural response. To this end, the orientation of the stimulus presented together with the cue was read out from the MEG response on a trial-by-trial level (by comparing the topography at each timepoint on each trial to

the topographies of all other trials with similar vs. orthogonal stimulus orientations, and calculating the difference in dissimilarities, see Fig. 8a). Trials were then averaged according to cue type, yielding separate orientation decoding timecourses for protect and update trials.

The initial visual response contained significant information about the stimulus, irrespective of cue type (Fig. 8b, update cues: 64-1,570 ms, cluster $p<0.0001$, protect cues: 78-686 ms, cluster $p<0.0001$). Importantly, the decoding returned to baseline levels after protect cues (where the stimulus did not need to be encoded in WM), while it was sustained into the maintenance delay after update cues. When comparing decoding timecourses between conditions, update cues showed better decoding in a relatively late period (1,104-1,230 ms, cluster $p=0.053$). This late difference in decoding corresponded well with the late differentiation of the two cue types (Fig. 6c). When collapsing over the time window in which the cues could be differentiated (Fig. 6c, 872-1,494 ms), the difference in stimulus decoding was also significant ($t_{19}=2.60$, $p=0.017$). In this time window, decoding was significant on update trials ($t_{19}=4.11$, $p=6.0*10^{-4}$), but not on protect trials ($t_{19}=1.85$, $p=0.081$).

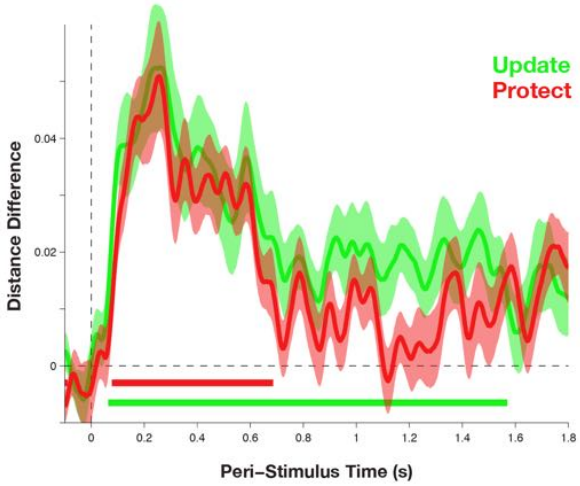


Figure 8. Stimulus decoding in the delay depends on cue type. Timecourses show trial-wise stimulus orientation decoding of the second stimulus, separately for Protect and Update trials. When the second stimulus was task-relevant, stimulus decoding remained significant well into the memory delay, whereas it returned to baseline on Protect trials, where the second stimulus acted as a distractor. Shading indicates standard error of the mean. Bars denote time windows of significant decoding (cluster-corrected) for the respective cue type.

Discussion

This study contrasted two WM control operations: updating and protecting the contents of WM. The main aim of the study was to identify brain regions that were more involved in one control operation than the other, and to examine the timecourse of these two kinds of control. In the fMRI experiment, cues instructing observers to protect the contents of WM and to ignore the current stimulus led to activation of the anterior insula and the dorsal anterior cingulate, whereas cues prompting observers to replace the WM contents with the current stimulus led to activation in extrastriate visual areas and the superior frontal gyrus. In neither case was the dorsolateral PFC selectively activated. MEG suggested that neural responses to the two cue types differed only fairly late, from approximately 1 s after the onset of the cue and stimulus. However, source-space analysis could not reproduce the fMRI results. Finally, decoding of the presented stimulus also differed between the two stimulus types in this late window, but not earlier. Despite these somewhat equivocal results, we tentatively suggest that 1) WM control comprises multiple functions that may engage different cortical networks, and that control cues presented with a stimulus need not have early effects on stimulus encoding, but can also influence WM in the delay period.

In this design, protecting the contents of WM may have been equivalent to directing internal attention to the stimulus that was already in WM. Protect cues, therefore, may have acted much like retrocues to

prioritize one item whose behavioral relevance has suddenly increased. While the fronto-opercular network that responded to protect cues is not typically seen in WM maintenance tasks, it becomes engaged in response to attention-directing cues presented during the maintenance delay (Nobre et al., 2004; Higo et al., 2011; Wallis et al., 2015; Nee and Jonides, 2009). This reinforces the possibility that protect cues here acted much like retrocues, prompting participants to bring the previously encoded item back into the focus of attention. This link is not typically made in other paradigms discussing the protection of WM interference. Instead, most models (Suzuki and Gottlieb, 2013; Jacob and Nieder, 2014) assume that WM is protected by inhibiting the processing of irrelevant stimuli to a high level. However, as mentioned in Chapter 1, early electrophysiological studies already showed that distractors do influence spike rates in PFC (Miller et al., 1996), albeit transiently. It is possible that interference is resolved in part by directing attention back to the relevant WM item, which automatically leads to a decay of information about the distractor. Simultaneously, moving an encoded item into the focus of attention may help position it so it can guide action during the next probe.

Compared to protect cues, updating led to activation in lateral extrastriate visual cortex. A probabilistic coordinate atlas indicated that it was localized in the lateral occipito-temporal cortex (Kourtzi and Kanwisher, 2000). An increased visual response seems reasonable given that the stimulus presented concurrently with the update cue needed to be encoded into WM and therefore demanded increased visual

processing. However, only relatively late stages in the visual processing hierarchy showed differences in activation, with no noticeable increase in early visual areas V1 to V3. This begs the question of what additional processing occurred in LOC. Given that differences in the MEG signal between update and protect trials emerged only late (or could only then be detected reliably), after the visual response had returned to baseline, it seems likely that the increased activity in visual cortex did not reflect increased initial processing of the visual stimulus itself, but rather reflected a process in the maintenance period. For instance, the increase in visual activity could reflect feedback-related responses driven by areas involved in top-down control. Whether the superior frontal gyrus was involved in this delay-period control over visual cortex remains to be examined further, because the SFG has not received much attention in the WM literature thus far and its function is unclear. Some researchers have suggested a role for anterior PFC in gating items into WM (Burgess et al., 2007), but this interpretation is made problematic in the current study because SFG responses were negative.

In addition to showing only a late dissociation between conditions, the MEG experiment further identified that stimulus orientation decoding during the early visual response was equivalent on protect and update trials. The quality of stimulus decoding differed only relatively late, again approximately 1 s after stimulus onset, well within the maintenance delay. It is possible that the increased activity in visual cortex reflected the reactivation or continued maintenance of new stimulus information.

In spite of the clear differentiation of conditions at the whole-sensor level, source-space signals extracted from the regions activated in the MRI experiment did not differentiate between cue conditions. A number of reasons could explain this null finding. Firstly, inaccuracy in the source localization due to spatial ambiguity or excessive head motion may have reduced the signal-to-noise ratio. Further, the frequency range included here (4-40 Hz) may have excluded signals relevant to the task. For one, high-frequency power (above 40 Hz) could have differed between update and protect trials, especially during encoding of a large visual stimulus, which is known to drive high-frequency synchronization in visual cortex (Gray et al., 1989; Fries et al., 2001). However, high-frequency synchronization typically requires sustained stimulus drive. Since stimulus presentation was relatively brief here (200 ms), high-frequency synchronization should have subsided quickly after stimulus offset. Given that the sensor-space data indicated only late differences, it is more likely that low-frequency fluctuations (below 4 Hz) were cue-selective. In more rapid visual updating tasks, delta-band (1-4 Hz) fluctuations have been shown to influence the quality of updating (Wyart et al., 2012). Finally, cue differences may not have driven changes in power in any frequency band. For example, temporary changes in inter-regional coherence could be sufficient to change a task network so that new information is routed into WM (or prevented from entering it). For example, Liebe and colleagues demonstrated changes in visual-prefrontal theta coherence during WM maintenance that corresponded with

behavioral success (Liebe et al., 2012). This could increase synaptic and spiking activity in both areas (and thus lead to increases in BOLD signal) without requiring increases in oscillatory power.

In addition to the categorical, cue-instructed effects of updating, the continuous recall measure made it possible to examine more graded effects of WM updating. In particular, on update trials the previously remembered (and replaced) orientation had a repulsive effect on recall of the updated item. This repulsion may have occurred to prevent accidental merging of the two representations. The most parsimonious interpretation of the neuroimaging results is that participants, on most trials, encoded the second item into WM and only then processed the cue to determine which item was relevant. If this strategy led to a temporary maintenance of both items in the same WM store, they may have interfered with each other, leading to a repulsion effect. Congruent with this interpretation is that the repulsion effect was also visible on protect trials (after misreporting had been taken into account). Nevertheless, some alternatives cannot be ruled out. For one, the preceding stimulus may have simply led to repulsion during initial visual processing. Similar effects are observed in the tilt after-effect (Gibson, 1937; Knapen et al., 2010). However, the tilt after-effect typically requires the first stimulus to be presented for much longer (at least 1 s) than it was here, making it less likely that this is the sole account of the repulsion. A second alternative is that active inhibition of the irrelevant item led to a suppression of neurons coding for that part of orientation space, causing repulsion. This has

recently been proposed as an explanation of repulsion effects following recall of items from WM (Kang and Choi, 2015).

In contrast to the repulsive effects within a trial, the item probed on the preceding trial had an *attractive* effect on recall. Since the biasing effect went in opposite directions within and between trials, it seems likely that this attractive bias is mediated by a different mechanism. One important factor appeared to be misreporting, given that excluding misreported trials reduced the attractive bias. A trivial explanation of the attractive effect could therefore be that the previous item has no effect on the current trial, unless the current item is forgotten and is therefore replaced during recall by the previous one. However, this explanation does not account for the entire phenomenon. Since the bias was still significant (and still attractive) even after excluding misreported trials, the previous item appears to have an attractive effect on the maintenance or recall of the current item even if it is remembered with high accuracy. It will be interesting to investigate whether within-trial and between-trial biases in updating have distinct neural correlates, and whether this distinction can help identify the multiplicity of WM-related information stores in the brain.

In conclusion, this study examined the neural basis of two simple control operations in WM: updating and protecting its contents. The operations drew on distinct brain networks, demonstrating that WM control is distributed across many cortical areas and highly dependent on the specific control operation. Future studies that draw clear distinctions

between different kinds of WM updating will help delineate the various forms of WM maintenance and how they are flexibly adapted to the current task.

Chapter 7: General Discussion

This thesis has examined some of the many ways in which humans can exert flexible control over working memory[3]. Flexibility allows us to structure action and cognition efficiently, and to react to salient news without losing track of what we were doing before. One contribution of this thesis has been the emphasis on the transient dynamics of control, both in modulating representations in sensory cortex and in affecting behavior. This may prompt a reevaluation of the temporal aspects of top-down control in memory representation using neuroimaging methods with high temporal resolution. These methods could help provide insight into the benefits of flexibility at all stages: encoding, maintenance, updating, and recall of items in WM.

Flexible and transient WM control operations during encoding, maintenance, and recall

Prior to, and during, encoding, modifications to the fidelity of WM items and the likelihood of remembering them likely depend on transient modulations in sensory cortex. Chapter 3 focused on how attention-directing cues desynchronize alpha oscillations over posterior sites and lead to enhanced visual responses to WM items that are more relevant to behavior. Not surprisingly, these effects mirror those of spatial attention

[3] Parts of this chapter (the subsection 'Within-trial effects of item similarity') relate to a previous publication: Myers N, Wallis G (2013) Constraining Theories of Working Memory with Biophysical Modeling. J Neurosci 33:385–386.

on visual perception. However, in many contexts, we are not forewarned about the behavioral relevance of upcoming information. As Chapter 6 showed, when an item turns out to be relevant to the task and therefore needs to be encoded in WM, this leads to increases in visual cortical activity (predominantly in LOC). The complementary MEG data indicated that this increase in visual activity may occur fairly late after stimulus presentation (although additional early differences could have occurred without being detected by MEG), and therefore could reflect a consolidation process rather than differences in the initial processing of the stimulus itself.

As is the case with improving perception, the prior state of excitability in sensory cortex may also be a good predictor of successful encoding into WM. In chapter 4, I showed that spontaneous variability in visual alpha-band synchronization and in alpha phase preceding the onset of the stimulus predicts the fidelity with which items are later recalled from WM. This could be one neurophysiological basis, along with slower fluctuations in alertness or motivation (Sadaghiani et al., 2009), that contributes to the observation that working memory capacity and fidelity fluctuate considerably from trial to trial, even within single observers. My thesis contributes to a growing acknowledgement that memory ability is not a static attribute, but that it can vary rapidly according to many deliberate and spontaneous factors.

One of the deliberate factors that can seemingly increase working-memory capacity is the cueing of relevant items during maintenance.

