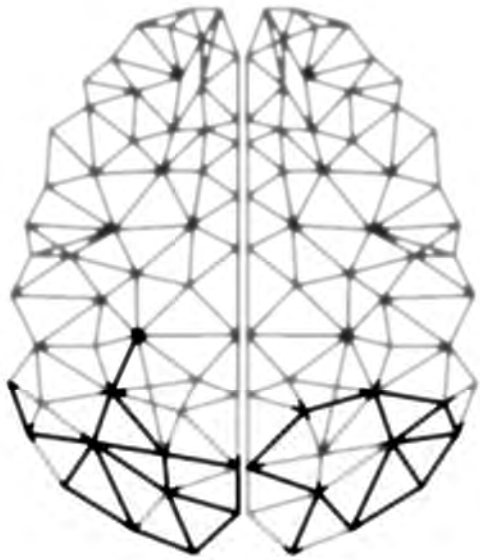
Pinging the brain to reveal hidden working memory states

Wolff, Michael

DOI:
10.33612/diss.151472370

Pinging the brain

to reveal hidden working memory states



Michael J. Wolff

Cover & layout design by Michael J. Wolff

Printed by Off Page (offpage.nl)

ISBN: 978-94-034-2515-0 (paperback)

ISBN: 978-94-034-2515-3 (eBook)

university of groningen

Pinging the brain to reveal hidden working memory states

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Thursday 18 February 2021 at 11.00 hours

by

Michael Josef Wolff

born on 24 September 1986
in Bad Driburg, Germany

Supervisor

Prof. M. M. Lorist

Co-supervisors

Dr. E.G. Akyürek

Assessment Committee

Prof. R. de Jong

Prof. C. Olivers

Prof. A. Compte

Contents

Chapter 1

General Introduction

1

Working memory (WM) comprises the temporary storage and manipulation of sensory information (Baddeley & Hitch, 1974). The ability to keep past experience "online" even though the sensory information is no longer present in the environment is a remarkable achievement of the brain. It provides the functional basis that moves past simple reflexive actions, and towards complex, goal-directed behaviours, and has been studied extensively by psychologists and neuroscientists for more than 50 years.

The "working" in working memory implies that information are not only memorized, but also "worked" with; manipulated, transformed, and used to best suit future behavioural demands. WM is thus a core cognitive function and essential for a wide range of cognitive tasks, such as learning, planning, language comprehension, and problem solving (Baddeley, 1992; Kane & Engle, 2002) and WM impairments are present in many cognitive disorders (Devinsky & D'Esposito, 2003) as well as healthy aging (Gazzaley, Cooney, Rissman, & D'Esposito, 2005).

The amount of information that can be maintained in WM at any one time is surprisingly limited, however. Its capacity has been estimated to only span between 3 to 5 independent pieces of information (Cowan, 2010), which is in stark contrast to the essentially limitless storage of long-term memory. WM capacity can vary substantially between individuals and is a strong predictor of intelligence and academic achievement (Conway, Kane, & Engle, 2003; Rohde & Thompson, 2007). Given this limited capacity, it is important that only behaviourally relevant information enters WM, and irrelevant information is filtered out. WM is thus closely intertwined with attentional processes, which play a role at all WM stages to use the limited resources most efficiently (e.g., Awh, Vogel, & Oh, 2006; Cowan, 2011). Researching the neuroscience of WM therefore requires taking into account its interplay with attention, which can often change the interpretation of neural data (e.g. Lewis-Peacock, Drysdale, Oberauer, & Postle, 2011).

Given its essential role in functional behaviour, it is of no surprise that establishing the neural basis of WM has been one of the main goals in cognitive neuroscience. Of interest is where and how in the brain information is maintained and potentially manipulated. Early neural recordings pointed towards a relatively simple and intuitive explanation, suggesting that persistent neural activity in the prefrontal cortex (PFC) maintains information in WM (Fuster & Alexander, 1971; Kubota & Niki, 1971), a view that has dominated the WM research literature for many decades (Curtis & D'Esposito, 2003). However, the recent surge of sophisticated analysis techniques that can be employed with the ever increasing computational power of modern hardware, and which can - directly link neural activity patterns to specific WM content, has cast doubt on this classic theory (Miller, Lundqvist, & Bastos, 2018; Stokes, 2015).

Finding the neural correlate of WM

Elevated neural activity in PFC

Almost half a decade ago, a group of researchers independently made discoveries (Fuster, 1973; Fuster & Alexander, 1971; Kubota & Niki, 1971), that set the stage for decades of research on persistent neuronal activity during WM.

Fuster and Alexander (1971) had monkeys perform a classic version of a now popular spatial WM task: An object is shown at a random location, in full view of the subject. Shortly after, the visibility of the object at the specific location is obstructed for a short time. At the end of this delay/maintenance period, location choices are presented, and the subject is instructed to select the location at which the object had appeared previously. While the monkeys performed many trials of this task, the activity of individual neurons in the PFC was recorded. The researchers found that a subset of the recorded neurons showed elevated and persistent spiking activity from the onset of the to-be-remembered location until the end of the maintenance period. The researchers proposed that this sustained activity is playing a major role in keeping the location-specific information in WM, until it is no longer needed.

These and similar findings and interpretations highlighted the importance of persistent neural activity in the PFC for WM and has shaped the WM research literature over the years accordingly. Numerous studies show elevated PFC activity during a variety of WM tasks (Curtis & D'Esposito, 2003; Funahashi, Bruce, & Goldman-Rakic, 1989; Goldman-Rakic, 1995) and lesion studies have found a causal relationship between specific parts of the PFC and WM tasks (Bauer & Fuster, 1976; Chao & Knight, 1998; Funahashi, Bruce, & Goldman-Rakic, 1993; M. H. Miller & Orbach, 1972). Thus, over the years, the PFC has clearly been the focus of WM research, which, without a doubt, has established the fundamental importance of the PFC for WM. However, it has not been entirely clear in which of the many aspects of WM it plays a role. During a WM task, subjects do not only need to maintain the object or location in WM, but pay continuous attention to the task, prepare to respond, encode only relevant information and disregard irrelevant information, and remember the tasks rules. For example, it has been found that it is not necessarily the retention of information that patients with frontal lesions struggle with, but rather the inhibition of irrelevant information, leading to a decrease in WM task performance (Chao & Knight, 1998). Furthermore, prefrontal activity has been found to ramp up in anticipation of the probe at the end of the trial (K. Watanabe & Funahashi, 2007), as if reflecting the preparation of the response. Similarly, in a human fMRI study it was found that activity in PFC only showed an increase when participants had to memorize a sequence of actions, not when memorizing a simple visuospatial stimulus (Pochon et al., 2001). Findings such as these suggest a more complex role of PFC activity, but the mixed nature of the results to date make it a challenge to define its exact role in WM.

Thus, even though it had to be acknowledged that PFC activity can vary substantially depending on what kind of WM task is performed, it was largely assumed

that PFC is the neural substrate of WM maintenance (Curtis & D'Esposito, 2003). This was called into question however with the first paper that successfully used multivariate pattern analysis (MVPA) to "decode" the content of WM from recorded brain activity (Harrison & Tong, 2009).

PFC: the executive control station?

Research in the cognitive neuroscience has largely depended on finding neural activity in specific brain regions that increases or decreases in response to experimental manipulation. However, with the introduction of MVPA in cognitive neuroscience, spearheaded by Kamitani and Tong (2005), researchers did not have to rely on finding univariate difference in neural activity levels between experimental conditions, but could instead look for differences in neural activation *patterns* between tasks or conditions. By relating the neural activation patterns to individual items in WM, MVPA can be used to find brain areas that represent the actual content of WM.

Numerous non-human primate, single-unit studies found that spatial locations, dissociated from motor preparations, are coded in PFC (Mendoza-Halliday, Torres, & Martinez-Trujillo, 2014; Qi, Meyer, Stanford, & Constantinidis, 2011; Rainer, Asaad, & Miller, 1998), as well as colour (Buschman, Siegel, Roy, & Miller, 2011), and natural images (Meyer, Qi, Stanford, & Constantinidis, 2011; Rao, Rainer, & Miller, 1997). However, these neural representations might not necessarily reflect the low-level sensory information, but rather more abstract neural representations of task-relevant dimensions and task requirements (Lara & Wallis, 2014). For example, Riggall and Postle (2012) found that while BOLD activity obtained in the PFC did not represent the visual information of the WM task, it did reflect the current task rules. Lee and colleagues (Lee, Kravitz, & Baker, 2013) also found that PFC activity reflected semantic information and not low level visual information. Buschman and colleagues (Buschman, Denovellis, Diogo, Bullock, & Miller, 2012) found that neural oscillations in the beta and alpha range represent specific and dynamically changing task rules during a perceptual task. In general, it seems that abstract, sensory independent information are coded in PFC, suggesting an executive role of the PFC during WM tasks (Postle, 2016). Indeed, many PFC neurons display mixed selectivity, demonstrating heterogeneous coding of different features of a cognitive task, and non-linear interactions between task relevant aspects (Fusi, Miller, & Rigotti, 2016; Mante, Sussillo, Shenoy, & Newsome, 2013; Rigotti et al., 2013). The PFC is thus able to flexibly code for task categories, even when the presented stimuli are exactly the same (Freedman, Riesenhuber, Poggio, & Miller, 2001; McKee, Riesenhuber, Miller, & Freedman, 2014). Similarly, arbitrary boundaries on continuous scales are flexibly adapted in neural population activity in PFC (Wutz, Loonis, Roy, Donoghue, & Miller, 2018).

Due to this highly heterogeneous coding of PFC activity across and within WM tasks, and the inconsistent findings regarding low level feature coding (Christophel, Klink, Spitzer, Roelfsema, & Haynes, 2017), PFC activity might more appropriately be regarded as the "central executive" of the classic model of WM, and not one of the

storage modules (Baddeley, 1992; Serences, 2016). The central executive is thought to be responsible for controlling and regulating the WM system, by (among other things) tracking and updating task demands and utilizing the limited cognitive resources efficiently, which corresponds with the heterogeneous findings of PFC activity during WM tasks discussed above.

Persistent, content-specific activity in sensory cortex

In their spearheading study, published a decade ago, Harrison and Tong set out to find the brain area that maintains the neural representation of information in WM (Harrison & Tong, 2009). Human participants in their study undertook a now widely used retro-cue WM task designed to dissociate stimulus driven effects from working-memory processes while fMRI was recorded. In each trial, two randomly orientated gratings were presented serially. After a short delay a retro-cue (a number) indicated which of those two orientations is relevant and would be tested later, rendering the other one irrelevant. It was found that while the blood oxygen level dependent (BOLD) brain signal obtained from the visual cortex increased during grating presentation, it returned almost to baseline levels during the delay period, as if playing no role during WM maintenance. Even so, the pattern of activity across voxels in the visual cortex coded for the relevant orientation grating throughout the delay, but not for the irrelevant item. This provides evidence that relevant, low level visual information in WM seem to be maintained in the same part of the brain that is also responsible for its processing, which is referred to as the sensory recruitment hypothesis (Serences, Ester, Vogel, & Awh, 2009). Findings such as these have led to a revision of neurophysiological network of WM maintenance, and it has been proposed the visual cortex is integral in the maintenance of visual information (Gayet, Paffen, & Van der Stigchel, 2018; Scimeca, Kiyonaga, & D'Esposito, 2018).

While WM research is largely dominated by the visual domain, it has more recently been found that persistent neural activity in the auditory cortex reflects the maintenance of specific tones in WM (Huang, Matysiak, Heil, König, & Brosch, 2016; Kumar et al., 2016; Uluç, Schmidt, Wu, & Blankenburg, 2018). Additionally, some limited evidence for persistent stimulus-selective activity for vibration in the secondary somatosensory cortex has also been found (Hernández et al., 2010), providing evidence for the sensory recruitment hypothesis from non-visual modalities (Christophel et al., 2017; Serences, 2016).

It makes intuitive sense for the brain to recruit the same areas during WM that are also optimized to process the fine details of the sensory information during perception, enabling the retention of detailed sensory information that do not afford an easy semantic transformation (Ester, Sprague, & Serences, 2015). This negates the need to make a copy of the initial perceptual response in another equally sensitive brain area, by exploiting already existing structures for multiple purposes.