Chapter 3 showed that these cues also lead to desynchronization of alpha band oscillations at posterior sites that are presumably involved in storing the cued item. While the consequence of cueing in sensory cortex appears to be similar to prospective cueing, the control networks involved may differ substantially. For example, chapter 6 showed that instructions directing attention back to a remembered item activated the anterior insula and anterior cingulate, two regions that are not typically seen as controllers of top-down attention for perception. Likewise, other studies have found that retrospective cues (Nobre et al., 2004; Nee and Jonides, 2009; Wallis et al., 2015) and feature-based cues (Bollinger et al., 2010; Zanto et al., 2011) that are used in the service of WM activate frontal brain areas beyond the typical spatial attention network (Nobre et al., 2000; Corbetta and Shulman, 2002).

Furthermore, even sensory effects of retrospective cues seem to differ upon closer inspection. For example, the alpha-band lateralization observed in chapter 3 was only transient. After a distracting stimulus abolished the lateralization, it did not return until shortly before the expected onset of the memory probe. This could be a hint that sustained top-down attention is not necessary for a flexible re-prioritization event to take place in WM. Instead, once the representational change of an item has been completed, attention may be diverted to other tasks until the prioritized item is finally needed. Behavioral evidence from chapter 2 belongs to a growing body of evidence confirming this. A recent proposal has argued that cognition can be thought of as a temporally segregated

stream of attentional episodes (Duncan, 2013). In this framework, flexibility in WM might take the form of branching control (Atkinson and Shiffrin, 1968; Norman and Shallice, 1986; Badre, 2008), ensuring that information which will become important later is laid down in a format that can be accessed when needed, after which attention can be directed to the control of an intervening task.

While this thesis did not look at flexibility in preparing to inhibit stimuli from cluttering WM, the experiments in Chapter 6 did involve some degree of distractor inhibition when attention was directed back to the contents of WM. Since neural responses were selective mainly in areas associated with retrospective cueing, it seems plausible that some tasks that purport to measure resistance to interference could be reinterpreted as orienting attention to information in WM. It will be interesting to see whether future work can disentangle the separate notions of *actively* inhibiting the processing of irrelevant stimuli vs. focusing attention on those stimuli that are relevant (even when those stimuli are already encoded in memory). Again, an examination of the temporal dynamics of engaging versus disengaging attention to retrospectively cued versus inhibited items could be illuminating.

Transient reactivation of silent WM representations

Methods with high temporal resolution may also become useful in examining the neural basis of representation in WM. Recent work has already shown that pattern analysis of EEG signals can potentially be

used to dissociate silent (or passive) from active (or attended) representational states in WM (LaRocque et al., 2013). Broadly in line with this view are the findings from Chapter 5. There, I used pattern analysis of MEG data to show that template orientations held in WM can be decoded only transiently, in a brief time window around the onset time of comparison stimuli. Two alternative mechanisms could account for this phenomenon. Both are consistent with the flexible reactivation of WM items when they are needed for behavior. The first possibility is that WM templates are reactivated around the time when a comparison stimulus is expected. Such reactivation based on temporal expectation could be tested by occasionally skipping comparison stimuli and testing whether the template pattern re-emerges in the absence of a stimulus. Alternatively, the onset of the comparison stimulus could provide bottom-up drive into the system, causing a neural response that is patterned by the silently maintained WM template (Stokes, 2015). The task in Chapter 5 did not allow a dissociation of the two possibilities. However, in line with the latter, a recent study has found that decoding of WM contents from EEG improves after presentation of a task-irrelevant driving stimulus 'probes' the hidden state of the system (Wolff et al., in press). Beyond the possibility that WM contents may be represented in a silent code whenever possible, it will be interesting to test under what conditions new stimuli merely drive the system to respond as a matched filter reflecting the WM content, and when new stimuli lead to a reconfiguration of the hidden state itself (D'Esposito and Postle, 2015; Postle, 2015).

Cooperation of working memory with other memory systems

The thesis so far has focused on working memory as a system that maintains information about distinct objects, in isolation. This definition implies that item representations do not interact with one another. In fact, however, they do. Moreover, there is a wide range of influences that bias the representation of single items in memory. The most prominent sources of interference are other items remembered at the same time (which can be misreported outright or create an ensemble statistic or prototype), and other items presented in the (recent) past (by creating proactive interference, and a long-run average). While most WM research has seen these biases as sources of noise and has tried to minimize their influence (e.g., Luck, 2008; Lin, 2012), their prevalence might actually be an indication that they serve an adaptive purpose.

Trial-history effects on visual WM

Verbal working-memory researchers have long remarked on the influence of preceding trials on current performance (Keppel and Underwood, 1962). Words which appeared on recent trials, but not on the current one, have an intrusive effect, and create proactive interference when they are used as negative probes (Atkinson and Juola, 1974; Monsell, 1978). Nee and Jonides (2006) have argued that proactive interference is a by-product of lingering activation of recent stimuli, which is generally detrimental to performance. As a result, whenever a recently

(but not currently) remembered probe appears, the latent recognition must be suppressed by the ventrolateral prefrontal cortex (Nee et al., 2007; Nee and Jonides, 2008). An initial aim of research on visual working memory was to estimate capacity without the possible contaminating effects of trial history (Luck, 2008). In order to prevent the use of a sense of recency in memory judgments, Luck and Vogel (1997) initially used very simple colored shapes. These had the advantage, compared to images of real-world objects, of carrying few associations, and presumably creating no long-lasting activations in long-term memory that might contribute to proactive interference on future trials. By repeating them in different configurations across trials, they also eliminated recency as an alternative strategy.

However, this design choice ignored the possibility that recent stimuli might still have an influence. A number of studies have begun to measure this effect. They generally presume that, under normal circumstances, recent stimuli are predictive of current stimuli because our environments tend to be temporally autocorrelated: much recent research on the effects of sensory expectations in perception is predicated on this observation (Summerfield and de Lange, 2014). Therefore, keeping recent stimuli in an activated state that biases the current memory item could be adaptive. This assumption was initially confirmed by Huttenlocher and colleagues (Huttenlocher et al., 2000). The authors found that delayed recall estimates of simple visual features (stimulus width, line length, or luminance) were biased towards the mean of the distribution of stimuli in

the preceding block. Importantly, they were able to rule out that this effect was simply regression to the mean. Since then, this effect has been replicated for spatial frequency memory (Huang and Sekuler, 2010), and for orientation judgments (Sims et al., 2012). Two recent studies additionally found that orientation recall is biased towards the tested orientation on the preceding trial (Fischer and Whitney, 2014), and that delayed saccades are biased towards the previous saccade target (Papadimitriou et al., 2015), demonstrating that individual items, and not just the long-run average statistics, can exert a bias.

Within-trial effects of item similarity

Similarly, when multiple items are remembered concurrently, they may influence each other. The most replicated example of this are swap errors in delayed estimation paradigms: on some proportion of trials, observers mistakenly report the feature of an unprobed item from the same trial (Bays et al., 2009; Ma et al., 2014). However, items may have a more subtle influence, where they simply bias feature representations toward one another, especially if features are similar across items (Mate and Baqués, 2009; Huang and Sekuler, 2010; Sanocki and Sulman, 2011).

Biophysical models of working memory that propose the storage of multiple items in a single neural ‘resource’ pool predict that this attractive effect occurs when two bump attractors representing two distinct but similar features merge, creating one wider attractor that peaks

at a feature value somewhere between the two original ones (Johnson et al., 2009; Wei et al., 2012). While elegant, these models may provide an incomplete picture. For one, in most visual WM paradigms, stimuli are presented and encoded in spatially distinct locations, leading to their representation in retinotopically distinct brain areas. Most biophysical models neglect the simple possibility that different items might be represented in non-overlapping cortical areas, making it questionable whether a single resource pool model could ever provide a realistic account of the representation of multiple items (Myers and Wallis, 2013). As mentioned in Chapter 2, one exception might be when two or more items crowd into the limited-capacity focus of attention, but this has not been tested. More recent studies have attempted to acknowledge this relative independence of retinotopic areas while still accounting for capacity limits in working memory when items are sufficiently far apart (Bays, 2014; 2015; Matthey et al., 2015; Orhan and Ma, 2015). One model has solved this contradiction by proposing that visual cortex, across a wide, retinotopically distinct territory, is normalized via lateral inhibitory connections so that the summed activity in the entire area remains fairly constant (Bays, 2014). While this kind of normalization has been observed at smaller spatial scales in many brain areas (Carandini and Heeger, 2012), it remains to be seen whether it can help account for capacity limits in working memory.

Returning to behavioral effects, one study found that similar colors were remembered with higher accuracy, and argued that this was

compatible with a mutual repulsion mechanism (Lin and Luck, 2009), rather than by mutual attraction. By repelling each other, similar colors in WM were not only more distinguishable from each other, but also from the similar (but mismatching) test color, leading to improved performance. However, the authors did not decisively rule out that mutual attraction might lead to improvements in their task because they tested only for effects on discrimination accuracy, rather than on bias itself.

The present thesis also found evidence of mutual repulsion between concurrently processed items (in Chapter 6). This finding seems to contradict the mutual attraction effects seen in other studies. One possibility is that repulsion occurred because items were presented in the same retinotopic location, so that normalization effects created repulsion rather than attraction. However, a similar effect was seen with respect to a retrocued item in Chapter 2, even though here items were presented in distinct locations. Here, the capacity-limited focus of attention might be the cause of the repulsive effect. Future studies that systematically vary the retinotopic distance and behavioral relevance of items in WM could help map out when items attract each other, and when they push each other apart.

Going beyond mutual interactions between individual items, a number of studies has found that WM items are not only encoded individually, but also as an ensemble (Brady et al., 2011). When large numbers of items need to be remembered, recall of individual items is biased towards the average values of the entire array (Brady and Alvarez,

2011). Ensemble statistics seem to be represented distinctly from individual items (Brady and Tenenbaum, 2013). It might be revealing to see whether prospective and retrospective cueing can influence the contribution of an item to the ensemble statistic, or conversely make recall of that item less dependent on the average of other items.

Mutual attraction between memory items, along with improved memory for similar items, is compatible with a hierarchical, Bayesian memory architecture (Huttenlocher et al., 2000; Sims et al., 2012): since single-item representations are noisy, it is optimal to integrate them with all available feature priors. The noisier the single-item representation on a given trial, the more we should draw on the prior. This prediction was confirmed by Sims and colleagues (2012), who showed that the strength of the bias increases with set size (as individual item representations get noisier). If they are not taken into account, these biases may go undetected or simply look like an increase in response variability (noise or random guessing). Conversely, manipulations that improve WM performance, such as retrospective cues or refreshing cues, should lower the bias incurred from priors, since the cued item itself is more reliable. This effect was observed in Chapter 2. Retrocues modulate item priority only after encoding, so this finding further indicates that regularities are not integrated into the individual item representations directly during encoding and maintenance. Instead, biases may be integrated only at the recall stage – this would be further evidence of the high degree of (rapid) flexibility in WM.

In sum, working memory seems to draw on multiple memory systems acting as reservoirs of information about the immediate (and longer-term) past (Brady et al., 2011; Hasson et al., 2015). Depending on the neural system (early visual cortex, mid-level visual areas such as LOC, parietal areas, or dlPFC), different aspects of the scene are represented (single features, feature-bundled objects, ensemble statistics, long-run averages). In different mnemonic networks, memories might decay with different time constants (Bernacchia et al., 2011; Murray et al., 2014) and be more or less sensitive to interference (Jacob and Nieder, 2014). I speculate that the output system driving behavior is responsible for integrating all sources of information and weighting them according to their reliability. The weighting will depend on an estimation of noise in each memory system, but also on an estimation of the task structure and the regularity of the environment.

Of course, understanding how these memory systems relate, and how our conception of working memory can be expanded in scope to accommodate its adaptability, will keep researchers in work for many more decades. My hope is that future work will focus less on the limits of memory, and more on whether these apparent limits might reflect useful adaptations, and how humans can cheat these limits intelligently. Knowing when to remember something, and when to forget it, might be a start.

Bibliography

AEEGS (1991) American Electroencephalographic Society guidelines for standard electrode position nomenclature. J Clin Neurophysiol 8:200–202.

Akam T, Kullmann DM (2014) Oscillatory multiplexing of population codes for selective communication in the mammalian brain. Nat Rev Neurosci 15:111–122.

Allport DA, Styles E, Hsieh S (1994) Shifting intentional set: Exploring the dynamic control of tasks. In: Umiltà, C, and Moscovitch, M (ed.). Attention and performance XV: Conscious and nonconscious information processing. pp. 421-452.

Anderson DE, Vogel EK, Awh E (2011) Precision in Visual Working Memory Reaches a Stable Plateau When Individual Item Limits Are Exceeded. Journal of Neuroscience 31:1128–1138.

Anderson JA (1973) A Theory for the Recognition of Items from Short Memorized Lists. Psychological Review 80:417–438.

Ashburner J, Friston K (1997) Multimodal image coregistration and partitioning—a unified framework. Neuroimage, 6(3):209-217.

Astle DE, Nobre AC, Scerif G (2010) Subliminally Presented and Stored Objects Capture Spatial Attention. Journal of Neuroscience 30:3567–3571.

Astle DE, Summerfield J, Griffin I, Nobre AC (2012) Orienting attention to locations in mental representations. Atten Percept Psychophys 74:146–162.

Astrand E, Ibos G, Duhamel J-R, Ben Hamed S (2015) Differential dynamics of spatial attention, position, and color coding within the parietofrontal network. Journal of Neuroscience 35:3174–3189.

Atkinson RC, Juola JF (1974) Search and decision processes in recognition memory. In Krantz DH, Atkinson RC, Luce RD, Suppes P (ed.): Contemporary developments in mathematical psychology I. Learning, memory, and thinking. Oxford, UK: W. H. Freeman.

Atkinson RC, Shiffrin RM (1968) Human memory: A proposed system and its control processes. In Spence KW, Spence JT: The psychology of learning and motivation: Advances in Research and Theory. Volume 2. New York: Academic Press. pp. 90-197.

Awh E, Anllo-Vento L, Hillyard SA (2000) The role of spatial selective attention in working memory for locations: evidence from event-related potentials. Journal of Cognitive Neuroscience 12:840–847.

Awh E, Jonides J, Reuter-Lorenz PA (1998) Rehearsal in spatial working memory. Journal of Experimental Psychology: Human Perception and Performance 24:780–790.

Awh E, Vogel EK, Oh SH (2006) Interactions between attention and working memory. Neuroscience 139:201–208.

Baddeley A, Della Sala S, Robbins TW (1996) Working Memory and Executive Control [and Discussion]. Philosophical Transactions of the Royal Society B: Biological Sciences 351:1397–1404.

Baddeley A, Logie R, Bressi S, Sala Della S, Spinnler H (1986) Dementia and working memory. Q J Exp Psychol A 38:603–618.
Badre D (2008) Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. Trends in Cognitive Sciences 12:193–200.
Badre D (2012) Opening the gate to working memory. PNAS 109:19878–19879.
Badre D, D'Esposito M (2009) Is the rostro-caudal axis of the frontal lobe hierarchical? Nat Rev Neurosci 10:659–669.
Barbey AK, Koenigs M, Grafman J (2013) Dorsolateral prefrontal contributions to human working memory. Cortex 49:1195–1205.
Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ (2012) Canonical Microcircuits for Predictive Coding. Neuron 76:695–711.
Bastos AM, Vezoli J, Bosman CA, Schoffelen J-M, Oostenveld R, Dowdall JR, De Weerd P, Kennedy H, Fries P (2015) Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. Neuron 85:390–401.
Bays PM (2014) Noise in Neural Populations Accounts for Errors in Working Memory. Journal of Neuroscience 34:3632–3645.
Bays PM (2015) Spikes not slots: noise in neural populations limits working memory. Trends in Cognitive Sciences 19(8):431-438.
Bays PM, Catalao RFG, Husain M (2009) The precision of visual working memory is set by allocation of a shared resource. Journal of Vision 9:7–7.
Bays PM, Gorgoraptis N, Wee N, Marshall L, Husain M (2011) Temporal dynamics of encoding, storage, and reallocation of visual working memory. Journal of Vision 11:6–6.
Bays PM, Husain M (2008) Dynamic Shifts of Limited Working Memory Resources in Human Vision. Science 321:851–854.
Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, Shadlen MN, Latham PE, Pouget A (2008) Probabilistic Population Codes for Bayesian Decision Making. Neuron 60:1142–1152.
Becke A, Müller N, Vellage A, Schoenfeld MA, Hopf J-M (2015) Neural sources of visual working memory maintenance in human parietal and ventral extrastriate visual cortex. NeuroImage 110:78–86.
Beckstead RM, Domesick VB, Nauta WJ (1979) Efferent connections of the substantia nigra and ventral tegmental area in the rat. Brain Research 175:191–217.
Benoit RG, Gilbert SJ, Frith CD, Burgess PW (2012) Rostral Prefrontal Cortex and the Focus of Attention in Prospective Memory. Cerebral Cortex 22:1876–1886.
Bernacchia A, Seo H, Lee D, Wang X-J (2011) A reservoir of time constants for memory traces in cortical neurons. Nature Neuroscience 14:366–372.
Berryhill ME, Richmond LL, Shay CS, Olson IR (2012) Shifting attention among working memory representations: Testing cue type, awareness, and strategic control. The Quarterly Journal of

Experimental Psychology 65:426–438.
Bianchi L (1895) The Functions of the Frontal Lobes. Brain 18:497–522.
Bisley JW, Goldberg ME (2010) Attention, Intention, and Priority in the Parietal Lobe. Annu Rev Neurosci 33:1–21.
Bledowski C, Rahm B, Rowe JB (2009) What "Works" in Working Memory? Separate Systems for Selection and Updating of Critical Information. Journal of Neuroscience 29:13735–13741.
Bliss TV, Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. Nature 361:31–39.
Bollimunta A, Mo J, Schroeder CE, Ding M (2011) Neuronal mechanisms and attentional modulation of corticothalamic α oscillations. Journal of Neuroscience 31:4935–4943.
Bollinger J, Rubens MT, Zanto TP, Gazzaley A (2010) Expectation-driven changes in cortical functional connectivity influence working memory and long-term memory performance. Journal of Neuroscience 30:14399–14410.
Bonnefond M, Jensen O (2012) Alpha Oscillations Serve to Protect Working Memory Maintenance against Anticipated Distracters. Current Biology 22:1969–1974.
Botvinick MM, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulate cortex: an update. Trends in Cognitive Sciences 8:539–546.
Boynton GM (2005) Attention and visual perception. Current Opinion in Neurobiology 15:465–469.
Brady TF, Alvarez GA (2011) Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. Psychological Science 22:384–392.
Brady TF, Konkle T, Alvarez GA (2011) A review of visual memory capacity: Beyond individual items and toward structured representations. Journal of Vision 11:4–4.
Brady TF, Tenenbaum JB (2013) A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. Psychological Review 120:85–109.
Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433–436.
Brandt ME, Ben H Jansen (1991) The Relationship Between Prestimulus Alpha Amplitude and Visual Evoked Potential Amplitude. International Journal of Neuroscience 61:261–268.
Brandt ME, Jansen BH, Carbonari JP (1991) Pre-stimulus spectral EEG patterns and the visual evoked response. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section 80:16–20.
Braver TS, Barch DM, Cohen JD (1999) Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. Biological Psychiatry 46:312–328.
Bressler SL, Tang W, Sylvester CM, Shulman GL, Corbetta M (2008) Top-Down Control of Human Visual Cortex by Frontal and Parietal Cortex in Anticipatory Visual Spatial Attention. Journal of Neuroscience 28:10056–10061.
Brignani D, Lepsien J, Rushworth MFS, Nobre AC (2009) The Timing of

Neural Activity during Shifts of Spatial Attention. Journal of Cognitive Neuroscience 21:2369–2383.
Brouwer GJ, Heeger DJ (2009) Decoding and Reconstructing Color from Responses in Human Visual Cortex. Journal of Neuroscience 29:13992–14003.
Brouwer GJ, Heeger DJ (2011) Cross-orientation suppression in human visual cortex. Journal of Neurophysiology 106(5):2108-2119.
Bundesen C, Habekost T, Kyllingsbæk S (2005) A neural theory of visual attention: bridging cognition and neurophysiology. Psychological Review 112:291–328.
Buonomano DV, Maass W (2009) State-dependent computations: spatiotemporal processing in cortical networks. Nat Rev Neurosci 10:113–125.
Burgess PW, Dumontheil I, Gilbert SJ (2007) The gateway hypothesis of rostral prefrontal cortex (area 10) function. Trends in Cognitive Sciences 11:290–298.
Busch NA, VanRullen R (2010) Spontaneous EEG oscillations reveal periodic sampling of visual attention. PNAS 107:16048–16053.
Buschman TJ, Miller EK (2007) Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. Science 315:1860–1862.
Buschman TJ, Siegel M, Roy JE, Miller EK (2011) Neural substrates of cognitive capacity limitations. PNAS 108:11252–11255.
Butters N, Pandya D (1969) Retention of Delayed-Alternation: Effect of Selective Lesions of Sulcus Principalis. Science 165:1271–1273.
Buzsáki G, Moser EI (2013) Memory, navigation and theta rhythm in the hippocampal-entorhinal system. Nature Neuroscience, 16(2):130–138.
Carandini M, Heeger DJ (2012) Normalization as a canonical neural computation. Nat Rev Neurosci 13:51–62.
Carrasco M, Eckstein M, Verghese P, Boynton G, Treue S (2009) Visual attention: Neurophysiology, psychophysics and cognitive neuroscience. Vision Research 49:1033–1036.
Chafee MV, Goldman-Rakic PS (2000) Inactivation of Parietal and Prefrontal Cortex Reveals Interdependence of Neural Activity During Memory-Guided Saccades. Journal of Neurophysiology 83:1550–1566.
Chakravarthi R, VanRullen R (2012) Conscious updating is a rhythmic process. PNAS 109:10599–10604.
Chawla D, Rees G, Friston KJ (1999) The physiological basis of attentional modulation in extrastriate visual areas. Nature Neuroscience 2:671–676.
Chelazzi L, Duncan J, Miller EK, Desimone R (1998) Responses of Neurons in Inferior Temporal Cortex During Memory-Guided Visual Search. Journal of Neurophysiology 80:2918–2940.
Chelazzi L, Libera Della C, Sani I, Santandrea E (2011) Neural basis of visual selective attention. WIREs Cogn Sci 2:392–407.
Chelazzi L, Miller EK, Duncan J, Desimone R (1993) A neural basis for visual search in inferior temporal cortex. Nature 363:345–347.

Chelazzi L, Miller EK, Duncan J, Desimone R (2001) Responses of neurons in macaque area V4 during memory-guided visual search. Cerebral Cortex 11:761–772.
Christophel TB, Hebart MN, Haynes J-D (2012) Decoding the contents of visual short-term memory from human visual and parietal cortex. Journal of Neuroscience 32:12983–12989.
Chun MM (2011) Visual working memory as visual attention sustained internally over time. Neuropsychologia 49:1407–1409.
Chun MM, Golomb JD, Turk-Browne NB (2011) A Taxonomy of External and Internal Attention. Annu Rev Psychol 62:73–101.
Chun MM, Johnson MK (2011) Memory: Enduring Traces of Perceptual and Reflective Attention. Neuron 72:520–535.
Churchland AK, Kiani R, Chaudhuri R, Wang X-J, Pouget A, Shadlen MN (2011) Variance as a Signature of Neural Computations during Decision Making. Neuron 69:818–831.
Churchland MM et al. (2010) Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nature 13:369–378.
Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. Nature Neuroscience 17:455–462.
Clapp WC, Rubens MT, Gazzaley A (2010) Mechanisms of Working Memory Disruption by External Interference. Cerebral Cortex 20:859–872.
Cohen JD, Braver TS, Brown JW (2002) Computational perspectives on dopamine function in prefrontal cortex. Current Opinion in Neurobiology 12:223–229.
Cohen MR, Maunsell JHR (2010) A Neuronal Population Measure of Attention Predicts Behavioral Performance on Individual Trials. Journal of Neuroscience 30:15241–15253.
Colclough, G. L., Brookes, M. J., Smith, S. M., & Woolrich, M. W. (2015). A symmetric multivariate leakage correction for MEG connectomes. NeuroImage 117:439-448.
Cole MW, Bassett DS, Power JD, Braver TS, Petersen SE (2014) Intrinsic and Task-Evoked Network Architectures of the Human Brain. Neuron 83:238–251.
Cole MW, Reynolds JR, Power JD, Repovs G, Anticevic A, Braver TS (2013) Multi-task connectivity reveals flexible hubs for adaptive task control. Nature Neuroscience 16:1348–1355.
Constantinidis C, Steinmetz MA (1996) Neuronal activity in posterior parietal area 7a during the delay periods of a spatial memory task. Journal of Neurophysiology 76(2):1352-1355.
Corbetta M, Miezin FM, Shulman GL, Petersen SE (1993) A PET study of visuospatial attention. J Neurosci 13:1202–1226.
Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3:215–229.
Cornelissen FW, Peters EM, Palmer J (2002) The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. Behavior Research Methods, Instruments, & Computers 34:613–617.
Courtney SM (1998) An Area Specialized for Spatial Working Memory in