However, the sensory-recruitment of WM hypothesis does not go unchallenged. It has been argued that WM maintenance should be extremely vulnerable to external distraction if the very area that processes externally presented information, also

maintains internally held information (Xu, 2017), and it is therefore necessary to make a copy of the information elsewhere. Indeed, it has been found that the presentation of a irrelevant visual distractor during the maintenance period of visual information disrupts and abolishes the visual WM related signal in the visual cortex, while the visual WM related signal remained robust in the superior intraparietal sulcus (Bettencourt & Xu, 2016), suggesting that the visual cortex is not necessary for maintenance. However, it has been suggested, that this can be regarded as flexible change in coding schemes, from an exact visual representation to a higher level of abstraction (Scimeca et al., 2018). Furthermore, visual distractors passively viewed during the maintenance period of visual information impact behavioural precision (Kiyonaga & Egner, 2016), and distractors that match WM content easily capture attention (Soto, Hodsoll, Rotshtein, & Humphreys, 2008), which indeed suggests an overlap and cost between stimulus maintenance and stimulus processing within the same neural network.

Not so persistent delay activity

As explained above, the classic model of WM that has dominated the literature for several decades, posits that persistent neural activity keeps information "online" in WM until it is no longer needed. This is not surprising, given that the PFC, as well as the sensory cortices and the parietal cortex have been shown to seemingly exhibit persistent WM related activity (Curtis & D'Esposito, 2003; Ester et al., 2015; Goldman-Rakic, 1995; Harrison & Tong, 2009; Wimmer, Nykamp, Constantinidis, & Compte, 2014). Recently, this has been called into question however (Lundqvist, Herman, & Miller, 2018) and it has been proposed that WM can be maintained in an "activity-silent" neural network (Stokes, 2015). But how can this hypothesis be reconciled with the overwhelming evidence of persistent delay activity? Three design-related arguments may be brought forward.

First, individual neurons that exhibit persistent delay activity reported and highlighted in many classic studies (e.g. Bauer & Fuster, 1976; Fuster, 1973), only make up a small proportion of the neurons that are modulated by WM operations, most of which only spike sporadically (Shafi et al., 2007).

Second, as discussed above, a lot of the evidence for persistent activity comes from studies that employed the memory-guided saccade task (Funahashi, Bruce, & Goldman-Rakic, 1989), which cannot dissociate between motor preparation and WM maintenance. Even to this day, it is still used to falsely assert that WM maintenance is mediated through persistent neural activity (Inagaki, Fontolan, Romani, & Svoboda, 2019; Wimmer et al., 2014), which may be so for motor preparation, but not necessarily WM maintenance.

Third, persistent delay activity can be an artefact of trial averaging (Stokes & Spaak, 2016). In cognitive neuroscience, subjects usually complete many trials of the same task and the recorded brain activity is averaged over trials. This is often necessary in order to isolate consistent task related signals from the noisy data. However, this assumes that the neural signal in question is time-locked to a specific event across all trials, which does not necessarily need to be the case. Thus, if a specific neural event occurs at a slightly

different time-point in each trial, averaging over trials will create the illusion of a sustained signal. This is exactly what Lundqvist and colleagues (Lundqvist et al., 2016) found. The averaged neural delay activity recorded from the PFC while monkeys performed WM tasks suggested sustained, content-specific activity in a broad range in the gamma-frequency, replicating previous findings of sustained gamma-power during WM maintenance (e.g. Howard et al., 2003). However, individual trials exhibited narrow, short-lived bursts of gamma-activity that carried information about WM content, interleaved by baseline-level activity states, providing counter-evidence to the sustained activity of WM account (Bastos, Loonis, Kornblith, Lundqvist, & Miller, 2018; Lundqvist, Herman, Warden, Brincat, & Miller, 2018)

Delay activity: Nothing but attention?

There is another, even more fundamental issue with the interpretation of persistent delay activity as reflecting WM maintenance. Most studies discussed thus far have employed simple experimental paradigms to measure WM related neural activity, where subjects usually maintained a single piece of information over short period of time before it is tested. However, WM is more than a simple single-item storage and it is now clear that attentional processes play a major role in WM, which cannot be dissociated when only a single item is maintained in WM. Two now classic behavioural experiments have found that attention can be used to focus on an individual item in WM (Griffin & Nobre, 2003; Landman, Spekreijse, & Lamme, 2003). More specifically, participants performed a now widely used retro-cue WM task, where the locations of multiple items need to be maintained over a short period of time. During the maintenance period, the retro-cue indicates with above chance probability which of the multiple items held in WM is most likely to be tested at the end of the trial. It was found that valid retro-cues (i.e. the cued item was tested) lead to an increase in performance, suggesting some sort of attentional enhancement or reformatting (Myers, Stokes, & Nobre, 2017) of the cued item during maintenance that cannot be explained by differential processing during item presentation. These results have sparked an enormous interest in the interplay between attention and WM (Gazzaley & Nobre, 2012; Souza & Oberauer, 2016) and has led to WM models that propose multiple states in WM (e.g. Oberauer & Hein, 2012), where information can be maintained in an "attended" and "unattended" WM state. Since these states cannot be dissociated when only a single item is memorized, and which is likely in the "attended" state, it begs the question what the neural representation of WM items that are not in the focus of attention is.

Single item experiments have been a popular choice in the quest to find the neural correlate of WM maintenance for a reason; their simple nature increases the chances of finding WM related neural traces. More complex paradigms are clearly needed to dissociate the neural signature of memoranda and attention, however. This is indeed what Watanabe and Funahashi (Watanabe & Funahashi, 2014) did. They recorded neural activity from the PFC from monkeys that completed a dual task that added an attentional task to the classic memory-guided saccade task (Goldman-Rakic, 1995). Each trial began

with the cueing of a random location (attention cue), and monkeys were instructed to release a lever when the colour changed, but not look at it. After the onset of the attention cue, but before its colour change, another random location was cued for a short time (memory cue) and the monkeys had memorize its location after it disappeared until the end of the trial when it was reported with a saccade. Crucially, during the overlap of the two tasks, monkeys had to maintain the location, while at the same time pay close attention to the cued location in order to be ready for its colour change. Neural activity in PFC exhibited clear location discrimination of the memory cue during and shortly after its presentation. However, it decreased to almost baselines levels during the retention and attention period. Interestingly, the memory-specific signal in PFC "reawakened" immediately after the completion of the attention task (i.e. after lever release), and presumably when monkeys could fully focus on maintaining the memorized location. Thus, even though the location was stored in memory, its code in PFC activity was almost non-existent as long as attention was preoccupied with another task, suggesting that PFC activity mainly reflects *attended* WM content.

But what about other brain areas? Lewis-Peacock and colleagues (Lewis-Peacock et al., 2011) used BOLD activity from the whole brain, obtained while human participants completed visual WM tasks, to decode the categorical memory items, without having an explicit hypothesis about the location of the neural representation of WM content. Participants had to memorize two memory items, both of which were tested. During the delay, retro-cues guided the internal focus of attention towards one item by indicating which of the two would be tested by the upcoming probe. Only the item that was in the focus of attention exhibited a corresponding neural trace, while the unattended item did not, as if forgotten. However, once a retro-cue redirected attention to the previously unattended item, its neural trace was reactivated. These findings were later replicated using electroencephalography (LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012). Similarly, a sophisticated analyses technique that can reconstruct the remembered location from BOLD activity in visual areas (including IPS), found that the neural representation of two locations in WM degraded gradually over time (Sprague, Ester, & Serences, 2016). However, once one of the two items could be dropped, the neural reconstruction of the remaining location in memory recovered, presumably because attentional resources could be focused on one item, instead of two.

Synaptic model of WM

It is intuitively appealing to assume that the neural mechanism of maintaining a specific item in WM is stable, item-specific neural activity, as if to keep a freeze-frame snapshot of past stimulation "online" until it is no longer needed. However, even without considering the recently emerging evidence that this may not be so, some issues with this theory are evident. If a minuscule interruption of the item-specific activity would mean an inevitable loss of it from WM, WM maintenance would be expected to be extremely prone to distractors. Additionally, computational models that are based on persistent activity have difficulty with the maintenance of more than one item, in particular when

there is overlap in their neural representations (Edin et al., 2009). Finally, persistent neural activity is metabolically expensive.

As discussed above, WM-related activity is heavily modulated by attention and essentially non-existent for unattended WM content (Larocque, Lewis-Peacock, & Postle, 2014), and even when it is attended, WM maintenance is accompanied only by sparse, short-lived activity bursts (Lundqvist, Herman, & Miller, 2018). Due to these observations it has been proposed that WM maintenance can occur within an "activity-silent" network (Miller et al., 2018; Stokes, 2015), which could be accomplished via transient changes in connectivity in the WM network (Mongillo, Barak, & Tsodyks, 2008). In the synaptic model of WM, relevant information that is encoded in WM leave behind an "impression" in the wiring pattern of the WM network. A biologically viable mechanism for this are calcium kinetics that afford short-term synaptic plasticity (STSP; Zucker & Regehr, 2002), that could last for approximately ~ 1 second (Catterall, Leal, & Nanou, 2013; Mongillo et al., 2008), rendering continuous neural activity unnecessary for maintenance. Item-specific activity-bursts strengthen or reinstate this connectivity periodically before it dissipates (Lundqvist et al., 2016). The behavioural relevance of individual items in WM dictates how often the corresponding item-specific connections are refreshed (Larocque et al., 2014), so that currently irrelevant and thus unattended items are maintained in an almost exclusively activity-silent state, while currently relevant information are maintained in an active neural state (Fig. 1.1).

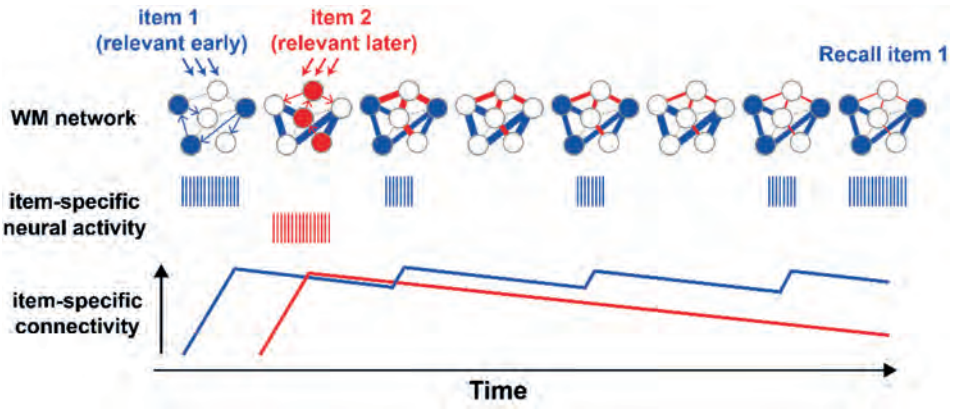


Figure 1.1. Synaptic model of WM highlighting the interplay between synaptic- and activity-states as a function of behavioural relevance. Item-specific neural activity triggers item-specific connectivity. While the attended and imminently relevant item (in blue) is maintained in an active state that periodically strengthens the item-specific connectivity, the unattended item exhibits no item-specific activity during maintenance (red).

The presence of STSP in the brain has been well established (e.g. Hempel, Hartman, Wang, Turrigiano, & Nelson, 2000; Sugase-Miyamoto, Liu, Wiener, Optican, & Richmond, 2008; Zanos, Rembado, Chen, & Fetz, 2018), and numerous (computational) models of WM have been proposed that depend on STSP (Barak & Tsodyks, 2014; Barak, Tsodyks, & Romo, 2010; Buonomano & Maass, 2009; Lundqvist, Herman, & Lansner, 2011; Manohar, Zokaei, Fallon, Vogels, & Husain, 2017; Miller et al., 2018; Stokes, 2015). Providing a direct link between STSP and WM is difficult, however. In order to show that two neurons are connected, the demonstration of correlated sinking activity between those neurons is necessary, which only an extremely small proportion of recorded neurons demonstrate (Fujisawa, Amarasingham, Harrison, & Buzsáki, 2008), making it unfeasible to establish WM-related connectivity changes in the non-human primate model, where only a limited number of neurons can be recorded simultaneously. However, researchers were able to demonstrate short-term plasticity in rat PFC during a WM maze task (Fujisawa et al., 2008).

WM maintenance as a state-dependent neural response

The synaptic model proposes not only that any neural activation pattern elicited either internally or externally leaves behind a transient neural trace of said pattern, but that it is also modulated by the current state of the network (Buonomano & Maass, 2009; Mongillo et al., 2008; Sugase-Miyamoto et al., 2008). That is, each activity state leaves behind a neural trace that in turn modifies the activity pattern of subsequent neural activity occurring in the same neural network within a short time-span, leading to a unique impulse response that is an interaction between the input and the current state of the neural system. Indeed, it has been found that the evoked neural activity in cat visual cortex not only codes the currently presented visual stimulus, but also the stimulus presented a few hundred milliseconds earlier (Nikolic, Haeusler, Singer,, & Maass, 2007). It has also been found that the presentation of a neutral stimulus presented during the delay of a WM task resulted in a neural response in the PFC that reflected the content of WM (Stokes et al., 2013).