Human Frontal Cortex. Science 279:1347–1351.
Cowan N (2000) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci 24:87–114–discussion114–85.
Cravo AM, Rohenkohl G, Wyart V, Nobre AC (2013) Temporal Expectation Enhances Contrast Sensitivity by Phase Entrainment of Low-Frequency Oscillations in Visual Cortex. Journal of Neuroscience 33:4002–4010.
Crittenden BM, Duncan J (2014) Task difficulty manipulation reveals multiple demand activity but no frontal lobe hierarchy. Cereb Cortex 24:532–540.
Crowe DA, Averbeck BB, Chafee MV (2010) Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex. Journal of Neuroscience 30:11640–11653.
Curtis CE, D'Esposito M (2003) Persistent activity in the prefrontal cortex during working memory. Trends in Cognitive Sciences 7:415–423.
D'Ardenne K, Eshel N, Luka J, Lenartowicz A, Nystrom LE, Cohen JD (2012) Role of prefrontal cortex and the midbrain dopamine system in working memory updating. PNAS 109(49):19900–19909.
David SV, Hayden BY, Mazer JA, Gallant JL (2008) Attention to Stimulus Features Shifts Spectral Tuning of V4 Neurons during Natural Vision. Neuron 59:509–521.
Dell'Acqua R, Sessa P, Toffanin P, Luria R, Jolicœur P (2010) Orienting attention to objects in visual short-term memory. Neuropsychologia 48:419–428.
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 134:9–21.
Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193–222.
D'Esposito M, Postle BR, Jonides J, Smith EE (1999) The Neural Substrate and Temporal Dynamics of Interference Effects in Working Memory as Revealed by Event-Related Functional MRI. PNAS 96:7514–7519.
D'Esposito M, Postle BR (2015) The cognitive neuroscience of working memory. Annu Rev Psychol 66:115–142.
Diedrichsen J, Ridgway GR, Friston KJ, Wiestler T (2011). Comparing the similarity and spatial structure of neural representations: a pattern-component model. NeuroImage 55(4):1665–1678.
Dienes Z (2011) Bayesian Versus Orthodox Statistics: Which Side Are You On? Perspectives on Psychological Science 6:274–290.
Donahue CH, Lee D (2015) Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. Nature Neuroscience 18:295–301.
Dosher BA, Lu Z-L (1999) Mechanisms of perceptual learning. Vision Research 39:3197–3221.
Duncan J (2013) The Structure of Cognition: Attentional Episodes in Mind

and Brain. Neuron 80:35–50.
Edin F, Klingberg T, Johansson P, McNab F, Tegner J, Compte A (2009) Mechanism for top-down control of working memory capacity. PNAS 106:6802–6807.
Eichenbaum H (2013) Memory on time. Trends in Cognitive Sciences, 17(2):81–88.
Eimer M (1996) The N2pc component as an indicator of attentional selectivity. Electroencephalography and Clinical Neurophysiology 99:225–234.
Eimer M, Velzen JV, Driver J (2002) Cross-Modal Interactions between Audition, Touch, and Vision in Endogenous Spatial Attention: ERP Evidence on Preparatory States and Sensory Modulations. Journal of Cognitive Neuroscience 14:254–271.
Eimer M, Forster B, van Velzen J (2003) Anterior and posterior attentional control systems use different spatial reference frames: ERP evidence from covert tactile-spatial orienting. Psychophysiol 40:924–933.
Eimer M, Kiss M (2010) An electrophysiological measure of access to representations in visual working memory. Psychophysiol 47:197–200.
Emrich SM, Riggall AC, LaRocque JJ, Postle BR (2013) Distributed Patterns of Activity in Sensory Cortex Reflect the Precision of Multiple Items Maintained in Visual Short-Term Memory. Journal of Neuroscience 33:6516–6523.
Engel TA, Chaisangmongkon W, Freedman DJ, Wang X-J (2015) Choice-correlated activity fluctuations underlie learning of neuronal category representation. Nature Communications 6:6454.
Engle RW, Tuholski SW, Laughlin JE, Conway ARA (1999) Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. Journal of Experimental Psychology: General 128:309–331.
Ester EF, Anderson DE, Serences JT, Awh E (2013) A Neural Measure of Precision in Visual Working Memory. Journal of Cognitive Neuroscience 25:754–761.
Ester EF, Serences JT, Awh E (2009) Spatially Global Representations in Human Primary Visual Cortex during Working Memory Maintenance. Journal of Neuroscience 29:15258–15265.
Evans AC, Collins DL, and Milner B (1992). "An MRI-based stereotactic atlas from 250 young normal subjects", Journal Soc. Neurosci. Abstr. 18: 408, 1992
Feldman H, Friston KJ (2010) Attention, Uncertainty, and Free-Energy. Front Hum Neurosci 4.
Feredoes E, Heinen K, Weiskopf N, Ruff C, Driver J (2011) Causal evidence for frontal involvement in memory target maintenance by posterior brain areas during distracter interference of visual working memory. PNAS 108:17510–17515.
Fischer J, Whitney D (2014) Serial dependence in visual perception. Nat Neuro 17:738–743.
Fontolan L, Morillon B, Liegeois-Chauvel C, Giraud A-L (2014) The

contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nature Communications 5:4694.
Fougnie D, Suchow JW, Alvarez GA (2012) Variability in the quality of visual working memory. Nature Communications 3:1229.
Foxe JJ, Simpson GV, Ahlfors SP (1998) Parieto-occipital ~10 Hz activity reflects anticipatory state of visual attention mechanisms. Neuroreport 9:3929–3933.
Frank MJ, Loughry B, O'Reilly RC (2001) Interactions between frontal cortex and basal ganglia in working memory: a computational model. Cognitive, Affective, & Behavioral Neuroscience 1:137–160.
Freedman DJ, Assad JA (2006) Experience-dependent representation of visual categories in parietal cortex. Nature 443:85–88.
Friedman HR, Goldman-Rakic PS (1994) Coactivation of prefrontal cortex and inferior parietal cortex in working memory tasks revealed by 2DG functional mapping in the rhesus monkey. J Neurosci 14:2775–2788.
Fries P, Reynolds JH, Rorie AE, Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. Science 291:1560–1563.
Friston K (2009) The free-energy principle: a rough guide to the brain? Trends in Cognitive Sciences 13:293–301.
Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD, and Frackowiak RSJ (1995). Statistical parametric maps in functional imaging: A general linear approach. Human Brain Mapping, 2:189-210.
Fujisawa S, Amarasingham A, Harrison MT, Buzsáki G (2008) Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nature Neuroscience 11:823–833.
Fukuda K, Vogel E, Mayr U, Awh E (2010) Quantity, not quality: the relationship between fluid intelligence and working memory capacity. Psychonomic Bulletin & Review 17:673–679.
Funahashi S, Bruce CJ, Goldman-Rakic PS (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. Journal of Neurophysiology 61:331–349.
Funahashi S, Bruce CJ, Goldman-Rakic PS (1990) Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. Journal of Neurophysiology 63:814–831.
Funahashi S, Bruce CJ, Goldman-Rakic PS (1993) Dorsolateral prefrontal lesions and oculomotor delayed-response performance: evidence for mnemonic "scotomas". J Neurosci 13:1479–1497.
Fuster JM (1993) Frontal lobes. Current Opinion in Neurobiology 3:160–165.
Fuster JM, Alexander GE (1971) Neuron activity related to short-term memory. Science 173(3997):652-654.
Fuster JM, Jervey JP (1981) Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science 212:952–955.
Garcia JO, Srinivasan R, Serences JT (2013) Near-Real-Time Feature-Selective Modulations in Human Cortex. Current Biology 23:515–522.
Gathercole SE, Pickering SJ, Knight C, Stegmann Z (2003) Working

memory skills and educational attainment: evidence from national curriculum assessments at 7 and 14 years of age. Appl Cognit Psychol 18:1–16.
Gazzaley A (2011) Influence of early attentional modulation on working memory. Neuropsychologia 49:1410–1424.
Gazzaley A, Nobre AC (2012) Top-down modulation: bridging selective attention and working memory. Trends in Cognitive Sciences 16:129–135.
Gazzaley A, Rissman J, Cooney J, Rutman A, Seibert T, Clapp W, D'Esposito M (2007) Functional interactions between prefrontal and visual association cortex contribute to top-down modulation of visual processing. Cerebral Cortex 17 Suppl 1:i125–i135.
Gazzaley A, Rissman J, D'Esposito M (2004) Functional connectivity during working memory maintenance. Cognitive, Affective, & Behavioral Neuroscience 4:580–599.
Ghose GM, Yang T, Maunsell JHR (2002) Physiological correlates of perceptual learning in monkey V1 and V2. Journal of Neurophysiology 87(4):1867–1888.
Gibson JJ (1937) Adaptation, after-effect, and contrast in the perception of tilted lines. II. Simultaneous contrast and the areal restriction of the after-effect. Journal of Experimental Psychology 20:553–569.
Gitelman DR, Nobre AC, Parrish TB, LaBar KS, Kim Y-H, Meyer JR, Mesulam MM (1999) A large-scale distributed network for covert spatial attention: Further anatomical delineation based on stringent behavioural and cognitive controls. Brain 122:1093–1106.
Gold JI, Shadlen MN (2007) The Neural Basis of Decision Making. Annual Review of Neuroscience, 30(1):535–574.
Goldman MS (2009) Memory without Feedback in a Neural Network. Neuron 61:621–634.
Goldman PS, Rosvold HE, Vest B, Galkin TW (1971) Analysis of the delayed-alternation deficit produced by dorsolateral prefrontal lesions in the rhesus monkey. J Comp Physiol Psychol 77:212–220.
Goldman-Rakic PS (1987) Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory. Hoboken, NJ, USA: John Wiley & Sons.
Goldman-Rakic PS, Bates JF, Chafee MV (1992) The prefrontal cortex and internally generated motor acts. Current Opinion in Neurobiology 2:830–835.
Gorgoraptis N, Catalao RFG, Bays PM, Husain M (2011) Dynamic Updating of Working Memory Resources for Visual Objects. Journal of Neuroscience 31:8502–8511.
Gould IC, Rushworth MF, Nobre AC (2011) Indexing the graded allocation of visuospatial attention using anticipatory alpha oscillations. Journal of Neurophysiology 105:1318–1326.
Gözenman F, Tanoue RT, Metoyer T, Berryhill ME (2014) Invalid retro-cues can eliminate the retro-cue benefit: Evidence for a hybridized account. Journal of Experimental Psychology: Human Perception and Performance 40:1748–1754.

Gray CM, König P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338(6213):334-337.
Green JJ, McDonald JJ (2006) An event-related potential study of supramodal attentional control and crossmodal attention effects. Psychophysiol 43:161–171.
Gregoriou GG, Gotts SJ, Zhou H, Desimone R (2009) High-Frequency, Long-Range Coupling Between Prefrontal and Visual Cortex During Attention. Science 324:1207–1210.
Griffin IC, Nobre AC (2003) Orienting attention to locations in internal representations. Journal of Cognitive Neuroscience 15:1176–1194.
Haegens S, Luther L, Jensen O (2012) Somatosensory Anticipatory Alpha Activity Increases to Suppress Distracting Input. Journal of Cognitive Neuroscience 24:677–685.
Haegens S, Nacher V, Luna R, Romo R, Jensen O (2011) α-Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. PNAS 108:19377–19382.
Haegens S, Osipova D, Oostenveld R, Jensen O (2010) Somatosensory working memory performance in humans depends on both engagement and disengagement of regions in a distributed network. Human Brain Mapping 31(1):26-35.
Hanslmayr S, Aslan A, Staudigl T, Klimesch W, Herrmann CS, Bäuml K-H (2007) Prestimulus oscillations predict visual perception performance between and within subjects. NeuroImage 37:1465–1473.
Hanslmayr S, Volberg G, Wimber M, Dalal SS, Greenlee MW (2013) Prestimulus Oscillatory Phase at 7 Hz Gates Cortical Information Flow and Visual Perception. Current Biology 23:2273–2278.
Harris JA, Miniussi C, Harris IM, Diamond ME (2002) Transient storage of a tactile memory trace in primary somatosensory cortex. Journal of Neuroscience 22:8720–8725.
Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual areas. Nature 458:632–635.
Harter MR, Miller SL, Price NJ, LaLonde ME, Keyes AL (1989) Neural Processes Involved in Directing Attention. Journal of Cognitive Neuroscience 1:223–237.
Harvey CD, Coen P, Tank DW (2012) Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature 484:62–68.
Hasson U, Chen J, Honey CJ (2015) Hierarchical process memory: memory as an integral component of information processing. Trends in Cognitive Sciences 19:304–313.
Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding neural representational spaces using multivariate pattern analysis. Annu Rev Neurosci 37:435–456.
Hayden BY, Gallant JL (2013) Working memory and decision processes in visual area V4. Front Neurosci 7:1–10.
Heath RA (1982) A model for signal detection based on an adaptive filter.