It has been suggested that this property is not just an inevitable side-effect of fast synaptic modulation, but could also serve an efficient read-out mechanism of the WM network in response to external stimulation. Indeed, it was found that while the neural response to the target stimulus in a WM task at first reflected the physical properties of the stimulus, it quickly evolved into a context-depended neural signal reflecting the behaviourally relevant dimension (Stokes et al., 2013). Additionally, it has been found that the WM state can act like a matched filter, filtering external stimulation in such a way that it automatically leads to a behaviourally relevant output signal (Myers et al., 2015).

To sum up, research has thus far mainly relied on measurable, item-specific neural activity to find the neural correlate of WM. As argued, this will likely not draw the complete picture, however, as WM maintenance should not only be understood as the literal maintenance of sensory information through neural activity, but also as the

reconfiguration of the WM network to best reflect future behavioural demands as a state-dependent neural response (Myers et al., 2017; Stokes, 2015). This thesis explicitly tests and exploits this state-dependent neural response during WM maintenance to reveal potentially hidden WM states.

Thesis overview

This thesis employs MVPA on electrophysiological data obtained through EEG to investigate item-specific neural responses during perception and maintenance. However, by the time the PhD that culminated in this thesis began, this method was mainly used in fMRI research but rarely considered for MEG or EEG, which have notoriously bad spatial resolution. Chapter 2 highlights research (Cichy, Ramirez, & Pantazis, 2015) that provides evidence that the MEG signal is nevertheless spatially specific enough to uncover fine-grained neural differences elicited by the cortical columns that respond to tilted lines in the visual cortex. The chapter employs simple modelling to further demonstrates that the same may hold true for EEG, thus highlighting the feasibility of employing MVPA on EEG data, which is exploited in all subsequent chapters.

Chapter 3 explicitly tests the proposed network-specific neural response to external stimulation. During the delay period of a simple, single item WM task, the same high contrast "impulse" stimulus was presented in every trial, and of interest was if its evoked neural response measured with EEG contained information about the previously presented and memorized randomly orientated grating. Using MVPA it was found that this was indeed the case, providing simple proof of principle for a powerful and relatively simple approach to infer possibly hidden neural states through external perturbation.

Chapter 4 uses the "impulse" approach introduced in the previous chapter to further explore the hidden WM state across multiple experiments. Using a retro-cue paradigm that dissociates stimulus-driven from WM-related neural effects (Harrison & Tong, 2009), it is tested whether the impulse response actually reflects WM content and is not simply a reactivation of stimulus-history. It is furthermore tested if the WM-related impulse is dependent on WM-related delay activity, or if it is also reflects unattended, but nevertheless memorized WM content, using a attentional priority paradigm (Lewis-Peacock et al., 2011).

The previous two chapters established that a visual impulse stimulus present during the delay period of a visual WM task results in a neural response that reflects WM content. Chapter 5 tests on the one hand if the same holds true for the auditory counterpart, i.e. testing the hypothesis that an auditory impulse stimulus presented during the delay of an auditory WM reflects auditory WM content. It furthermore tests if auditory and visual WM content is maintained in a sensory-specific neural network. This done not by looking for confirmatory WM-specific delay activity in sensory areas (Kumar et al., 2016; Scimeca et al., 2018; Xu, 2017), but rather by assessing if sensory specific and sensory non-specific bottom-up neural responses are WM-specific.

One of the mysteries of WM is its limitation (Cowan, 2010). The quality of even a single remembered item gradually decays over time, which can be measured in free-recall

paradigms (Rademaker, Park, Sack, & Tong, 2018). Modelling work suggests that this is due to random drift along the continuous item dimension in the neural population code (Schneegans & Bays, 2018) but neurophysiological evidence is limited (Wimmer et al., 2014). Chapter 6 employs the impulse approach to enhance the neural representation of orientations at different times during the delay period of a free-recall WM task. It is explicitly tested if reports that are clockwise or counter-clockwise relative to the correct orientation are accompanied by a corresponding shift in the neural representation.

Finally, Chapter 7 summarizes and discusses the research results presented in this thesis.

Chapter 2

Decoding rich spatial information with high temporal resolution

This chapter was previously published as:

Stokes, M. G., Wolff, M. J., & Spaak, E. (2015). Decoding rich spatial information with high temporal resolution. *Trends in cognitive sciences*, *19*(11), 636-638.

2

Abstract

New research suggests that magnetoencephalography (MEG) contains sufficient spatial information for decoding the orientation of visual stimuli. As with multivariate pattern analysis in functional magnetic resonance imaging, subtle but consistent differences in the distribution of orientation columns generate subject-specific patterns of activity. This implies MEG (and electroencephalography: EEG) is ideal for decoding neural states in the human brain.

Keywords: Neural decoding; multivariate pattern analysis; orientation tuning; magnetoencephalography; electroencephalography; spatiotemporal information.

A major challenge in cognitive neuroscience is to discriminate brain states with high spatial and temporal resolution. These two dimensions of high resolution are often considered mutually exclusive for non-invasive human studies. Functional magnetic resonance imaging (*fMRI*) can resolve detailed spatial patterns of activity, but has notoriously poor temporal resolution; whereas methods that track electrical activity provide rich temporal information, but lack spatial precision. However, a recent paper by Cichy et al invites us to re-evaluate this classic dichotomy. Using a combination of empirical data and theoretical modelling, they argue that the signals measured with magnetoencephalography (MEG) actually contain rich spatial information that can be used to differentiate extremely subtle neural states (Cichy et al., 2015). This could be a game changer for high-temporal resolution methodologies that have been long considered too coarse to resolve fine-scale neural coding.

Just over a decade ago, fMRI experienced a minor revolution inspired by a relatively simple idea: idiosyncratic patterns of activity carry important information. The test case was orientation decoding. It turns out that activity patterns in visual cortex can reliably predict the orientation of a grating stimulus presented to the subject (e.g. Kamitani & Tong, 2005). The general importance of this finding lies in its broader implication. Different orientations are not represented in different brain areas, but within narrow cortical columns that are distributed throughout the retinotopic landscape of visual cortex. Therefore, if it is possible to decode the orientation of a grating stimulus in visual cortex, perhaps it is also possible to decode other distributed, and spatially overlapping neural states, and in other brain areas. In the extreme, fMRI suddenly appeared to carry informational content comparable to the gold standard single unit recordings in non-human primates (Kriegeskorte, Mur, Ruff, et al., 2008).

The key insight for the fMRI community was that subtle biases in the distribution of neurons tuned to one feature or another could lead to subtle differences in the activity of a sampled voxel (schematised in Fig. 2.1 A). Although such biases would be weak, they could be pooled together over a number of samples (i.e., voxels) to statistically differentiate activity patterns. This approach has come to be known as multivariate pattern analysis (Haxby, Connolly, & Guntupalli, 2014), and has changed the way people think about fMRI. Decoding overlapping population codes for orientation encouraged the field to think more about information coded in a pattern of activity rather than differences in mean activity in certain brain areas (Kriegeskorte, Goebel, & Bandettini, 2006).

As orientation decoding was the test-ground for fine-scale pattern decoding in fMRI, Cichy et al. set out to show that MEG could also be used to decode spatially overlapping neural states. Other studies have shown that orientation information can be decoded from the visual evoked response in MEG and EEG using multivariate pattern analysis (Ramkumar, Jas, Pannasch, Hari, & Parkkonen, 2013; Wolff, Ding, Myers, & Stokes, 2015/Chapter 3). However, there are a number of possible confounds that were raised in the fMRI literature that could potentially explain orientation decoding based on coarse spatial differences (e.g., coarse-scale activity differences due to the over-

representation of cells tuned to particular orientations; (Freeman, Heeger, & Merriam, 2013)). Cichy and colleagues systematically address a large number of such possible confounds, concluding each time that MEG is able to decode genuine information about the orientation of presented stimuli. The authors concede that it is impossible to claim that their efforts were exhaustive. Indeed, just like the fMRI debate, it is likely that other potential explanations will surface, and would need to be addressed in future studies. Notwithstanding this caveat, Cichy et al present an impressive set of experiments all seemingly pointing to an important conclusion: MEG can resolve spatially overlapping representations.

As reviewed above, previous fMRI studies argued that orientation decoding is driven by subtle differences in sampling small-scale biases in the distribution of tuned cells. However, the spatial resolution of MEG is far coarser than fMRI. So what is the mechanism that could explain genuine orientation decoding? Cichy et al propose a surprisingly simple idea (schematised in Fig. 2.1 B).

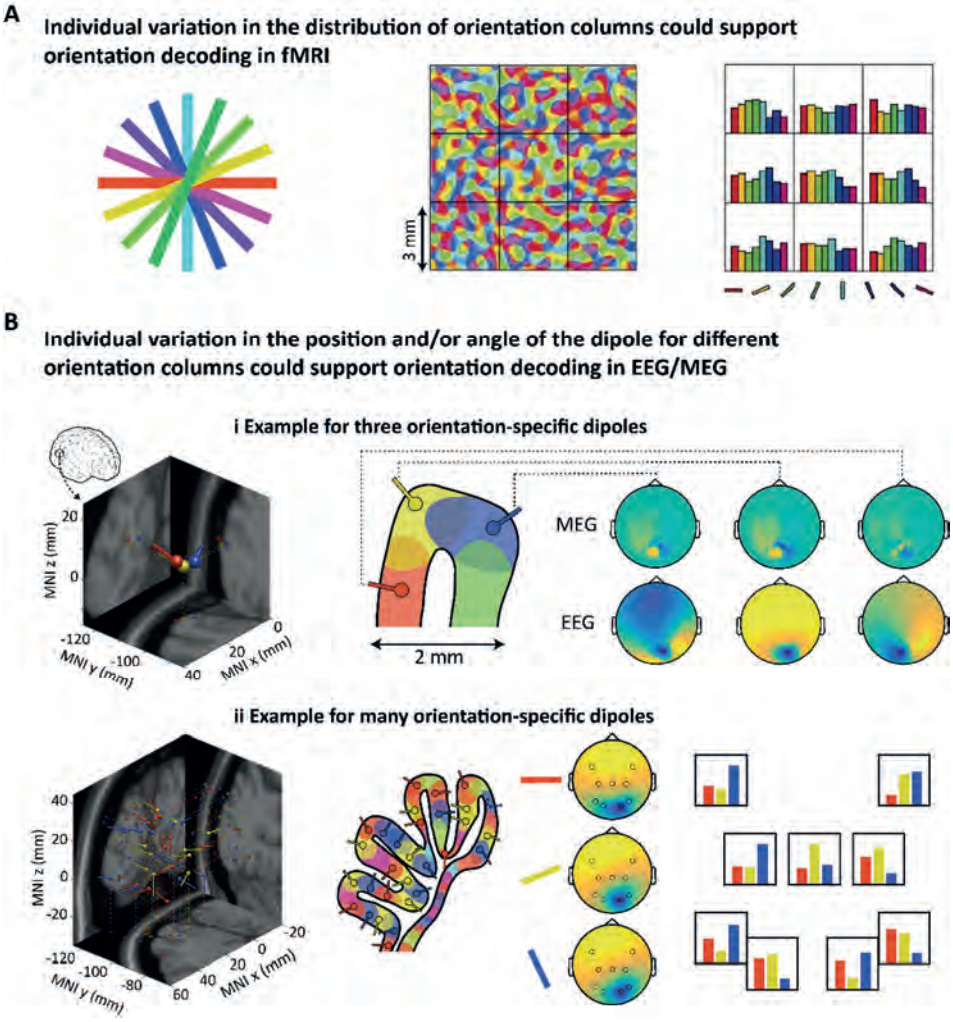


Figure 2.1. Condition specific activity patterns in fMRI, EEG, and MEG. **(A)** Figure adapted taken from (Norman, Polyn, Detre, & Haxby, 2006). Simulated orientation map in V1 (middle). Even though each voxel samples the activity of many orientation columns, the activity patterns across voxels are orientation specific (right). **(B) i.** Three dipoles approximately 2 mm apart result in distinguishable MEG/EEG topographies, due to their different orientations. **ii.** Thirty dipoles are randomly placed within a 40 mm^3 cube, of which 10 belong to one of three orientation conditions. Each condition yields highly similar EEG topographies. However, the relative activity histograms of 9 sensors over the three orientations are separable.

It is well-established that electrical activity in aligned cells generates a dipole which projects to the scalp surface. EEG measures the resultant electric potential at the scalp surface, whereas MEG measures the magnetic field. The spatial distribution of the field depends on the location of the dipole, but critically, also on its angle. Cichy and colleagues argue that because the surface of the cortex is irregular, even dipoles from neighbouring clusters of cells will have different angles, resulting in separable field patterns at the scalp surface (see Fig. 2.1 B i). Although these patterns will be idiosyncratic to a given subject (depending on subtle differences in cortical folding), systematic differences within participants can be differentiated using multivariate classification. So exactly like MVPA for fMRI, it should be possible to differentiate spatially overlapping brain states by analysing subject-specific patterns (see Fig. 2.1 B ii), even though group differences would typically just average out.

If MEG/EEG can be a source of such rich spatial information, then why are these non-invasive methods so often considered to have poor spatial resolution? The classic problem limiting spatial resolution in MEG/EEG is source ambiguity. Strictly, it is not possible to localise with certainty the source of the field measured at the scalp surface. There is no unique solution, but theoretically infinitely many solutions that could generate the same pattern of observed activity. To reverse engineer the location of the source from the observed scalp distribution runs up against the obstinate inverse problem. Although sophisticated methods have been developed to constrain probabilistic solutions (e.g. López, Litvak, Espinosa, Friston, & Barnes, 2014), the inherent uncertainty results in a relatively coarse estimate of the underlying source. However, if the purpose of the analysis is to track differential brain states over time, rather than localise activity differences, then the inherent ambiguity hardly matters.

We predict that multivariate decoding will revolutionise MEG/EEG just as it did fMRI. The key insight is that these measures contain rich spatial information, even if the source localisation is inherently ambiguous. As the fMRI community has moved from localising blobs of condition-specific differences to measuring information coded in activity patterns, so the MEG/EEG will embrace MVPA for decoding neural states. Moreover, coupled with the exquisite temporal resolution inherent to electromagnetic indices of brain activity, MEG/EEG could really become the method of choice for exploring the spatiotemporal dynamics of human brain activity.

Acknowledgments

We would like to thank the Biotechnology & Biological Sciences Research Council (to M.G.S), and Frederik van Ede for helpful discussion. The dipole simulation was performed using the FieldTrip toolbox (http://www.ru.nl/neuroimaging/fieldtrip).

Chapter 3

Revealing hidden states in visual working memory using electroencephalography

This chapter was previously published as:

Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience*, *9*, 123.

Data and code available at osf.io/g2jen

Abstract

It is often assumed that information in visual working memory (vWM) is maintained via persistent activity. However, recent evidence indicates that information in vWM could be maintained in an effectively "activity-silent" neural state. Silent vWM is consistent with recent cognitive and neural models, but poses an important experimental problem: how can we study these silent states using conventional measures of brain activity? We propose a novel approach that is analogous to echolocation: using a high-contrast visual stimulus, it may be possible to drive brain activity during vWM maintenance and measure the vWM-dependent impulse response. We recorded electroencephalography (EEG) while participants performed a vWM task in which a randomly oriented grating was remembered. Crucially, a high-contrast, task-irrelevant stimulus was shown in the maintenance period in half of the trials. The electrophysiological response from posterior channels was used to decode the orientations of the gratings. While orientations could be decoded during and shortly after stimulus presentation, decoding accuracy dropped back close to baseline in the delay. However, the visual evoked response from the task-irrelevant stimulus resulted in a clear re-emergence in decodability. This result provides important proof-of-concept for a promising and relatively simple approach to decode "activity-silent" vWM content using non-invasive EEG.

Introduction

Visual Working memory (vWM) is essential for high-level cognition. By keeping task-relevant information in mind, vWM provides a functional basis for complex behaviors based on timeextended goals and contextual contingencies. Some of the most influential models of vWM are built on the intuitive notion that maintenance is directly related to the persistence of stationary activity states, representing specific content in vWM from the moment of encoding until that content is needed for behaviour (Curtis & D'Esposito, 2003; Goldman-Rakic, 1995). Persistent activity models have obvious appeal - vWM effectively preserves a freeze-frame snapshot of past experience until it is no longer required. However, there are gaps in the argument for persistent activity models of vWM.

Accumulating evidence suggests that vWM is not always accompanied by persistent delay activity (Sreenivasan, Curtis, & D'Esposito, 2014). For example, a recent study in non-human primates showed that content-specific delay activity can be effectively abolished during dual task interference, even though vWM-guided behavior is relatively spared (Watanabe & Funahashi, 2014). Robust delay activity returned when attention was refocused on the vWM- task. Similarly, human studies using non-invasive brain imaging suggest that activity patterns during maintenance delays correspond only to attended items (Lewis-Peacock et al., 2011). Unattended items do not seem to have a corresponding activity state, even though such unattended items are still maintained in vWM (Larocque et al., 2014; Olivers, Peters, Houtkamp, & Roelfsema, 2011). As in the non-human primate study, the activity state of unattended items becomes apparent once attention is directed to them (Lewis-Peacock et al., 2011; Lewis-Peacock & Postle, 2012).

These results suggest that delay activity is not strictly necessary for maintenance in vWM. Dissociating vWM-performance from persistent delay activity implies that some form of "activity-silent" neural state contributes to maintenance in vWM (Stokes, 2015). For example, a synaptic model of vWM proposes that information is encoded in item-specific patterns of functional connectivity (Mongillo et al., 2008; Sugase-Miyamoto et al., 2008). Essentially, activity patterns during encoding drive content-specific changes in short-term synaptic plasticity (Zucker & Regehr, 2002). Although the temporary synaptic trace is effectively "activity silent," this hidden neural state can be read out from the network during processing of a memory probe. Mongillo et al. (2008) focused on known mechanisms of short-term synaptic plasticity; however, other neurophysiological factors could also pattern hidden states for vWM-guided behavior (Buonomano & Maass, 2009). The key principle is that activity-dependent changes in the hidden neural state could be important for maintaining information in vWM.

One reason that persistent-activity models of vWM have been so pervasive in the past is that it is much easier to find confirmatory evidence with conventional measures, such as elevated delay-period firing (Fuster & Alexander, 1971) or pattern decoding during the delay period (Harrison & Tong, 2009). Disconfirmatory evidence is essentially

a null effect. Therefore, to evaluate the possible contributions of hidden states to vWM maintenance, it is necessary to develop measures that are capable of revealing them. Previously, it was found that a neutral task-irrelevant stimulus presented during a vWM delay period generated vWM-specific patterns of activity in monkey prefrontal cortex (Stokes et al., 2013). We suggested that this context-dependent response pattern could reflect differences in hidden state. For illustration, consider echolocation (e.g., sonar), where a simple impulse (e.g., "ping") is used to probe hidden contours of unseen structure. Analogously, the impulse response to neural perturbation should co-depend on the pattern of input activity and the hidden state of the network. If the input pattern is held constant, we can attribute differences in the output to underlying changes in hidden state.

In the current study, we develop this idea further using a task-irrelevant visual stimulus (or "impulse stimulus") to drive a vWM-specific impulse response function that could be measured non-invasively using EEG. Participants performed a two-alternative vWM discrimination task that requires precise maintenance of the orientation of a memory item during a delay interval (Bays & Husain, 2008). Critically, on a subset of trials we presented a fixed high-contrast impulse stimulus designed to drive neural activity in the visual system. We predicted that the evoked response should differentiate the memory condition (i.e., the remembered orientation), even in the absence of vWM-discriminative delay activity.

To anticipate the results, multivariate decoding at posterior electrodes accurately discriminated the orientation of the memory item during stimulus encoding. Consistent with previous evidence for dynamic coding in neural populations (Meyers, Freedman, Kreiman, Miller, & Poggio, 2008; Stokes et al., 2013) and scalp-level patterns (Cichy, Chen, & Haynes, 2011), the discriminative patterns were dynamic during stimulus processing. After the initial dynamic trajectory, discrimination decayed to near-baseline levels during the delay period. Importantly, the impulse stimulus reactivated vWM-specific activity patterns, consistent with the hypothesis that vWM content could be stored in an "activity-silent" neural format. Interestingly, although the impulse response pattern differentiated the vWM-stimulus, the discriminative pattern did not match the patterns during memory encoding. This experiment provides a novel proof-of-concept of a potentially powerful method for inferring hidden neural states.

Methods

Participants

Twenty-four healthy adults (12 female, mean age 22.2 years, range 18 – 38 years) were included in the experiment and analyses. During recruitment, four additional participants were excluded from all analyses due to excessive eye-movements and eye-blinks (more than 20 % of trials were contaminated). All participants received a monetary

compensation of £10/hour and gave written informed consent. The study was approved by the Central University Research Ethics Committee of the University of Oxford.

Apparatus and Stimuli

The experimental stimuli were generated and controlled with the freely available MATLAB extension Psychophysics Toolbox (Brainard, 1997) and presented at a 100 Hz refresh rate and a resolution of 1680 x 1050 on a 17" Samsung SyncMaster 2233. A USB keyboard was used for response input. The viewing distance was set at 64 cm.

A grey background (RGB = [150 150 150]) was maintained throughout the experiment. Memory items were circular sine-wave gratings presented at a 20 % contrast. The memory probes were circular, 100 % contrast gratings underlying a square-form function. The radius and spatial frequency was fixed for both types of stimuli (2.88°, and 0.62 cycles per degrees), and the phase was randomized. The memory items' orientations were uniformly distributed, and angle difference between memory item and probe within each trial was uniformly distributed across 20 angle differences (±4°, ±5°, ±7°, ±9°, ±12°, ±15°, ±20°, ±26°, ±34°, ±45°). The impulse item was a high-contrast, black-and-white round "bull's-eye" in the same size and spatial frequency as the memory items and probes. All stimuli were presented centrally. Accuracy feedback was given with high (880 Hz) and low (220 Hz) tones for correct and incorrect responses, respectively.

Procedure

Participants were seated in a comfortable chair and the keyboard was placed either on their lap or on a table in front of the participants. The participants' task was to memorize the orientation of the presented low-contrast grating and to press the "m" key with the right index finger if the probe was rotated clockwise and the "c" key with the left index finger if the probe was rotated counter-clockwise relative to the previously presented memory item. They were instructed to respond as quickly and as accurately as possible.

Each trial began with the presentation of a fixation cross, which stayed on the screen until probe presentation. After 1,000 ms the memory item was presented for 200 ms. In half of the trials (long), the following delay period was 2600 ms, after which the probe was presented for 200 ms. In the delay period at either 1,170 (early-impulse) or 1,230 ms (late-impulse) after the memory item, the impulse stimulus was presented for 200 ms (Fig. 3.1 A), which the participants were instructed to ignore. The temporal jitter was introduced to allow us to test whether any effect on stimulus decoding was specifically time-locked to the impulse. In the other half of trials (short), the response probe was presented 1200 ms after memory item (Fig. 3.1 B). This trial length condition was included to ensure that participants were paying attention in the delay period of the long trials, thus increasing the potential effect of the impulse. After probe offset, the screen remained blank until response-input. A feedback tone was then played for 100 ms and the next trial automatically began after 500 ms. Every 24 trials a performance summary screen, with the average accuracy and median reaction of all trials thus far, was shown. Participants could use this moment to take short breaks. The trial conditions

were randomized across the entire session and participants completed 1600 trials in total (400 early-impulse trials, 400 late-impulse trials, and 800 short trials) over a time period of approximately 165 min (including breaks).

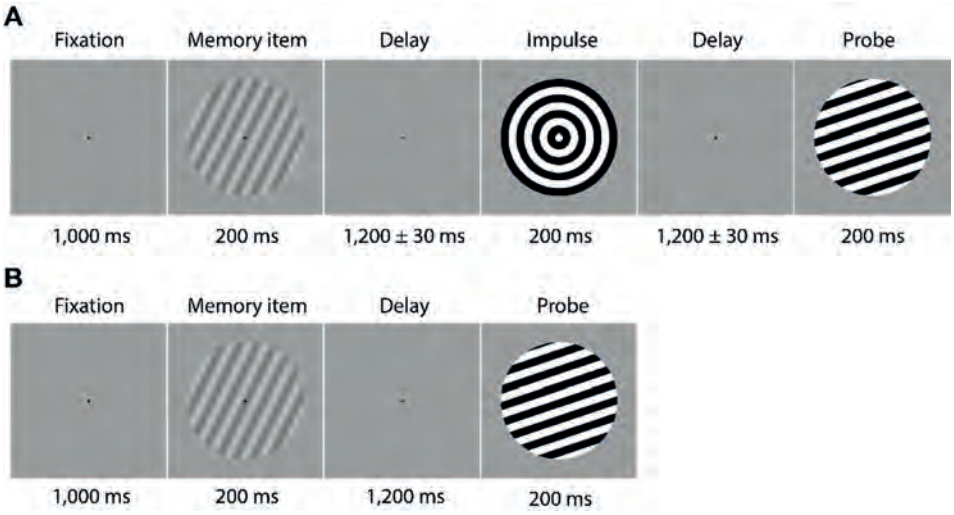


Figure 3.1. Trial structure. Participants memorized the orientation of a low contrast sine-wave grating. **(A)** In half of the trials a neutral impulse stimulus was shown after the initial delay. The onset of the impulse was jittered by ±30 ms. The force-choice discrimination memory probe was presented after a second delay period. **(B)** In the other half of the trials, determined randomly, the probe was presented instead of the impulse after the first delay.