Biol Cybern 45:95–100.
Hebb DO (1949) The organization of behavior: A neuropsychological approach. New York: Wiley & Sons.
Henseler I, Krüger S, Dechent P, Gruber O (2011) A gateway system in rostral PFC? Evidence from biasing attention to perceptual information and internal representations. NeuroImage 56:1666–1676.
Higo T, Mars RB, Boorman ED, Buch ER, Rushworth MFS (2011) Distributed and causal influence of frontal operculum in task control. PNAS 108:4230–4235.
Hillyard SA, Vogel EK, Luck SJ (1998) Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philosophical Transactions of the Royal Society B: Biological Sciences 353:1257–1270.
Hopf JM, Mangun GR (2000) Shifting visual attention in space: an electrophysiological analysis using high spatial resolution mapping. Clinical Neurophysiology 111(7):1241-1257.
Hopf JM, Boelmans K, Schoenfeld AM, Heinze HJ, Luck SJ (2002) How does attention attenuate target–distractor interference in vision? Cognitive Brain Research 15:17–29.
Hopfield JJ (1980) The energy relay: a proofreading scheme based on dynamic cooperativity and lacking all characteristic symptoms of kinetic proofreading in DNA replication and protein synthesis. PNAS 77:5248–5252.
Huang J (2010) Distortions in recall from visual memory: Two classes of attractors at work. Journal of Vision 10:1–27.
Huang J, Sekuler R (2010) Distortions in recall from visual memory: two classes of attractors at work. Journal of Vision 10(24):1–27.
Hussar CR, Pasternak T (2012) Memory-guided sensory comparisons in the prefrontal cortex: contribution of putative pyramidal cells and interneurons. Journal of Neuroscience 32:2747–2761.
Hussar CR, Pasternak T (2013) Common rules guide comparisons of speed and direction of motion in the dorsolateral prefrontal cortex. Journal of Neuroscience 33:972–986.
Huttenlocher J, Hedges LV, Vevea JL (2000) Why do categories affect stimulus judgment? Journal of Experimental Psychology: General 129:220–241.
Jacob SN, Nieder A (2014) Complementary Roles for Primate Frontal and Parietal Cortex in Guarding Working Memory from Distractor Stimuli. Neuron 83:226–237.
Jacobsen CF (1935) Functions of Frontal Association Area in Primates. Arch Neur Psych 33:558–569.
Janczyk M, Berryhill ME (2014) Orienting attention in visual working memory requires central capacity: Decreased retro-cue effects under dual-task conditions. Atten Percept Psychophys 76:715–724.
Jensen O, Bonnefond M, VanRullen R (2012) An oscillatory mechanism for prioritizing salient unattended stimuli. Trends in Cognitive Sciences 16:200–206.
Jensen O, Gelfand J, Kounios J, Lisman JE (2002) Oscillations in the

Alpha Band (9-12 Hz) Increase with Memory Load during Retention in a Short-term Memory Task. Cerebral Cortex 12:877–882.
Jensen O, Mazaheri A (2010) Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition. Front Hum Neurosci 4:1–8.
Jensen O, Tesche CD (2002) Frontal theta activity in humans increases with memory load in a working memory task. Eur J Neurosci 15:1395–1399.
Johnson A, Redish AD (2007) Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. Journal of Neuroscience, 27(45):12176–12189.
Johnson JS, Spencer JP, Luck SJ, Schöner G (2009) A dynamic neural field model of visual working memory and change detection. Psychological Science 20(5):568-577.
Johnson MK, Raye CL, Mitchell KJ, Greene EJ, Cunningham WA, Sanislow CA (2005) Using fMRI to investigate a component process of reflection: Prefrontal correlates of refreshing a just-activated representation. Cognitive, Affective, & Behavioral Neuroscience 5:339–361.
Johnson MK, Reeder JA, Raye CL, Mitchell KJ (2002) Second Thoughts versus Second Looks: An Age-Related Deficit in Reflectively Refreshing Just-Activated Information. Psychological Science 13:64–67.
Johnson MR, Higgins JA, Norman KA, Sederberg PB, Smith TA, Johnson MK (2013) Foraging for Thought: An Inhibition-of-Return-Like Effect Resulting From Directing Attention Within Working Memory. Psychological Science 24:1104–1112.
Johnson MR, Mitchell KJ, Raye CL, D'Esposito M, Johnson MK (2007) A brief thought can modulate activity in extrastriate visual areas: Top-down effects of refreshing just-seen visual stimuli. Neuroimage 37:290–299.
Jokisch D, Jensen O (2007) Modulation of Gamma and Alpha Activity during a Working Memory Task Engaging the Dorsal or Ventral Stream. Journal of Neuroscience 27:3244–3251.
Jonides J, Nee DE (2006) Brain mechanisms of proactive interference in working memory. Neuroscience 139:181–193.
Kang MS, Choi J (2015) Retrieval-Induced Inhibition in Short-Term Memory. Psychological Science 26:1014–1025.
Kass RE, Raftery AE (2012) Bayes Factors. Journal of the American Statistical Association 90:773–795.
Kastner S, McMains SA, Beck D (2009) Mechanisms of selective attention in the human visual system: Evidence from neuroimaging. The cognitive neurosciences. pp. 205-217.
Kastner S, Pinsk MA, De Weerd P, Desimone R, Ungerleider LG (1999) Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron 22:751–761.
Kelly SP, Gomez-Ramirez M, Foxe JJ (2009) The strength of anticipatory spatial biasing predicts target discrimination at attended locations: a

high-density EEG study. Eur J Neurosci 30:2224–2234.
Keppel G, Underwood BJ (1962) Proactive inhibition in short-term retention of single items. Journal of Verbal Learning and Verbal Behavior 1:153–161.
Kastner S, Ungerleider LG (2000) Mechanisms of visual attention in the human cortex. Annu Rev Neurosci 23:315–341.
Keshvari S, van den Berg R, Ma WJ (2013) No Evidence for an Item Limit in Change Detection. PLoS Comput Biol 9:e1002927.
Kinchla RA, Chen Z, Evert D (1995) Precue effects in visual search: data or resource limited? Perception & Psychophysics 57(4):441–450.
King J-R, Dehaene S (2014) Characterizing the dynamics of mental representations: the temporal generalization method. Trends in Cognitive Sciences 18:203–210.
King J-R, Faugeras F, Gramfort A, Schurger A, Karoui El I, Sitt JD, Rohaut B, Wacongne C, Labyt E, Bekinschtein T, Cohen L, Naccache L, Dehaene S (2013) Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. NeuroImage 83:726–738.
Kiss M, Jolicœur P, Dell'Acqua R, Eimer M (2008) Attentional capture by visual singletons is mediated by top-down task set: New evidence from the N2pc component. Psychophysiol 45:1013–1024.
Kiyonaga A, Egner T (2013) Working memory as internal attention: toward an integrative account of internal and external selection processes. Psychonomic Bulletin & Review 20:228–242.
Klimesch W, Doppelmayr M, Schwaiger J, Auinger P, Winkler T (1999) “Paradoxical” alpha synchronization in a memory task. Cognitive Brain Research 7:493–501.
Klimesch W, Sauseng P, Hanslmayr S (2007) EEG alpha oscillations: The inhibition–timing hypothesis. Brain Research Reviews 53:63–88.
Knapen T, Rolfs M, Wexler M, Cavanagh P (2010) The reference frame of the tilt aftereffect. Journal of Vision 10(8):1–13.
Koechlin E, Ody C, Kouneiher F (2003) The architecture of cognitive control in the human prefrontal cortex. Science 302:1181–1185.
Kohonen T, Oja E (1976) Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. Biol Cybern 21(2):85-95.
Kourtzi Z, Kanwisher N (2000) Cortical regions involved in perceiving object shape. Journal of Neuroscience 20:3310–3318.
Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. Trends in Cognitive Sciences 17:401–412.
Kubota K, Niki H (1971) Prefrontal cortical unit activity and delayed alternation performance in monkeys. Journal of Neurophysiology 34:337–347.
Kuo B-C, Stokes MG, Murray AM, Nobre AC (2014) Attention Biases Visual Activity in Visual STM. Journal of Cognitive Neuroscience 26(7):1377-1389.
Kuo B-C, Stokes MG, Nobre AC (2012) Attention Modulates Maintenance

of Representations in Visual Short-term Memory. Journal of Cognitive Neuroscience 24:51–60.
Kuo BC, Rao A, Lepsien J, Nobre AC (2009) Searching for Targets within the Spatial Layout of Visual Short-Term Memory. Journal of Neuroscience 29:8032–8038.
Kuo B-C, Yeh Y-Y, Chen AJ-W, D'Esposito M (2011) Functional connectivity during top-down modulation of visual short-term memory representations. Neuropsychologia 49:1589–1596.
Landman R, Spekreijse H, Lamme VAF (2003) Large capacity storage of integrated objects before change blindness. Vision Research 43:149–164.
Lange J, Oostenveld R, Fries P (2013) Reduced Occipital Alpha Power Indexes Enhanced Excitability Rather than Improved Visual Perception. Journal of Neuroscience 33:3212–3220.
Lara AH, Wallis JD (2012) Capacity and precision in an animal model of visual short-term memory. Journal of Vision 12:13–13.
Lara AH, Wallis JD (2014) Executive control processes underlying multi-item working memory. Nat Neuro 17:876–883.
LaRocque JJ, Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR (2013) Decoding attended information in short-term memory: an EEG study. Journal of Cognitive Neuroscience 25:127–142.
LaRocque JJ, Lewis-Peacock JA, Postle BR (2014) Multiple neural states of representation in short-term memory? It's a matter of attention. Front Hum Neurosci 8:5.
Laurent G (2002) Olfactory network dynamics and the coding of multidimensional signals. Nat Rev Neurosci 3:884–895.
Lee S-H, Kravitz DJ, Baker CI (2013) Goal-dependent dissociation of visual and prefrontal cortices during working memory. Nat Neuro 16:997–999.
Lepsien J, Griffin IC, Devlin JT, Nobre AC (2005) Directing spatial attention in mental representations: Interactions between attentional orienting and working-memory load. NeuroImage 26:733–743.
Lepsien J, Nobre AC (2006) Cognitive control of attention in the human brain: Insights from orienting attention to mental representations. Brain Research 1105:20–31.
Lepsien J, Nobre AC (2007) Attentional Modulation of Object Representations in Working Memory. Cerebral Cortex 17:2072–2083.
Lepsien J, Thornton I, Nobre AC (2011) Modulation of working-memory maintenance by directed attention. Neuropsychologia 49:1569–1577.
Leszczyński M, Myers NE, Akyürek EG, Schubö A (2012) Recoding between two types of stm representation revealed by the dynamics of memory search. Journal of Cognitive Neuroscience 24:653–663.
Leutgeb S, Leutgeb JK, Treves A, Moser M-B, Moser EI (2004) Distinct ensemble codes in hippocampal areas CA3 and CA1. Science 305:1295–1298.
Lewis-Peacock JA, Drysdale AT, Oberauer K, Postle BR (2012) Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. Journal of Cognitive Neuroscience 24:61–79.

Lewis-Peacock JA, Drysdale AT, Postle BR (2014) Neural Evidence for the Flexible Control of Mental Representations. Cerebral Cortex:1–11.
Lewis-Peacock JA, Postle BR (2012) Decoding the internal focus of attention. Neuropsychologia 50:470–478.
Liberman A, Fischer J, Whitney D (2014) Serial Dependence in the Perception of Faces. Current Biology 24(21):2569-2574.
Liebe S, Hoerzer GM, Logothetis NK, Rainer G (2012) Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance. Nature Neuroscience 15:456–462.
Lin P-H, Luck SJ (2009) The Influence of Similarity on Visual Working Memory Representations. Vis Cogn 17:356–372.
Lin P-H, Luck SJ (2012) Proactive Interference Does Not Meaningfully Distort Visual Working Memory Capacity Estimates in the Canonical Change Detection Task. Front Psychology 3:42.
Lisman JE, Jensen O (2013) The Theta-Gamma Neural Code. Neuron, 77(6):1002–1016.
Liu T, Abrams J, Carrasco M (2009) Voluntary Attention Enhances Contrast Appearance. Psychological Science 20:354–362.
Luck SJ, Chelazzi L, Hillyard SA, Desimone R (1997) Neural Mechanisms of Spatial Selective Attention in Areas V1, V2, and V4 of Macaque Visual Cortex. Journal of Neurophysiology 77:24–42.
Luck SJ, Hillyard SA (1994) Electrophysiological correlates of feature analysis during visual search. Psychophysiol 31:291–308.
Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390:279–281.
Luck SJ, Vogel EK (2013) Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends in Cognitive Sciences 17:391–400.
Luria AR (1966) Higher Cortical Functions in Man. New York: Basic.
Luria AR, Homskaya ED (1966) The Frontal Lobes and Regulation of Psychological Processes. Moscow: MGU.
Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nature Neuroscience 9:1432–1438.
Ma WJ, Husain M, Bays PM (2014) Changing concepts of working memory. Nature Neuroscience 17:347–356.
Machizawa MG, Goh CCW, Driver J (2012) Human Visual Short-Term Memory Precision Can Be Varied at Will When the Number of Retained Items Is Low. Psychological Science 23:554–559.
Magnussen S (2000) Low-level memory processes in vision. Trends in Neurosciences 23:247–251.
Magnussen S, Greenlee MW (1992) Retention and disruption of motion information in visual short-term memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 18:151–156.
Magnussen S, Greenlee MW, Asplund R, Dyrnes S (1991) Stimulus-specific mechanisms of visual short-term memory. Vision Research 31:1213–1219.
Makovski T (2012) Are multiple visual short-term memory storages necessary to explain the retro-cue effect? Psychonomic Bulletin &