Behavioural Analysis

Memory performance was analysed with the freely available MATLAB extension MemToolbox (Suchow, Brady, Fougnie, & Alvarez, 2013). The standard mixture model of visual working memory (Zhang & Luck, 2008) was fit separately for each participant ($N = 24$) and trial-length condition. The model assumes that the distribution of response errors has two distinct causes: (1) Pure guesses, which result in a uniform distribution of errors across all angle differences in the forced-choice paradigm. (2) Variability in the precision of the remembered item, which, even though the item is memorized, can result in errors at particularly small angle differences between memory item and probe. Although the main purpose of this analysis was simply to confirm that our participants could reliably memorize the low-contrast memory item in this experiment, for

completeness we also performed paired-samples t-tests on guess rate and memory variability between trial-length conditions.

EEG Acquisition

The EEG was recorded using NeuroScan SynAmps RT amplifier and Scan 4.5 software (Compumedics NeuroScan, Charlotte, NC) from 61 Ag/AgCl sintered surface electrodes (EasyCap, Herrsching, Germany) laid out according the to the extended international 10–20 system (Sharbrough et al., 1991) at 1000Hz. The anterior midline frontal electrode (AFz) was reserved as the ground. Electrooculography (EOG) was recorded from electrodes placed below and above the right eye and from electrodes placed to the left of the left eye and to the right of the right eye. Impedances were kept below 5 kΩ. Data were filtered online using a 200 Hz low-pass filter and the electrodes were referenced to the right mastoid.

EEG Preprocessing

Offline, the signal was re-referenced to the average of both mastoids, down-sampled to 250 Hz with 16-bit precision and band pass filtered (0.1 Hz high-pass and 40 Hz low-pass) using EEGLAB (Delorme & Makeig, 2004). The data were then epoched from -200 ms to 1,400 ms relative to the onset of the memory item for the short, no-impulse trials, and from -200 ms to 2,800 ms for the long, impulse trials. Both long and short epochs were then baseline-corrected using the 200 ms prior to memory item onset. Subsequent artefact detection and trial rejection was performed via visual inspection and focused exclusively on the EOG channels and the 17 posterior channels of interest included in the analyses (P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, O2). Trials containing saccadic eye-movements at any point in time, blinks during stimulus presentation, or other non-stereotyped artefacts were rejected from all further analyses. Impulse trials were subsequently re-epoched to two shorter epochs, time-locked to the memory item (-200 ms to 1,400 ms) or to the impulse stimulus (-200 ms to 1,400 ms). Finally, the data were smoothed with a Gaussian kernel (SD = 8 ms).

Multivariate Pattern Analysis

To determine whether the pattern of the EEG signal across the posterior channels of interest contained information about the remembered item, we used the Mahalanobis distance (De Maesschalck, Jouan-Rimbaud, & Massart, 2000; Mahalanobis, 1936) to perform pair-wise comparisons between sets of trials in which orthogonal orientations were presented.

Trials were divided across four angle bins two times and only orthogonal angle bins were compared in the multivariate analysis (0º to 45º versus 90º to 135º; 45º to 90º versus 135º to 180º; -22.5º to 22.5º versus 67.5º to 112.5º and 22.5º to 67.5º versus 112.5º to 157.5º). For illustration, see Fig. 3.2 for the event-related potentials of occipital electrodes (O1, Oz and O2) for each pairwise comparison between orthogonal angle-bins.

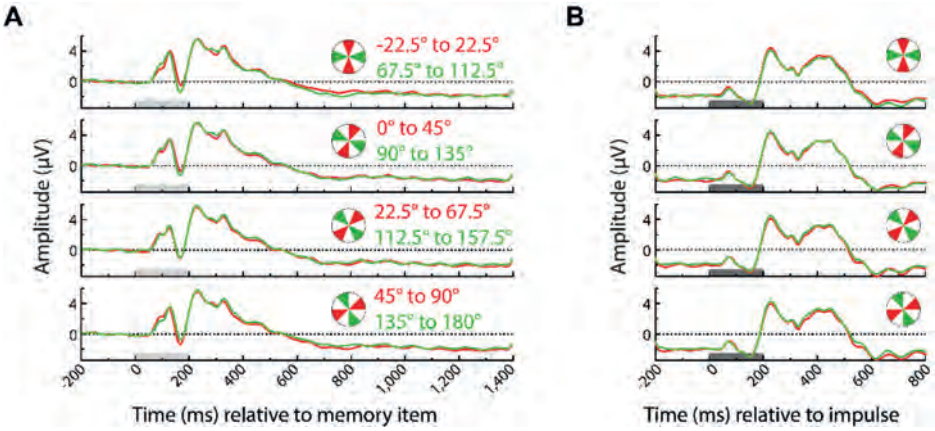


Figure 3.2. Event-related potentials of each angle bin averaged over the occipital channels (O1, Oz, and O2). Illustrated are all pairwise orthogonal angle bin comparisons that were made in the multivariate analysis of the memory item segment **(A)** and impulse segment **(B)**. Light-gray and dark-gray bars represent memory item and impulse presentations, respectively.

We used a leave-one-trial-out cross-validation approach to calculate, on each trial, the multivariate dissimilarity (Mahalanobis distance) of that trial to the average of all other trials in the same angle bin, relative to the dissimilarity of that trial to the average of all trials in the orthogonal angle bin. Mahalanobis distances of the test trial were computed for each time point as follows:

$$D1=\sqrt{(\text{Train angle 1} - \text{Test trial})^T * pC^+ * (\text{Train angle 1} - \text{Test trial})}$$

$$D2=\sqrt{(\text{Train angle 2} - \text{Test trial})^T * pC^+ * (\text{Train angle 2} - \text{Test trial})}$$

where "Train angle 1" and "Train angle 2" are row vectors containing the average signals of angle bins 1 and 2 (excluding the test trial) of each channel, and "pC^+" is the pseudo inverse of the error covariance matrix. The error covariance was estimated by pooling over the covariances of each angle condition, estimated from all trials within each condition (excluding the test trial) using a shrinkage estimator that is more robust than the sample covariance for data sets with many variables and/or few observations (Kriegeskorte et al., 2006; Ledoit & Wolf, 2004). The variables "Train angle 1", " Train angle2" and "pC^+" are all part of the training set, on which "Test trial", a row vector containing the signal of each channel of the left-out test-trial, is tested on. This was done by computing the difference between the two Mahalanobis distances between "Test trial" and "Train angle 1" (D1) and "Test trial" and "Train angle 2" (D2). The same-angle bin distance was always subtracted from the orthogonal-angle bin difference (so if the "Test trial" was part of angle bin 1 then D1 would be subtracted from D2). If the signal indeed contained information about the memory item at that time point, this

distance difference should be positive (because the orthogonal-angle bin distance should be higher than the same-angle bin distance). See Figure 3.3 for a schematic overview of the analysis. This procedure was performed for all trials and all previously defined angle bin comparisons, resulting in two equivalent estimates of distance differences per trial. Observed distances were then averaged over the two estimates, and across trials, to derive a single value for each time point and each participant for subsequent statistical testing and plotting.

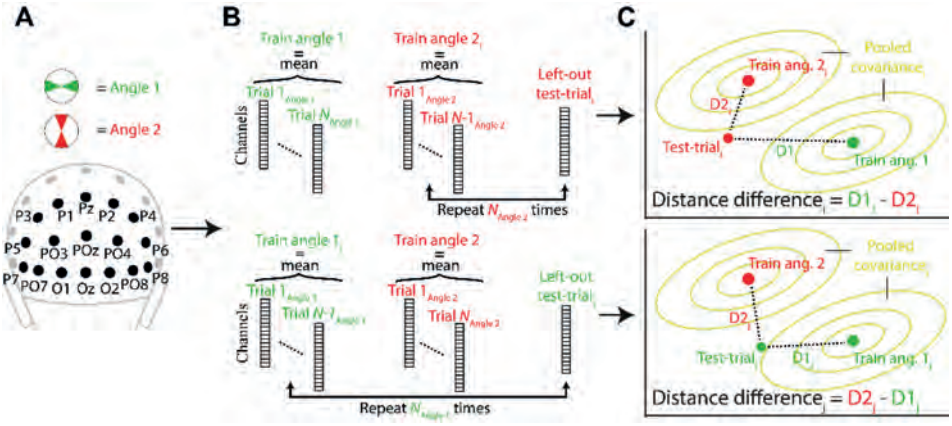


Figure 3.3. A schematic representation of the trial-wise Mahalanobis distance analysis. **(A)** The signal for two orthogonal angle bins (angle 1 and angle 2) was extracted from 17 posterior channels at a specific time point. **(B)** A single trial was either removed from angle 2 (top; test-trial$_i$) or angle 1 (bottom; test-trial$_j$) and the mean signal for each angle condition of all other trials made up the training set (train angle 1, train angle 2). **(C)** The Mahalanobis distances of the left-out test-trial to train angle 1 (D1) and train angle 2 (D2) illustrated in two-dimensional space. The pooled covariance is computed from the trials underlying train angle 1 and 2 and is recomputed for each new test. When the test trial belongs to angle bin 2, D2$_i$ is subtracted from D1$_i$ (top), when it belongs to angle bin 2, D1$_j$ is subtracted from D2$_j$ (bottom). This procedure is repeated for each trial and time-point and the resulting distance differences are averaged across all trials.

Cross-temporal Analysis

To explore the dynamics of information processing, and to test if the informative signal cross-generalizes to other time points (King & Dehaene, 2014), we computed a cross-temporal extension of the Mahalanobis analysis described above. The difference between condition-specific distances was computed as described above. However,

instead of training and testing only on the same equivalent time points, train/test sliding windows were decoupled: The training data consisting of "Train angle 1," "Train angle 2" and the corresponding pseudo inverse of the covariance matrix (as described above) at train time *Y* was used to compute the distances to the test-trial at test time *X* (e.g. Stokes et al., 2013). After computing the distance differences for all possible train-test time combinations and averaging across all test trials, the results were combined into a cross-temporal matrix in which differences along the diagonal correspond directly to the time-resolved analyses already discussed, but off- diagonal coordinates reflect the extent to which the underlying discriminative neural patterns cross-generalize between train-test time points. This cross-temporal analysis was carried out within each trial epoch separately (memory-item and impulse), as well as across epochs, where the train data was taken from the impulse epoch and tested on all trials within the memory item epoch and vice versa, resulting in four cross-temporal discrimination matrices.

Univariate Analysis

To explore to what extent the differences in the EEG signal between memory items is driven by amplitude rather than pattern differences, we performed the univariate equivalent to the multivariate analysis described above. Instead of calculating the difference between the same- and orthogonal-angle bin Mahalanobis distances, the difference between the absolute same- and orthogonal-angle bin voltage differences averaged across all 17 posterior channels was computed.

Significance Testing

Statistics of one-dimensional EEG-analyses were inferred non-parametrically (Maris & Oostenveld, 2007) with sign-permutation tests. For each time-point, the decoding value of each participant was randomly multiplied by 1 or -1. The resulting distribution was used to calculate the p-value of the null-hypothesis that the mean discrimination-value was equal to 0. Cluster-based permutation tests were then used to correct for multiple comparisons across time using 10,000 permutations, with a cluster-forming threshold of $p < 0.01$. The significance threshold was set at $p < 0.05$ and all tests were two-sided. Significance tests were carried out separately for the memory item (0 – 1,400 ms) and the impulse (0 – 800 ms). The sample size of all tests was 24.

Data Sharing

In accordance with the principles of open evaluation in science (Walther & van den Bosch, 2012), all data and fully annotated analysis scripts from this study are publicly available at

http://datasharedrive.blogspot.co.uk/2015/05/revealing-hidden-states-in-working.html.

We also hope these data and analyses will provide a valuable resource for future re-use by other researchers. In line with the OECD Principles and Guidelines for Access

to Research Data from Public Funding (Pilat & Fukasaku, 2007), we have made every effort to provide all necessary task/condition information within a self-contained format to maximise the re-use potential of our data. We also provide fully annotated analysis scripts that were used in this paper.

Results

Behavioural Results

Visual working memory performance (Fig. 3.4A) was modelled separately for short and long trials, each consisting of 800 trials. The difference in guess rates for short (M = 0.074, SD = 0.048) and long trials (M = 0.073, SD = 0.047) was not statistically different ($t(23) = 0.182$, $p = 0.858$). On the other hand, the standard deviation of remembered items (sd) was significantly different between trial length conditions ($t(23) = 2.458$, p = 0.022): sd was lower for short trials (M = 4.272, SD = 1.318) than for long trials (M = 4.927, SD = 1.292; Fig. 3.4B). Whether this decrease in precision in long trials is due to the increase in trial duration (Zhang & Luck, 2009) or the possible interference effect of the impulse stimulus (Magnussen, Greenlee, Asplund, & Dyrnes, 1991) cannot be concluded, as the present study was not designed to address this issue.

The very low guess rates in both conditions provided evidence that the participants had little difficulty to reliably memorize the low contrast angle stimuli. Because most errors were attributed to noise in mnemonic precision rather than absolute forgetting, we included both incorrect and correct trials in all EEG analyses.

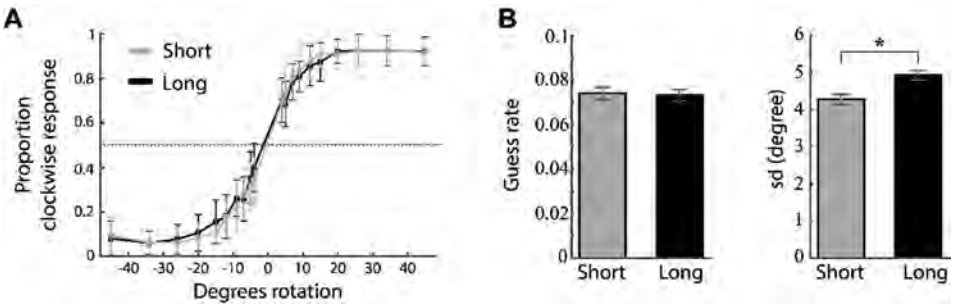


Figure 3.4. Behavioral performance and model parameters. **(A)** Mean proportion of clockwise responses as a function of angle difference between memory item and probe plotted separately for short (grey) and long (black) trials. Error bars are standard deviations. **(B)** Guess rates and memory variability (sd) for short and long trials estimated by the standard mixture model of working memory. Long trials result in significantly higher sd than short trials. Error bars are normalized standard errors.

Memory Item Discrimination during and after Item Presentation

The averaged trial-wise difference in Mahalanobis distances between across- and within-angle conditions enabled us to decode the memory items from the EEG signal of the posterior channels as a function of time. A statistically significant cluster emerged 68 ms after memory item onset, and lasted until the end of this epoch (1,400 ms, cluster $p < 0.001$; Fig. 3.5A, cyan). Because the impulse analysis was only based on 50% of trials, we also analysed the memory encoding effect only on corresponding long trials (Fig. 3.5A, blue), enabling a power-matched comparison between the memory item- and impulse-epoch. This revealed several significant decoding clusters: 76 to 632 ms ($p < 0.001$), 668 to 720 ms ($p = 0.023$), 756 to 788 ms ($p = 0.047$), 876 to 936 ms ($p = 0.016$), and 964 ms to 1,000 ms ($p = 0.036$).

Memory Item Discrimination during and after Impulse Presentation

The same analysis as above was performed on the subsequent epoch for long trials, time-locked to the impulse onset. Significant temporal clusters of above-chance discrimination were detected at 140 to 408 ms ($p < 0.001$) and 424 to 508 ms ($p = 0.005$ after impulse onset (Fig. 3.5B, blue, bottom).

Decoding Accuracy Increases Significantly after Impulse Presentation

Since the decoding accuracy does not seem to drop completely to chance levels in the initial delay period, we also tested whether the presentation of the impulse results in a significant *increase* in discriminability. To this end, we subtracted the mean discriminability between -100 ms and 0 ms prior to impulse onset from the discrimination values after impulse onset. Two significant clusters were identified: 188 to 232 ms ($p = 0.012$) and 364 to 404 ms ($p = 0.016$). These results confirm that discrimination accuracy increased significantly after impulse presentation (Fig. 3.5B, blue, top).

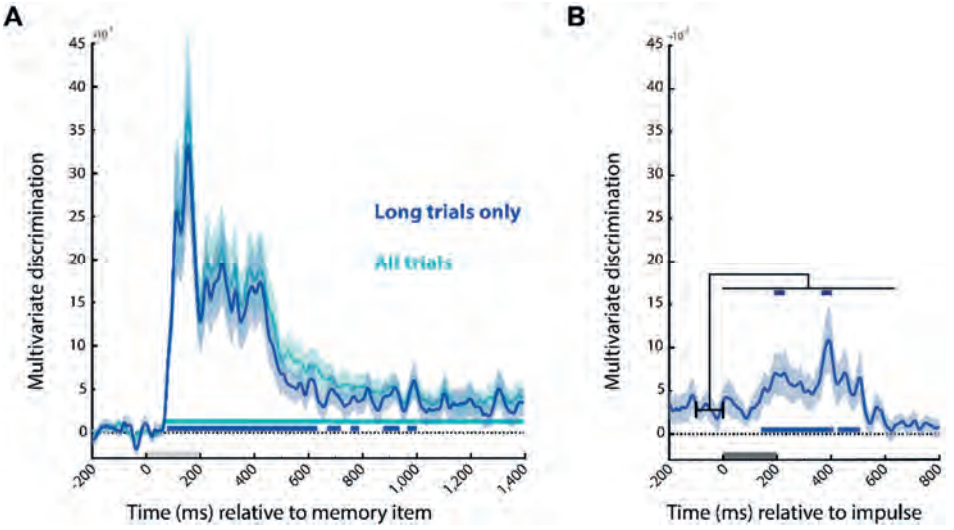


Figure 3.5. Multivariate discrimination of the memory item across time. **(A)** Memory item epoch. The discrimination for both trial types (in cyan), and exclusively for the long trials used in the impulse response analysis (in blue). Significant positive clusters are marked with bars in the corresponding colors. **(B)** Impulse epoch. The discrimination of memory item is shown for long trials (in blue), with positive clusters are marked in the corresponding significance bar along the bottom. Significant increases in discrimination compared to the mean discrimination 100 ms prior to impulse onset are indicated with dark-blue bars at the top. Light-gray and dark-gray bars represent memory item and impulse presentation, respectively. Error bars are standard deviations from the permuted null-distributions.

The Memory Item and Impulse Show Dynamic coding

The cross-temporal analysis of the memory item epoch using both long and short trials showed a dynamic coding pattern. Discrimination was greatest when trained and tested on the same time-points, as opposed to different time-points (Fig. 3.6A, lower left). The impulse response, though weaker than the memory item response, suggested a dynamic coding pattern as well (Fig. 3.6A, upper right).

Memory Item and Impulse Coding Do Not Cross-generalize

We saw no evidence for cross- generalization between the neural patterns evoked by the memory stimulus and the impulse response, either when the training set was taken from the impulse epoch and tested on the memory item epoch (Fig. 3.6A, top left), or the other way around (Fig. 3.6A, bottom right).

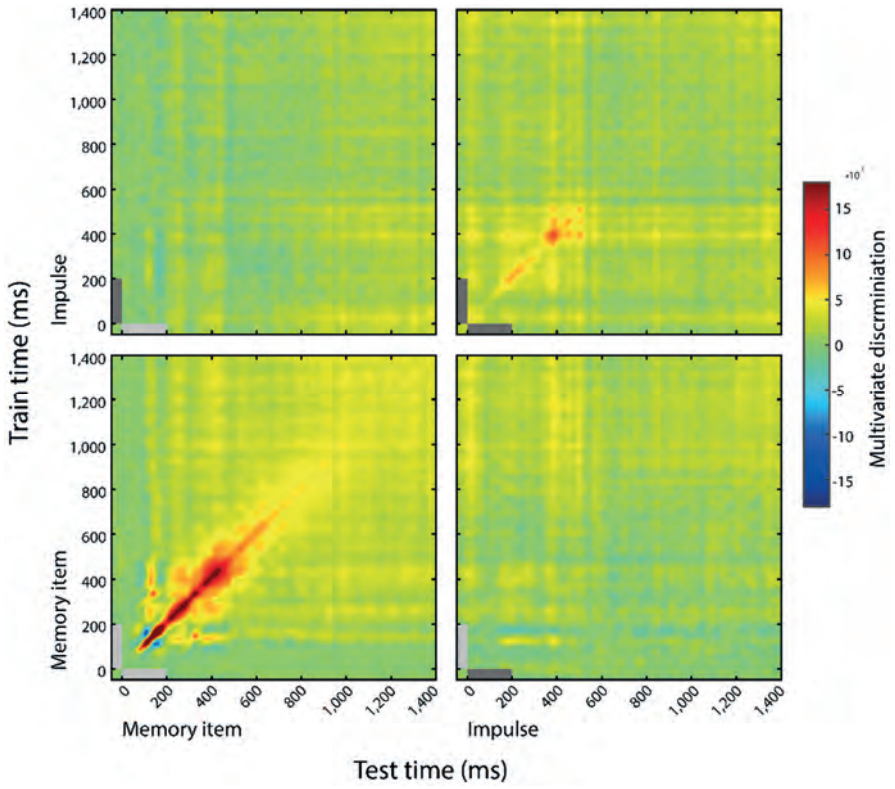


Figure 3.6. Dynamics of memory item discrimination. Mean discrimination matrices derived from training and testing on all time-point combinations. Light-gray and dark-gray bars represent memory item and impulse presentation, respectively.

Discrimination accuracy is time-locked to impulse onset

The increased discrimination accuracy shortly after the impulse could in principle be explained by a probe expectancy effect. Because the memory probe is presented on half the trials at this point, participants might prepare to respond to the probe. This could result in a more "active" maintenance of the memory item (e.g. K. Watanabe & Funahashi, 2007), which in turn could improve decoding accuracy. Although we do not find any evidence for a progressive ramp-up in discriminability at this time, this does not rule out a very precise form of temporal expectation.

To address this potential issue directly, we had introduced a very subtle temporal variability in the presentation of the impulse stimulus. Our reasoning was as follows: If discriminability is tightly time-locked to the variable onset of the impulse, rather than to the expected onset of the probe relative to the memory item, we can sensibly attribute the observed boost in discriminability to the presentation of the impulse stimulus.

We therefore plotted the cross-temporal matrices of the discrimination of the early and late impulse onset trials separately (Fig. 3.7A) time-locked to memory item onset, where the training data of both matrices was based on all impulse trials time-locked to impulse onset. As is apparent from the figure, the highest discrimination effect is not along the diagonal (where the test and train times correspond to the mean impulse onset and the actual impulse onset of all trials, respectively). Rather, for the early impulse trials, discrimination is highest when the training time is shifted by +30 ms, while a -30 ms shift is best for the late impulse trials. We then plotted and analysed the discriminations of the early and late impulse trials based on these shifted training times (Fig. 3.7B). Three positive significant clusters were found both in the early-onset condition (1,544 to 1,664 ms, $p = 0.003$; 1,704 to 1,776 ms, $p = 0.007$; 1,792 to 1,828 ms, $p = 0.028$) and in the late-onset condition (1,568 to 1,744 ms, $p < 0.001$; 1,784 to 1,836 ms, $p = 0.012$; 1,860 to 1,908 ms, $p = 0.016$). As is apparent from both the figure and the significant clusters, the time course of the late impulse onset trials is clearly later than the early onset trials.

To more directly test for the expected 60 ms latency shift in discrimination accuracy corresponding to the onset difference of the two impulse stimuli, we computed the Pearson's correlation between discrimination values of the time window from 1,370 to 2,170 ms of the early impulse onset condition with different time windows of the same length of the decoding values of the late impulse onset condition. Correlation coefficients were computed between the same time windows (0 ms difference) as well as for each 4 ms step up to a difference of 120 ms, resulting in 31 correlation values for each participant in total (Fig. 3.7C). The mean correlation clearly peaked at a 60 ms difference and a cluster-corrected permutation test on the Fisher transformed correlation values showed that only the correlation coefficients between a time-difference of 32 to 100 ms were significantly positive across subjects ($p < 0.001$). These results provide clear evidence that the decoding time-course was time-locked to the onset of the impulse.