Review 19:470–476.
Makovski T, Sussman R, Jiang YV (2008) Orienting attention in visual working memory reduces interference from memory probes. Journal of Experimental Psychology: Learning, Memory, and Cognition 34:369–380.
Mangun GR (1994) Orienting Attention in the Visual Fields: An Electrophysiological Analysis. In: Cognitive electrophysiology, pp 81–101. Boston, MA: Birkhäuser Boston.
Mante V, Sussillo D, Shenoy KV, Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503:78–84.
Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods 164:177–190.
Marr D, Willshaw D, McNaughton B (1991) Simple memory: a theory for archicortex. Boston: Birkhäuser.
Martínez-García M, Rolls ET, Deco G, Romo R (2011) Neural and computational mechanisms of postponed decisions. PNAS 108(28):11626–11631.
Mathewson KE, Lleras A, Beck DM, Fabiani M, Ro T, Gratton G (2011) Pulsed out of awareness: EEG alpha oscillations represent a pulsed-inhibition of ongoing cortical processing. Front Psychology 2.
Mate J, Baqués J (2009) Visual similarity at encoding and retrieval in an item recognition task. The Quarterly Journal of Experimental Psychology 62(7):1277-1284.
Matthey L, Bays PM, Dayan P (2015) A probabilistic palimpsest model of visual short-term memory. PLoS Comput Biol 11:e1004003.
Maunsell JHR, Treue S (2006) Feature-based attention in visual cortex. Trends in Neurosciences 29:317–322.
Mazaheri A, Jensen O (2008) Asymmetric Amplitude Modulations of Brain Oscillations Generate Slow Evoked Responses. J Neurosci 28(31):7781-7787.
McAdams CJ, Maunsell JH (1999) Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. The Journal of Neuroscience 19(1):431–441.
McCollough AW, Machizawa MG, Vogel EK (2007) Electrophysiological measures of maintaining representations in visual working memory. Cortex 43:77–94.
McNaughton BL, Morris RGM (1987) Hippocampal synaptic enhancement and information storage within a distributed memory system. Trends in Neurosciences 10:408–415.
Medendorp WP, Kramer GFI, Jensen O, Oostenveld R, Schoffelen JM, Fries P (2007) Oscillatory Activity in Human Parietal and Occipital Cortex Shows Hemispheric Lateralization and Memory Effects in a Delayed Double-Step Saccade Task. Cerebral Cortex 17:2364–2374.
Mendoza-Halliday D, Torres S, Martinez-Trujillo JC (2014) Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. Nature Neuroscience 17:1255–1262.
Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T (2008)

Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. Journal of Neurophysiology 100:1407–1419.
Miller BT, Vytlacil J, Fegen D, Pradhan S, D'Esposito M (2010) The Prefrontal Cortex Modulates Category Selectivity in Human Extrastriate Cortex. Journal of Cognitive Neuroscience 23(1):1-10.
Miller EK, Buschman TJ (2007) Response to Comment on “Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices.” Science 318:44c–44c.
Miller EK, Desimone R (1994) Parallel neuronal mechanisms for short-term memory. Science 263:520–522.
Miller EK, Erickson CA, Desimone R (1996) Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J Neurosci 16:5154–5167.
Miller EK, Li L, Desimone R (1993) Activity of neurons in anterior inferior temporal cortex during a short-term memory task. J Neurosci 13:1460–1478.
Miller, J., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. Psychophysiology, 35(1), 99–115.
Milner B (1964) Some effects of frontal lobectomy in man. Warren and Akert (eds.). The Frontal Granular Cortex and Behavior. New York: McGraw-Hill.
Mongillo G, Barak O, Tsodyks M (2008) Synaptic theory of working memory. Science 319:1543–1546.
Monsell S (1978) Recency, immediate recognition memory, and reaction time. Cognitive Psychology 10:465–501.
Munneke J, Belopolsky AV, Theeuwes J (2012) Shifting Attention within Memory Representations Involves Early Visual Areas. PLoS ONE 7:e35528.
Murray AM, Nobre AC, Clark IA, Cravo AM, Stokes MG (2013) Attention restores discrete items to visual short-term memory. Psychological Science 24:550–556.
Murray AM, Nobre AC, Stokes MG (2011) Markers of preparatory attention predict visual short-term memory performance. Neuropsychologia 49:1458–1465.
Murray JD, Bernacchia A, Freedman DJ, Romo R, Wallis JD, Cai X, Padoa-Schioppa C, Pasternak T, Seo H, Lee D, Wang X-J (2014) A hierarchy of intrinsic timescales across primate cortex. Nature Neuroscience 17:1661–1663.
Murray MM, Brunet D, Michel CM (2008) Topographic ERP analyses: a step-by-step tutorial review. Brain Topogr 20:249–264.
Murty VP, Sambataro F, Radulescu E, Altamura M, Iudicello J, Zoltick B, Weinberger DR, Goldberg TE, Mattay VS (2011) Selective updating of working memory content modulates meso-cortico-striatal activity. NeuroImage 57:1264–1272.
Myers N, Wallis G (2013) Constraining Theories of Working Memory with Biophysical Modeling. J Neurosci 33:385–386.

Myung IJ (2003) Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology 47:90–100.
Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. NeuroImage 56:400–410.
Navalpakkam V, Itti L (2007) Search goal tunes visual features optimally. Neuron 53:605–617.
Nee DE, Brown JW (2012a) Dissociable Frontal-Striatal and Frontal-Parietal Networks Involved in Updating Hierarchical Contexts in Working Memory. Cerebral Cortex.
Nee DE, Brown JW (2012b) Rostral–caudal gradients of abstraction revealed by multi-variate pattern analysis of working memory. NeuroImage 63:1285–1294.
Nee DE, Jonides J (2009) Common and distinct neural correlates of perceptual and memorial selection. NeuroImage 45:963–975.
Nee DE, Jonides J, Berman MG (2007) Neural mechanisms of proactive interference-resolution. NeuroImage 38:740–751.
Neisser U (1967) Cognitive Psychology. New York: Appleton.
Nelissen N, Stokes M, Nobre AC, Rushworth MFS (2013) Frontal and parietal cortical interactions with distributed visual representations during selective attention and action selection. Journal of Neuroscience 33:16443–16458.
Nikolić D, Häusler S, Singer W, Maass W (2009) Distributed fading memory for stimulus properties in the primary visual cortex. Plos Biol 7:e1000260.
Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A Toolbox for Representational Similarity Analysis. PLoS Comput Biol 10:e1003553.
Nobre AC (2004) Probing the flexibility of attentional orienting in the human brain. In M Posner (Ed.), Cognitive Neuroscience of Attention. New York: Guildford Press.
Nobre AC, Coull JT, Maquet P, Frith CD, Vandenberghe R, Mesulam MM (2004) Orienting Attention to Locations in Perceptual Versus Mental Representations. Journal of Cognitive Neuroscience 16:363–373.
Nobre AC, Sebestyen GN, Gitelman DR, Mesulam MM, Frackowiak RS, Frith CD (1997) Functional localization of the system for visuospatial attention using positron emission tomography. Brain 120:515–533.
Nobre AC, Sebestyen GN, Miniussi C (2000) The dynamics of shifting visuospatial attention revealed by event-related potentials. Neuropsychologia 38:964–974.
Nobre AC, Stokes MG (2011) Attention and short-term memory: crossroads. Neuropsychologia 49:1391–1392.
Nolte G (2003) The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. Phys Med Biol 48:3637–3652.
Norman DA, Shallice T (1986) Attention to Action. In: Consciousness and Self-Regulation, pp 1–18. Boston, MA: Springer US.
Oberauer K (2002) Access to information in working memory: Exploring

the focus of attention. Journal of Experimental Psychology: Learning, Memory, and Cognition 28:411–421.
Oberauer K (2005) Binding and inhibition in working memory individual and age differences in short-term recognition. Journal of Experimental Psychology: General 134:368–387.
Oberauer K (2013) The focus of attention in working memory—from metaphors to mechanisms. Front Hum Neurosci 7.
Oberauer K, Hein L (2012) Attention to Information in Working Memory. Current Directions in Psychological Science 21:164–169.
Oberauer K, Kliegl R (2006) A formal model of capacity limits in working memory. Journal of Memory and Language 55:601–626.
Olivers CNL, Peters J, Houtkamp R, Roelfsema PR (2011) Different states in visual working memory: when it guides attention and when it does not. Trends in Cognitive Sciences 15:327–334.
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.
Orhan AE, Jacobs RA (2013) A probabilistic clustering theory of the organization of visual short-term memory. Psychological Review 120:297–328.
Orhan AE, Ma WJ (2015) Neural population coding of multiple stimuli. Journal of Neuroscience 35:3825–3841.
Papadimitriou C, Ferdoash A, Snyder LH (2015) Ghosts in the machine: memory interference from the previous trial. Journal of Neurophysiology 113:567–577.
Park S, Chun MM, Johnson MK (2010) Refreshing and integrating visual scenes in scene-selective cortex. Journal of Cognitive Neuroscience 22(12):2813-2822.
Passingham RE (1985) Memory of monkeys (Macaca mulatta) with lesions in prefrontal cortex. Behavioral Neuroscience 99:3–21.
Pasternak T, Greenlee MW (2005) Working memory in primate sensory systems. Nat Rev Neurosci 6:97–107.
Pasternak T, Lui LL, Spinelli PM (2015) Unilateral prefrontal lesions impair memory-guided comparisons of contralateral visual motion. Journal of Neuroscience 35:7095–7105.
Pearson B, Raškevičius J, Bays PM, Pertzov Y, Husain M (2014) Working memory retrieval as a decision process. Journal of Vision 14:2–2.
Pertzov Y, Bays PM, Joseph S, Husain M (2013) Rapid forgetting prevented by retrospective attention cues. Journal of Experimental Psychology: Human Perception and Performance 39:1224.
Pesaran B, Pezaris JS, Sahani M, Mitra PP, Andersen RA (2002) Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nature Neuroscience 5:805–811.
Petrides M (2002) The Mid-ventrolateral Prefrontal Cortex and Active Mnemonic Retrieval. Neurobiology of Learning and Memory 78:528–538.
Petrides M, Baddeley A (1996) Specialized Systems for the Processing of

Mnemonic Information within the Primate Frontal Cortex [and Discussion]. Philosophical Transactions of the Royal Society B: Biological Sciences 351(1346):1455-1462.
Petrides M, Milner B (1982) Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. Neuropsychologia 20:249–262.
Pfurtscheller G, Stancák A Jr., Neuper C (1996) Event-related synchronization (ERS) in the alpha band — an electrophysiological correlate of cortical idling: A review. International Journal of Psychophysiology 24:39–46.
Phillips WA (1974) On the distinction between sensory storage and short-term visual memory. Perception & Psychophysics 16:283–290.
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.
Postle BR, Berger JS, D'Esposito M (1999) Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. PNAS 96(22):12959-12964.
Postle BR (2006) Working memory as an emergent property of the mind and brain. Neuroscience 139:23–38.
Postle BR (2015a) Neural Bases of the Short-term Retention of Visual Information. Mechanisms of Sensory Working Memory: Attention and Performance XXV. 43-68.
Postle BR (2015b) The cognitive neuroscience of visual short-term memory. Current Opinion in Behavioral Sciences 1:40–46.
Praamstra P, Boutsen L, Humphreys GW (2005) Frontoparietal control of spatial attention and motor intention in human EEG. Journal of Neurophysiology 94:764–774.
Purushothaman, G., & Bradley, D. C. (2005). Neural population code for fine perceptual decisions in area MT. Nature Neuroscience, 8(1), 99–106.
Rajagovindan R, Ding M (2011) From prestimulus alpha oscillation to visual-evoked response: an inverted-U function and its attentional modulation. Journal of Cognitive Neuroscience 23:1379–1394.
Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2:79–87.
Raye CL, Johnson MK, Mitchell KJ, Reeder JA, Greene EJ (2002) Neuroimaging a Single Thought: Dorsolateral PFC Activity Associated with Refreshing Just-Activated Information. NeuroImage 15:447–453.
Reddy L, Kanwisher NG, VanRullen R (2009) Attention and biased competition in multi-voxel object representations. PNAS 106:21447–21452.
Rerko L, Oberauer K (2013) Focused, unfocused, and defocused information in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 39:1075–1096.
Rerko L, Souza AS, Oberauer K (2014) Retro-cue benefits in working memory without sustained focal attention. Memory &

Cognition42(5):712–728.
Reynolds JH, Heeger DJ (2009) The Normalization Model of Attention. Neuron 61:168–185.
Riggall AC, Postle BR (2012) The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. Journal of Neuroscience 32:12990–12998.
Rigotti M, Barak O, Warden MR, Wang X-J, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497:585–590.
Roe AW, Chelazzi L, Connor CE, Conway BR, Fujita I, Gallant JL, Lu H, Vanduffel W (2012) Toward a Unified Theory of Visual Area V4. Neuron 74:12–29.
Rohenkohl G, Nobre AC (2011) Alpha Oscillations Related to Anticipatory Attention Follow Temporal Expectations. Journal of Neuroscience 31:14076–14084.
Romei V, Brodbeck V, Michel C, Amedi A, Pascual-Leone A, Thut G (2008a) Spontaneous Fluctuations in Posterior α-Band EEG Activity Reflect Variability in Excitability of Human Visual Areas. Cerebral Cortex 18(9):2010-2018.
Romei V, Rihs T, Brodbeck V, Thut G (2008b) Resting electroencephalogram alpha-power over posterior sites indexes baseline visual cortex excitability. Neuroreport 19:203–208.
Roth JK, Johnson MK, Raye CL, Constable RT (2009) Similar and dissociable mechanisms for attention to internal versus external information. NeuroImage 48:601–608.
Rushworth MFS, Hadland KA, Paus T, Sipila PK (2002a) Role of the Human Medial Frontal Cortex in Task Switching: A Combined fMRI and TMS Study. Journal of Neurophysiology 87(5):2577-2592.
Rushworth MFS, Passingham RE, Nobre AC (2002b) Components of switching intentional set. Journal of Cognitive Neuroscience 14:1139–1150.
Rushworth MFS, Passingham RE, Nobre AC (2005) Components of attentional set-switching. Experimental Psychology (formerly "Zeitschrift für Experimentelle Psychologie") 52:83–98.
Rutman AM, Clapp WC, Chadick JZ, Gazzaley A (2010) Early top–down control of visual processing predicts working memory performance. Journal of Cognitive Neuroscience 22:1224–1234.
Saalmann YB, Pinsk MA, Wang L, Li X, Kastner S (2012) The Pulvinar Regulates Information Transmission Between Cortical Areas Based on Attention Demands. Science 337:753–756.
Sadaghiani S, Hesselmann G, Kleinschmidt A (2009) Distributed and antagonistic contributions of ongoing activity fluctuations to auditory stimulus detection. The Journal of Neuroscience 29(42):13410-13417.
Saenz M, Buracas GT, Boynton GM (2002) Global effects of feature-based attention in human visual cortex. Nature Neuroscience 5:631–632.
Sakai K, Rowe JB, Passingham RE (2002) Active maintenance in

prefrontal area 46 creates distractor-resistant memory. Nature Neuroscience 5:479–484.
Salazar RF, Dotson NM, Bressler SL, Gray CM (2012) Content-Specific Fronto-Parietal Synchronization During Visual Working Memory. Science 338:1097–1100.
Sanocki T, Sulman N (2011) Color relations increase the capacity of visual short-term memory. Perception 40:635–648.
Saproo S, Serences JT (2010) Spatial attention improves the quality of population codes in human visual cortex. Journal of Neurophysiology 104(2): 885–895.
Sauseng P, Klimesch W, Heise KF, Gruber WR, Holz E, Karim AA, Glennon M, Gerloff C, Birbaumer N, Hummel FC (2009) Brain Oscillatory Substrates of Visual Short-Term Memory Capacity. Current Biology 19:1846–1852.
Sauseng P, Klimesch W, Stadler W, Schabus M, Doppelmayr M, Hanslmayr S, Gruber WR, Birbaumer N (2005) A shift of visual spatial attention is selectively associated with human EEG alpha activity. Eur J Neurosci 22:2917–2926.
Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. Trends in Neurosciences 32:9–18.
Scolari M, Byers A, Serences JT (2012) Optimal deployment of attentional gain during fine discriminations. Journal of Neuroscience 32:7723–7733.
Scolari M, Serences JT (2009) Adaptive allocation of attentional gain. Journal of Neuroscience 29:11933–11942.
Scolari M, Serences JT (2010). Basing perceptual decisions on the most informative sensory neurons. Journal of Neurophysiology 104(4):2266–2273.
Scoville WB, Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery & Psychiatry 20:11–21.
Serences JT, Ester EF, Vogel EK, Awh E (2009) Stimulus-Specific Delay Activity in Human Primary Visual Cortex. Psychological Science 20:207–214.
Serences JT, Saproo S (2012) Computational advances towards linking BOLD and behavior. Neuropsychologia 50:435–446.
Sharbrough F, Chatrian GE, Lesser RP, Lüders H (1991) American Electroencephalographic Society guidelines for standard electrode position nomenclature. J Clin Neurophysiol 8(2):200-202.
Silvanto J, Cattaneo Z (2010) Transcranial magnetic stimulation reveals the content of visual short-term memory in the visual cortex. NeuroImage 50:1683–1689.
Silver MA, Ress D, Heeger DJ (2007) Neural correlates of sustained spatial attention in human early visual cortex. Journal of Neurophysiology 97(1):229-237.
Sims CR, Jacobs RA, Knill DC (2012) An ideal observer analysis of visual working memory. Psychological Review 119:807–830.
Sligte IG, Scholte HS, Lamme VAF (2008) Are There Multiple Visual Short-

Term Memory Stores? PLoS One 3:e1699.
Sligte IG, Scholte HS, Lamme VAF (2009) V4 Activity Predicts the Strength of Visual Short-Term Memory Representations. Journal of Neuroscience 29:7432–7438.
Sneve MH, Alnæs D, Endestad T, Greenlee MW, Magnussen S (2012) Visual short-term memory: Activity supporting encoding and maintenance in retinotopic visual cortex. NeuroImage 63:166–178.
Souza AS, Rerko L, Oberauer K (2014) Unloading and reloading working memory: Attending to one item frees capacity. Journal of Experimental Psychology: Human Perception and Performance 40:1237–1256.
Soto D, Hodsoll J, Rotshtein P, Humphreys GW (2008) Automatic guidance of attention from working memory. Trends in Cognitive Sciences 12:342–348.
Spaak E, Bonnefond M, Maier A, Leopold DA, Jensen O (2012) Layer-specific entrainment of gamma-band neural activity by the alpha rhythm in monkey visual cortex. Curr Biol 22:2313–2318.
Sperling G (1960) The information available in brief visual presentations. Psychological Monographs: General and Applied 74:1–29.
Spitzer B, Blankenburg F (2012) Supramodal Parametric Working Memory Processing in Humans. Journal of Neuroscience 32:3287–3295.
Sreenivasan KK, Curtis CE, D'Esposito M (2014a) Revisiting the role of persistent neural activity during working memory. Trends in Cognitive Sciences 18:82–89.
Sreenivasan KK, Vytlacil J, D'Esposito M (2014b) Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. Journal of Cognitive Neuroscience 26:1141–1153.
Sternberg S (1966) High-speed scanning in human memory. Science 153:652–654.
Stokes MG (2011) Top-down visual activity underlying VSTM and preparatory attention. Neuropsychologia 49:1425–1427.
Stokes MG (2015) “Activity-silent” working memory in prefrontal cortex: a dynamic coding framework. Trends in Cognitive Sciences 19:394–405.
Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J (2013) Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron 78:364–375.
Stokes M, Thompson R, Nobre AC, Duncan J (2009) Shape-specific preparatory activity mediates attention to targets in human visual cortex. PNAS 106:19569–19574.
Sugase-Miyamoto Y, Liu Z, Wiener MC, Optican LM, Richmond BJ (2008) Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. PLoS Comput Biol 4:e1000073.
Summerfield C, de Lange FP (2014) Expectation in perceptual decision making: neural and computational mechanisms. Nat Rev Neurosci 15:745–756.
Summerfield C, Koechlin E (2008) A neural representation of prior information during perceptual inference. Neuron 59(2):336–347.

Super H, Spekreijse H, Lamme VA (2001) A neural correlate of working memory in the monkey primary visual cortex. Science 293:120–124.
Suzuki M, Gottlieb J (2013) Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. Nature Neuroscience 16:98–104.
Sylvester CM, Shulman GL, Jack AI, Corbetta M (2009) Anticipatory and stimulus-evoked blood oxygenation level-dependent modulations related to spatial attention reflect a common additive signal. Journal of Neuroscience 29:10671–10682.
Swanson LW (1982) The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. Brain Research Bulletin 9:321–353.
Szczepanski SM, Knight RT (2014) Insights into Human Behavior from Lesions to the Prefrontal Cortex. Neuron 83:1002–1018.
Taulu S, Kajola M, Simola J (2004) Suppression of Interference and Artifacts by the Signal Space Separation Method. Brain Topogr 16:269–275.
Todd JJ, Marois R (2004) Capacity limit of visual short-term memory in human posterior parietal cortex. Nature 428:751–754.
Töllner T, Müller HJ, Zehetleitner M (2012) Top-Down Dimensional Weight Set Determines the Capture of Visual Attention: Evidence from the PCN Component. Cerebral Cortex 22:1554–1563.
Tong F, Pratte MS (2012) Decoding patterns of human brain activity. Annu Rev Psychol 63:483–509.
Treue S, Trujillo JCM (1999) Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399:575–579.
Tsuchida A, Fellows LK (2009) Lesion Evidence That Two Distinct Regions within Prefrontal Cortex are Critical for n-Back Performance in Humans. Journal of Cognitive Neuroscience 21(12):2263-2275.
Tuladhar AM, Huurne NT, Schoffelen J-M, Maris E, Oostenveld R, Jensen O (2007) Parieto-occipital sources account for the increase in alpha activity with working memory load. Human Brain Mapping 28:785–792.
van de Ven V, Jacobs C, Sack AT (2012) Topographic contribution of early visual cortex to short-term memory consolidation: a transcranial magnetic stimulation study. Journal of Neuroscience 32:4–11.
van den Berg R, Shin H, Chou W, George R, Ma WJ (2012) Variability in encoding precision accounts for visual short-term memory limitations. PNAS 109(22): 8780–8785.
van Dijk H, Schoffelen J-M, Oostenveld R, Jensen O (2008) Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. Journal of Neuroscience 28:1816–1823.
van Dijk H, Van Der Werf J, Mazaheri A, Medendorp WP, Jensen O (2010) Modulations in oscillatory activity with amplitude asymmetry can produce cognitively relevant event-related responses. PNAS 107:900–905.
van Moorselaar D, Olivers CNL, Theeuwes J, Lamme VAF, Sligte IG

(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.
VanRullen R, Macdonald JSP (2012) Perceptual Echoes at 10 Hz in the Human Brain. Current Biology 22:1–5.
van Velzen J, Eimer M (2003) Early posterior ERP components do not reflect the control of attentional shifts toward expected peripheral events. Psychophysiol 40:827–831.
Verstynen TD, Badre D, Jarbo K, Schneider W (2012) Microstructural organizational patterns in the human corticostriatal system. Journal of Neurophysiology 107(11):2984-2995.
Vogel EK, Machizawa MG (2004) Neural activity predicts individual differences in visual working memory capacity. Nature 428:748–751.
Vogel EK, McCollough AW, Machizawa MG (2005) Neural measures reveal individual differences in controlling access to working memory. Nature 438:500–503.
Vogels R, Orban GA (1990) How well do response changes of striate neurons signal differences in orientation: a study in the discriminating monkey. The Journal of Neuroscience 10(11):3543–3558.
Voytek B, Knight RT (2010) Prefrontal cortex and basal ganglia contributions to visual working memory. PNAS 107:18167–18172.
Wallis G, Stokes M, Cousijn H, Woolrich M, Nobre AC (2015) Frontoparietal and Cingulo-opercular Networks Play Dissociable Roles in Control of Working Memory. Journal of Cognitive Neuroscience 27(10):2019–2034.
Warden MR, Miller EK (2007) The Representation of Multiple Objects in Prefrontal Neuronal Delay Activity. Cerebral Cortex 17:i41–i50.
Watanabe K, Funahashi S (2014) Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. Nature Neuroscience 17(4):601–611.
Wei ZZ, Wang X-JX, Wang D-HD (2012) From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. Journal of Neuroscience 32:11228–11240.
Williams M, Hong SW, Kang MS, Carlisle NB, Woodman GF (2013) The benefit of forgetting. Psychonomic Bulletin & Review 20:348–355.
Wolff, M. J., Ding, J., Myers, N. E., and Stokes, M. G. (in press). Revealing hidden states in working memory using EEG. Frontiers in Systems Neuroscience.
Woodman GF, Luck SJ (1999) Electrophysiological measurement of rapid shifts of attention during visual search. Nature 400:867–869.
Woodman GF, Luck SJ (2003) Serial deployment of attention during visual search. Journal of Experimental Psychology: Human Perception and Performance 29:121–138.
Woolrich M, Hunt L, Groves A, Barnes G (2011) MEG beamforming using Bayesian PCA for adaptive data covariance matrix regularization. NeuroImage 57:1466–1479.
Worden MS, Foxe JJ, Wang N, Simpson GV (2000) Anticipatory biasing of