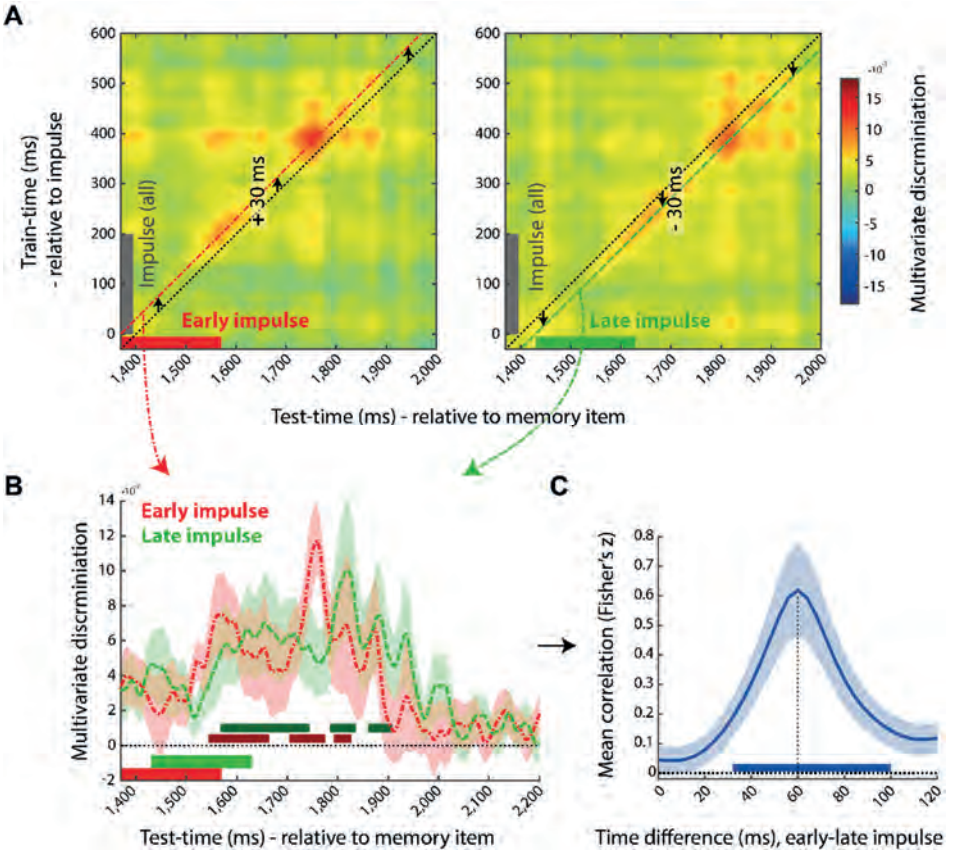


Figure 3.7. Effect of late impulse onset. **(A)** Mean discrimination matrices derived from training on all impulse trials, time-locked to impulse onset and testing separately on early (left, red) and late (right, green) impulse onset trials.. The black dotted lines illustrate the multivariate discrimination when tested on the average impulse onset relative to memory item (1,400 ms) but trained relative to the actual impulse onset (0 ms). Discrimination for early onset trials is highest when the training time is shifted by +30 ms (left, red line) and highest for late onset trials when shifted by -30 ms (right, green line). **(B)** A one dimensional plot of the early (red) and late (green) onset discriminations trained at +30 ms and -30 ms relative to impulse onset, respectively. Significant positive clusters of each onset condition are indicated by bars in a darker shade of the corresponding colors. Error bars are standard deviations of the permuted null distributions. **(C)** Mean correlations (Fisher's z) between the decoding time-course for the early and late impulse onset trials as a function of different temporal shifts. Mean correlation peaks at 60 ms. The blue bar illustrates the significant positive cluster of correlations. Error bars are standard deviations of the permuted null distributions.

Memory Item Discrimination is Not Simply Driven by Mean Amplitude Difference

The univariate analysis that was based on the averaged signal of all posterior electrodes showed significant memory item discrimination only shortly after memory item onset, where a single short significant cluster was present (140 to 168 ms, p = 0.022). No significant discrimination could be made within the impulse epoch (Fig. 3.8).

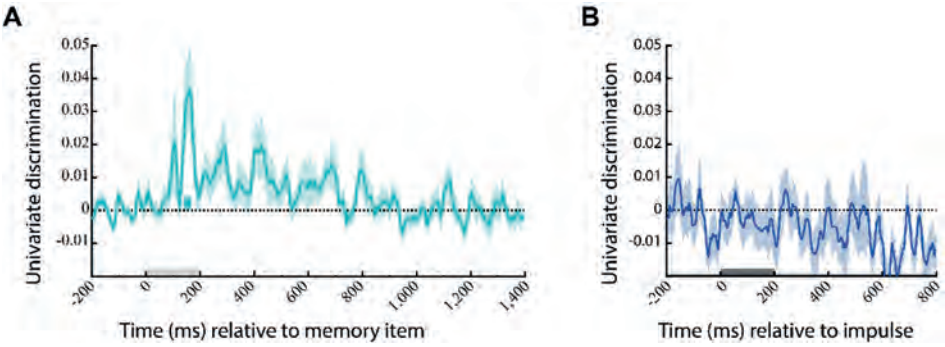


Figure 3.8. Univariate discrimination of the memory item. The cyan and blue lines show the univariate discrimination of the memory item of the **(A)** memory item and **(B)** impulse epoch, respectively. The cyan bar indicates the significantly positive discrimination cluster of the memory item epoch. Light-gray and dark-gray bars represent memory item and impulse presentation, respectively. Error bars are standard deviations of the permuted null-distributions.

Discussion

We report the results of a novel method to recover visual working memory states that are otherwise hidden to EEG using a functional perturbation approach. We presented a high-energy visual impulse stimulus during the vWM delay period and measured the visual evoked response. Critically, we found that the impulse response carried significant information about the contents in vWM. Using multivariate analysis, we could decode the orientation of the previous memory item from the impulse-driven visual response. This provides important proof-of-principle evidence for the feasibility of exploring hidden neural states with non-invasive EEG, with important implications for working memory (Stokes, 2015).

We used Mahalanobis distances to compute the multivariate dissimilarity between the evoked response during maintenance of specific orientations. The Mahalanobis distance is superior to Euclidean distance (Stokes et al., 2013) because it accounts for the covariance structure of the noise between features (Kriegeskorte et al., 2006). In the

current study, features were EEG sensors, which are known to be highly correlated. Analysis of the evoked response to the memory stimulus clearly validated this multivariate method as a powerful approach for decoding task-relevant parametric dimensions. Robust orientation discrimination was observed in the EEG activity as early as 68 ms after the presentation of the memory stimulus. Decoding peaked at around 160 ms, before decaying into the memory delay period. Despite returning almost to baseline prior to the onset of the impulse stimulus, we observed a robust 'reactivation' in decodability of the memory item that peaked at 200 ms and 360 ms after the impulse stimulus.

The impulse onset was temporally jittered by ±30 ms. The rationale for introducing this variability was to control for the possibility that reactivation could be explained by temporal expectation. On half the trials, the response probe was presented instead of the impulse stimulus. This was to ensure that participants were attending throughout the delay period. However, previous studies have shown that temporal expectation can also result in a ramp-up of item-specific delay activity (Barak et al., 2010; Takeda & Funahashi, 2004; Y. Watanabe, Takeda, & Funahashi, 2009). Ramp-up activity could reflect a build-up of temporal expectation (Nobre, Correa, & Coull, 2007), which could trigger attention-related pre-activation of the task-relevant template, as previously observed in monkey PFC (Rainer, Rao, & Miller, 1999) and the human visual system (Stokes, Thompson, Cusack, & Duncan, 2009). Jittering the impulse onset time allowed us to differentiate the relative contribution of temporal expectation and of the impulse response. This subtle temporal offset allowed us to test whether reactivation was indeed time-locked to the impulse stimulus, or whether decodability was better explained by the temporal structure of the task.

Visual inspection of the decodability time-course locked to the impulse probe already suggests that temporal expectation is not a plausible account. It would be surprising if template-reactivation could be so precise over an interval as long as 1.2 s. Moreover, plotting the impulse response for the different impulse onset times relative to the onset of the memory stimulus provides an estimate of the time-locking to the stimulus onset (Fig. 3.7B). As expected, the decodability profiles appear offset by approximately 60 ms. Finally, a correlation analysis of the decodability time-courses between impulse onsets confirmed that the correlation peaked at an offset of 60 ms. Overall, this pattern of results is consistent with the prediction that a neutral stimulus presented during the delay period drives activity in the memory network, resulting in a patterned response that systematically reflects the representational characteristics of the information in working memory (i.e., orientation).

Previous studies have argued that early visual cortex is important for vWM (Pasternak & Greenlee, 2005). For example, Harrison and Tong conducted an fMRI study using a very similar paradigm as the current design (Harrison & Tong, 2009). Using multivariate analyses, they found significant decoding during the delay period despite an absence of above-baseline activity levels. This suggests that subtle activity patterns in fMRI could also reflect hidden states (patterned spontaneous activity). Computational

modelling provides evidence that spontaneous spiking activity should be patterned by the hidden state (Sugase-Miyamoto et al., 2008). Moreover, we previously found evidence for significant pattern separation in monkey PFC, despite activity levels that were no greater than the pre-trial baseline (Stokes et al., 2013). Increasing the overall level of activity increased the pattern separation in that study. Future research could explore the relationship between spontaneous activity patterns measured with fMRI, single unit recording, and EEG.

It is also possible that the activity observed by Harrison and Tong (2009) actually reflected attentional preparation (Stokes et al., 2009) or imagery-related activity (Albers, Kok, Toni, Dijkerman, & de Lange, 2013; Stokes, Saraiva, Rohenkohl, & Nobre, 2011). Indeed, it is almost impossible to separate potential non-working memory contributions in their design (Stokes, 2011). In the current study, we clearly dissociate impulse-driven decoding from temporal expectation. Moreover, visual imagery is unlikely to be triggered so rapidly by the impulse stimulus. It would be important for future research to explore the relationship between discriminating stimulus-driven and non-driven activity as a function of attention and imagery to further pinpoint the relative contribution of different neural states to these separable, but interrelated cognitive functions.

We also observed evidence for dynamic coding of the memory stimulus. Cross-temporal analyses clearly revealed superior discrimination along the diagonal axis, reflecting within-time generalisation, relative to off-diagonal coordinates representing cross-temporal generalisation. This is the hallmark pattern for dynamic coding, indicating that the discriminative patterns vary over time (King & Dehaene 2014). Previously, Cichy and colleagues observed a similar pattern in MEG data during perceptual categorisation (Cichy et al., 2011), consistent with similar results from intracranial recordings in monkey visual (IT; Meyers et al., 2008), parietal (Crowe, Averbeck, & Chafee, 2010) and prefrontal cortices (Meyers et al., 2008; Stokes et al., 2013). There was also some evidence for a dynamic coding pattern in the impulse response, suggesting that the impulse response might be best conceptualised as a memory-specific trajectory, although future research would need to clarify this interpretation.

Interestingly, we found no evidence for cross-generalisation between the neural patterns evoked by the memory stimulus and the impulse response. Again, this could be interpreted as an extension of dynamic coding. The same task parameters are represented in both epochs (i.e., memory orientation), but using independent coding schemes. Epoch-independent coding schemes could be optimal for structured high-level representations (Sigala, Kusunoki, Nimmo-Smith, Gaffan, & Duncan, 2008). However, this result could also reflect a fundamental difference in patterns of activity that modulate hidden states, and the patterns of activity that are emitted from a particular impulse stimulus. Indeed, the current results are consistent with the hypothesis that the impulse response should be an interaction between the input pattern and the current hidden state, rather than a simple 'reactivation'. Readout of the hidden state from the EEG response only requires a systematic relationship between the impulse response and the hidden

state. By contrast, downstream cortical areas that read out the hidden state to generate a response might need to learn how to decode a time- and context-varying hidden state to access a memorized orientation. Recent theoretical models have shown that unsupervised read-out of dynamically changing states is in principle possible (Sussillo, 2014; Sussillo & Abbott, 2009).