visuospatial attention indexed by retinotopically specific-band electroencephalography increases over occipital cortex. J Neurosci 20:1–6.
Worsley KJ, Friston KJ (1995) Analysis of fMRI time-series revisited--again. NeuroImage 2:173–181.
Wyart V, de Gardelle V, Scholl J, Summerfield C (2012) Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. Neuron 76(4):847–858.
Wyart V, Sergent C (2009) The Phase of Ongoing EEG Oscillations Uncovers the Fine Temporal Structure of Conscious Perception. Journal of Neuroscience 29:12839–12841.
Xu Y, Chun MM (2006) Dissociable neural mechanisms supporting visual short-term memory for objects. Nature 440:91–95.
Yamaguchi S, Tsuchiya H, Kobayashi S (1994) Electrooencephalographic activity associated with shifts of visuospatial attention. Brain 117:553–562.
Yeh Y-Y, Kuo B-C, Liu H-L (2007) The neural correlates of attention orienting in visuospatial working memory for detecting feature and conjunction changes. Brain Research 1130:146–157.
Yi DJ, Turk-Browne NB, Chun MM (2008) When a thought equals a look: Refreshing enhances perceptual memory. Journal of Cognitive Neuroscience 20(8):1371-1380.
Zaksas D, Pasternak T (2001) Microstimulation of cortical area MT affects performance on a visual working memory task. Journal of Neurophysiology 85(1):187-196.
Zanto TP, Gazzaley A (2009) Neural Suppression of Irrelevant Information Underlies Optimal Working Memory Performance. Journal of Neuroscience 29:3059–3066.
Zanto TP, Rubens MT, Bollinger J, Gazzaley A (2010) Top-down modulation of visual feature processing: The role of the inferior frontal junction. NeuroImage 53:736–745.
Zanto TP, Rubens MT, Thangavel A, Gazzaley A (2011) Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. Nature Neuroscience 14:656–661.
Zemel RS, Dayan P, Pouget A (1998) Probabilistic interpretation of population codes. Neural Computation 10(2):403–430.
Zhang W, Luck SJ (2008) Discrete fixed-resolution representations in visual working memory. Nature 453:233–235.
Zokaei N, Gorgoraptis N, Bahrami B, Bays PM, Husain M (2011) Precision of working memory for visual motion sequences and transparent motion surfaces. Journal of Vision 11:2–2.
Zokaei N, Manohar S, Husain M, Feredoes E (2014) Causal evidence for a privileged working memory state in early visual cortex. Journal of Neuroscience 34:158–162.

Appendix to Chapter 5 (Testing sensory evidence against mnemonic templates)

Dissociating Neural Responses to Task-irrelevant Clockwise and Counter-clockwise Angular Distances

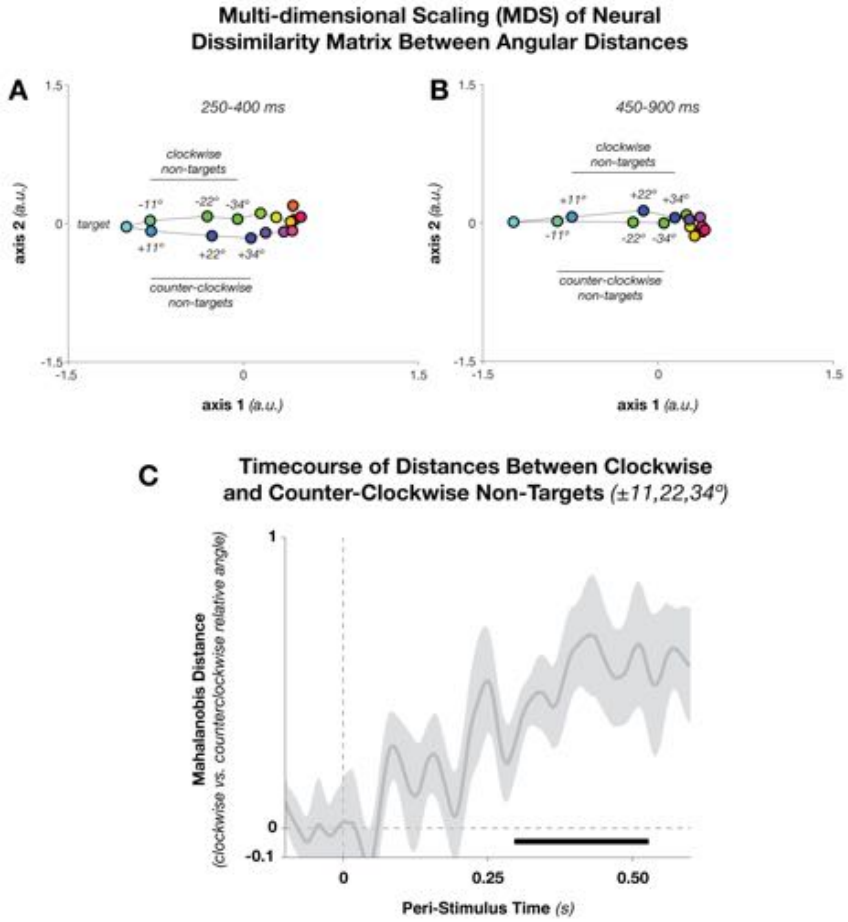


Figure A1. Multidimensional Scaling and Pairwise Mahalanobis distances between Angular Distances. A. Dissimilarity structure of angular distances. We used multi-dimensional scaling (MDS), which maps the multidimensional (32x32) Mahalanobis distance matrix between target-relative angles into two dimensions. During relatively early stimulus processing (250-400 ms after stimulus onset), geometry is elliptical – that is, in addition to conditions separating along the target-relative axis (horizontal), conditions separate along a task-irrelevant axis (vertical). During later processing stages (B: 450-900 ms), the task-related axis accounts for most of the condition differences. Since MDS is rotation-invariant, the solution in B happens to have flipped axis 2, without affecting the geometrical relationship between points. C. Mahalanobis distances (shuffle-corrected) between trials with equal target proximity, but different direction (i.e., clockwise vs. counter-clockwise deviations of the stimulus angle, with respect to the template angle). The figure shows the mean z-score (with respect to 250 random permutations of the trial labels) of pairwise distances between equal target proximities, averaged

over the pairs ±11.25°, ±22.5°, and ±33.75°. Shading indicates standard error of the mean. The black bar denotes significant time points (p<0.05, cluster-corrected).

Timecourse of Linear Decision Variable Geometry

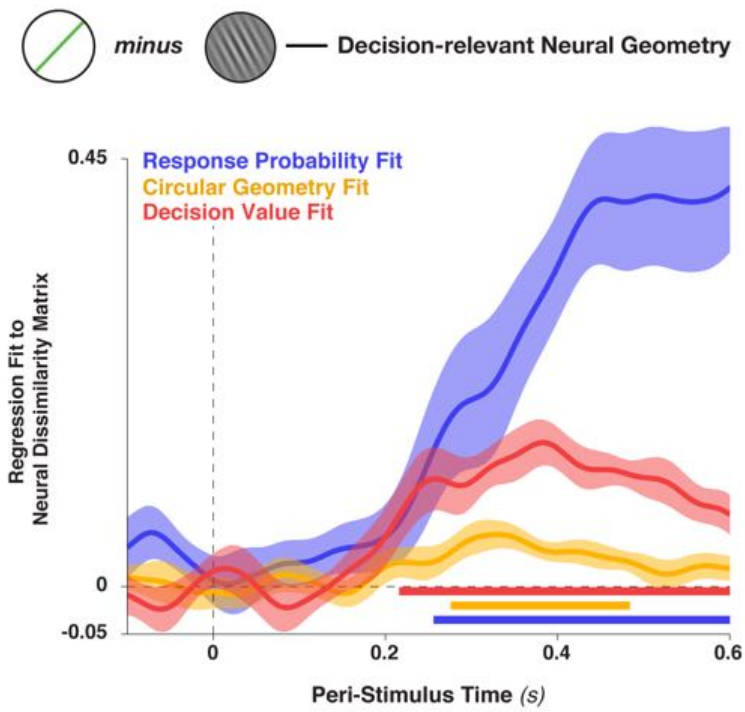


Figure A2. Figure is identical to Panel 8c, but includes in the graph the fit to the distance matrix provided by the linear decision value (i.e., the unsigned target proximity, |stimulus – target|). This variable was also included in the analysis described in the main text, but omitted from figure 8c for clarity (since it a nuisance variable). Shading indicates standard error of the mean, and colored bars at the bottom denote significant time points for each regressor (p<0.05, cluster-corrected).

Detailed description of toy model behavior.

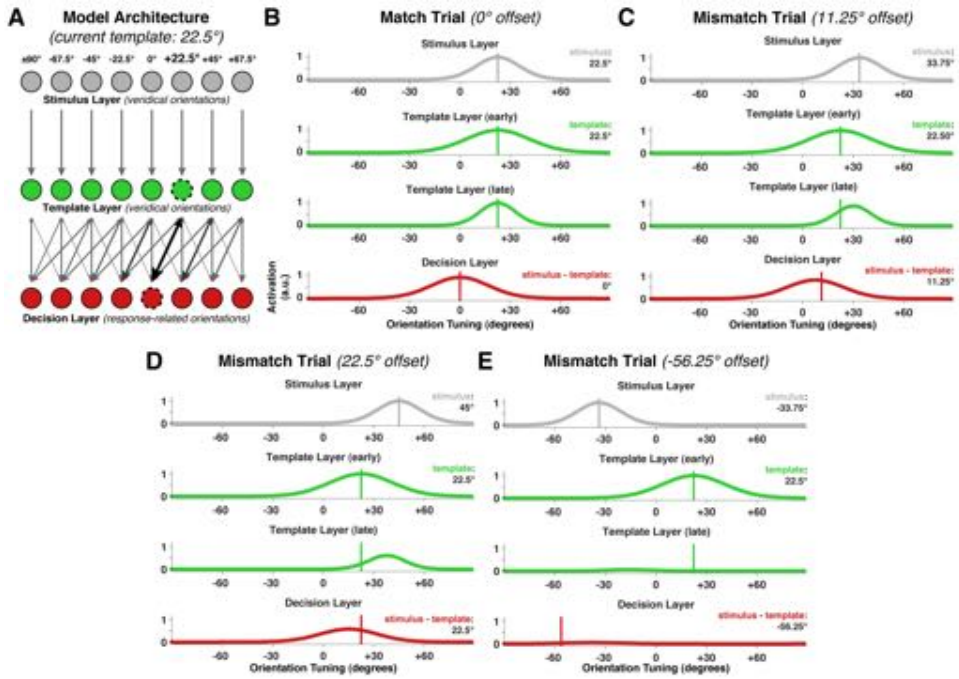


Figure A3. Neural Population Model. A. Model architecture, showing weight matrix reflecting the current template orientation. B. Model response on an exact template match trial. C-E. Model responses on mismatch trials.

The model behavior (Fig. A3b,c) followed a simple trajectory over the course of a hypothetical trial. At the beginning of the trial, before current stimulus input has been processed, the template layer encodes the current template layer, via a bump in activation in template-tuned neurons. This input can be instantiated in the template layer via top-down input from the 0° unit in the decision layer (although other mechanisms for activating the template are also conceivable, such as periodic reactivation, e.g. Buzsáki & Moser, 2013; Eichenbaum, 2013; Johnson & Redish, 2007; Lisman & Jensen, 2013; Schroeder & Lakatos, 2009). This activation of the template layer around the time of stimulus onset might correspond to the decoding profile for template information in the MEG data (Chapter 5, Fig. 3c). Next, stimulus input is represented in the stimulus layer, again via a population activity profile peaking at neurons tuned to the currently presented stimulus (again corresponding to the decoding profile, Chapter 5, Fig. 3a). Stimulus layer activation is then fed forward into the template layer, where activation is multiplied point-by-point with the existing activation state. This could happen, for example, if neurons in the template layer change their gain to all input depending on their proximity to the current target orientation (Carandini & Heeger, 2012; McAdams & Maunsell, 1999; Reynolds & Heeger, 2009; Silver, Ress, & Heeger, 2007; Treue & Trujillo, 1999). The resulting activation profile now represents the stimulus orientation, scaled by its similarity to the template

– while neurons tuned to the current orientation again have high activation, the height of that activation depends on stimulus-template similarity (and the peak is shifted towards the template orientation, with the magnitude of the shift depending on the ratio of stimulus and template tuning widths). This representation of the template and the stimulus in the same population (at slightly different but overlapping timepoints) might reflect why stimulus and template geometries cross-generalize (Chapter 5, Fig. 7g,h).

By passing the template layer profile on to the decision layer, it is shifted into a decision-relative (i.e. stimulus-invariant) space. Here, the response exhibits two decision-relevant features. First, the closer the current stimulus is to the template, the closer the decision layer peak is to the 0° neuron. Second, as in the template layer, the height of the decision layer profile also depends on the proximity between stimulus and template, with highest activation for targets (Fig. A3b), and the peak dropping off for increasingly distant non-targets (Fig. A3c-e). This decreasing amplitude may explain why near-targets were more separable than definite non-targets later in the trial epoch (Chapter 5, Fig. 7b). Downstream read-out units could use either the population's *peak location* (i.e. how close the maximum response is to the 0° unit, as in Beck et al., 2008) or its *peak activation* (as in more classical decision models) to determine whether a target is present or absent.