Although this proof-of-principle experiment does not provide the definitive test for 'activity-silent' working memory, the results are nonetheless consistent with a number of key predictions. First, memory-discriminative information effectively returns to baseline after initial encoding. Although this is essentially a null effect, the decay function is consistent with studies decoupling persistent content-specific delay activity and memory-guided behaviour (Sreenivasan et al., 2014). Secondly, impulse-driven reactivation is consistent with a context-dependent response of a memory-configured hidden state (Mongillo et al., 2008; Sugase-Miyamoto et al., 2008). Finally, the dynamic trajectory during memory encoding is also consistent with a more general dynamic coding framework for working memory (Stokes, 2015).

Irrespective of any particular theoretical framework, the current experiment also provides an important demonstration of combining a functional perturbation approach with multivariate decoding to reveal otherwise hidden neural states. Activity states that we usually measure with non-invasive recordings only provide an incomplete picture of the diversity of neural states underlying cognition. This might be especially true for more tonic cognitive states, such as working memory, attention, or task set. Activity-silent representations pose an obvious problem for contemporary neuroscience, which is dominated by measurement and analysis of activity states. The ultimate success of future research will depend on new approaches to existing measurement techniques to probe diverse neural states, including 'activity-silent' states. We believe that this paper provides an important proof-of-principle toward an accessible non-invasive approach. Non-invasive brain stimulation could be used in combination with EEG to probe hidden states (Bortoletto, Veniero, Thut, & Miniussi, 2015).The advantage of transcranial magnetic stimulation is that the response profile of distinct brain networks can be targeted specifically (Rosanova et al., 2009), but with the major disadvantage that the stimulation artefact effectively precludes analysis of the initial local response to the perturbation. While this is less problematic for measuring context-dependent changes in effective connectivity between distant brain areas (Taylor, Nobre, & Rushworth, 2007), this limitation could easily obscure the kind of effect studied here.

In conclusion, we provide useful proof-of-principle demonstration of the utility of combining a functional perturbation approach with EEG to reveal otherwise silent neural states. Although these results are consistent with a dynamic coding framework that suggests visual working memory could be encoded in an 'activity-silent' state, the main purpose of the experiment was to develop a powerful tool for exploring cognitive states that cannot otherwise be differentiated with EEG. Future experiments will be able to exploit this novel approach in more complex experimental designs to tease apart the key coding principles underlying visual working memory.

Acknowledgments

This study was funded by the Wellcome Trust (to NEM), Medical Research Council (to MGS), and the National Institute for Health Research Oxford Biomedical Research Centre Programme based at the Oxford University Hospitals Trust, Oxford University. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. We would like to thank Janina Jochim for assistance collecting EEG data.

Chapter 4

Dynamic hidden states underlying working memory guided behaviour

This chapter was previously published as:

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864.

Data and code available at osf.io/mt4w8

Abstract

Recent theoretical models propose that working memory is mediated by rapid transitions in 'activity-silent' neural states (for example, short-term synaptic plasticity). According to the dynamic coding framework, such hidden state transitions flexibly configure memory networks for memory-guided behaviour and dissolve them equally fast to allow forgetting. We developed a perturbation approach to measure mnemonic hidden states in an electroencephalogram. By 'pinging' the brain during maintenance, we show that memory-item-specific information is decodable from the impulse response, even in the absence of attention and lingering delay activity. Moreover, hidden memories are remarkably flexible: an instruction cue that directs people to forget one item is sufficient to wipe the corresponding trace from the hidden state. In contrast, temporarily unattended items remain robustly coded in the hidden state, decoupling attentional focus from cue-directed forgetting. Finally, the strength of hidden-state coding predicts the accuracy of working-memory-guided behaviour, including memory precision.

Introduction

Working memory (WM) is a core cognitive function critical for flexible, intelligent behaviour(Alan Baddeley, 2003). Until recently, it was widely assumed that information is maintained in WM by maintaining specific activity states that represent the specific memoranda (Curtis & D'Esposito, 2003; Goldman-Rakic, 1995). However, accumulating evidence increasingly shows that successful maintenance in WM is not strictly dependent on an unbroken chain of corresponding delay activity (Stokes, 2015), and that item-specific activity states could reflect other cognitive processes. For example, in monkey studies persistent activity ramps up with expectation of the probe (Barak et al., 2010; Miller, Erickson, & Desimone, 1996; Watanabe & Funahashi, 2007, 2014). Similarly, in the human it has been shown that unattended WM content is not reflected in the neural signal, even when it is still clearly maintained (LaRocque et al., 2012; Lewis-Peacock et al., 2011; Sprague et al., 2016). Evidence for WM in the absence of persistent delay activity suggests that WM can be maintained in 'activity silent' neural states (Stokes, 2015).

Recent theories acknowledge that brain activity is highly dynamic, even when the contents of working memory remain stable (Sreenivasan et al., 2014). Multiple neurophysiological mechanisms could underlie such dynamics (Barak & Tsodyks, 2014; Buonomano & Maass, 2009; J. D. Murray et al., 2017). According to a dynamic coding model of WM (Stokes, 2015), behaviourally relevant sensory input drives a memory item-specific neural response, which triggers an item-specific change in the functional state of the system. Depending on the precise neural mechanism, this functional state could be activity-silent (e.g., short-term synaptic plasticity (Barak & Tsodyks, 2014; Fujisawa et al., 2008; Hempel et al., 2000; Lundqvist et al., 2016; Mongillo et al., 2008)), and maintained throughout the memory delay to serve as the neural context for subsequent processing. Items in WM would be read-out via the context-dependent response to a probe stimulus during recall (Buonomano & Maass, 2009; Sugase-Miyamoto et al., 2008). Crucially, this model predicts that dynamic hidden states are constructed when new information is encoded, and dissolved as soon as it is forgotten. This model also predicts that dynamic hidden states should determine the quality of a representation maintained in WM.

To probe hidden neural states, we developed a functional perturbation approach to 'ping the brain'. Analogous to the idea of active sonar (or echolocation), the response to a well-characterised impulse stimulus can be used to infer the current state of the system (Buonomano & Maass, 2009; Stokes, 2015). We recently validated this general approach using non-invasive electroencephalography (EEG) in a proof of principle study (Wolff et al., 2015/Chapter 3). The presentation of a high contrast, neutral visual stimulus evoked neural activity that clearly discriminated the previously presented visual stimulus. Here, we exploit this approach to track the functional dynamics of hidden states for WM.

Across two experiments, we show that the content of WM can be decoded from the impulse response during the maintenance interval, while forgotten information

leaves effectively no trace. In Experiment 2, we also demonstrate robust hidden-state representation for unattended content in WM, providing a plausible mechanism for maintenance that is independent of the activity associated with the focus of attention. Finally, we also find evidence that the quality of working memory varies with the decodability of these hidden states.

Methods

Participants

Thirty healthy adults (13 female, mean age 24.9 years, range 18-38 years) were included in the analyses of Experiment 1, 19 (10 female, mean age 24.7 years, range 18-39 years) in Experiment 2, and 20 in Experiment 3 (13 female, mean age 21, range 18-29 years). During data collection and preprocessing, 4 additional participants of Experiment 1, 1 additional participant of Experiment 2, and 6 additional participants of Experiment 3 were excluded from all analyses due to either low average performance on the memory task (below 60% accuracy) or excessive eye-movements (more than 30% of trials contaminated). No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications (Wolff et al., 2015/Chapter 3; Zhang & Luck, 2008). All participants of Experiment 1 and 2 received monetary compensation of £10/h, and participation in Experiment 3 contributed to course credits. All participants gave written informed consent. Experiments 1 and 2 were approved by the Central University Research Ethics Committee of the University of Oxford and Experiment 3 was approved by the Departmental Ethical Committee of the University of Groningen.

Apparatus and Stimuli

The experimental stimuli were generated and controlled by Psychtoolbox (Brainard, 1997), a freely available MATLAB extension. The stimuli were presented on a 23" screen running at 100 Hz and a resolution of 1920 by 1080 in Experiment 1, on a 22" screen at a resolution of 1680 by 1050 in Experiment 2, and on a 19" CRT screen running at 100 Hz and a resolution of 1280 by 1024 in Experiment 3. Viewing distance was set at 64 cm in Experiment 1, 67.5 cm in Experiment 2 and approximately 60 cm (not controlled) in Experiment 3, to ensure that the visual angles of stimuli were the same across experiments even though the screen parameters were different. A standard keyboard was used for response input by the participants.

All reported stimuli were the same in all experiments, unless explicitly mentioned otherwise. A grey background (RGB = 128, 128, 128; 20.5 cd/m^2; 28.6 cd/m^2 in Experiment 3) was maintained throughout the experiments. A black fixation dot with a white outline (0.242°) was presented in the centre of the screen throughout all trials. Memory items and memory were sine-wave gratings presented at 20% contrast, with a diameter of 6.69° and spatial frequency of 0.65 cycles per degree. The phase was

randomized within and across trials. The memory items were presented at 6.69° eccentricity and for each trial the orientations were randomly selected without replacement from a uniform distribution of orientations. The impulse stimulus was 3 adjacent 'bullseyes' in Experiment 1. Each 'bullseye' was of the same size and spatial frequency as the memory items. To reduce strain on the eyes, and to minimise forward masking in Experiment 3, the impulse stimulus in Experiments 2 and 3 consisted of 3 adjacent white circles. In Experiment 1 and 2 the probes had the same contrast and spatial frequency as the memory items, and was presented in the centre of the screen. In Experiment 3 the probe screen included a high contrast black and white square-wave grating in the centre and two white lateralized circles on the outside (the same location and size as the preceding lateral impulse circles). The angle differences between a memory item and the corresponding memory probe were uniformly distributed across 7 angle differences in Experiment 1 (±3°, ±7°, ±12°, ±18°, ±25°, ±33°, ±42°), 6 angle differences in Experiment 2 (±5°, ±10°, ±16° ±24°, ±26°, ±32°, ±40°) and a single angle difference (±16°) in Experiment 3.

Procedure

Experiment 1

Participants completed a retro-cue visual working memory task. Each trial began with the onset of a fixation dot at the centre of the screen. After 1000 ms, the memory item array was shown for 250 ms, consisting of two randomly oriented low-contrast gratings left and right of fixation. After a delay of 800 ms an arrow was shown for 200 ms in the centre of the screen, pointing either to the left or to the right, and thus cueing which of the two previously presented items would be tested. The number of left and right cued trials was equal and the order was randomized for each participant. The impulse stimulus was presented for 100 ms, 900 ms after the offset of the retro-cue. After another delay of 400 ms, the memory probe was shown for 250 ms. Participants were instructed to indicate if the orientation of the probe relative to the orientation of the memory item was rotated clockwise by pressing the "m" key with the right index finger, or counter-clockwise by pressing the "c" key with the left index finger. A high or low frequency feedback tone was played after response, indicating if the answer was correct or incorrect, respectively. The next trial started within 400 to 700 ms (determined randomly). Participants completed 1344 trials in total, which lasted approximately 3 hours (including breaks). Trial conditions were randomized across the whole session. See Figure 4.1A for a trial schematic.

Experiment 2

Participants completed a visual working memory task where two items were serially tested. The experiment began by instructing the participant which of the two memory items would be tested early, and which one would be tested late. This rule never changed within a session. Each trial began with the onset of a fixation dot at the centre of the screen. After 1000 ms, the memory item array was shown for 250 ms, consisting of two randomly oriented low-contrast gratings left and right from fixation. After a delay of 950

ms, the first impulse was presented for 100 ms. After a delay of 500 ms, the first memory probe was presented for 250 ms, probing the first item. The response input was the same as in Experiment 1. After a fixed delay of 1750 ms after the offset of the first probe, the second impulse was shown for 100 ms. Following a delay of 400 ms, the second memory probe was presented for 250 ms, probing the late-tested item. After the second response, two feedback tones were played, one for each response, separately indicating whether the first and second answers were correct. Participants completed two sessions of the task on two separate days, separated by approximately 1-2 weeks. The testing order of the memory items was fixed within each sessions, and switched between sessions (i.e. left item tested first in one session, right item tested first in the other session). The order of the testing rule between sessions, (whether the left item would be tested first in the first or in the second session) was counterbalanced across participants (odd numbered left first, even numbered right first). Each session consisted of 864 trials, and lasted approximately 3 hours including breaks. See Figure 4.4A for a trial schematic.

Experiment 3

The task was almost the same as Experiment 1, including the same timings of the memory items, cue, probe and overall trial duration. The one key difference was the timing of the impulse stimulus. While the delay between cue offset and probe onset was held constant at 1,400 ms across all trials (the same as in Experiment 1), the SOA between impulse and probe onset was 0, 50, 100, 250 or 500 ms (determined pseudo randomly across the session). No impulse was shown in the 0 ms SOA condition. The impulse remained on the screen until the probe stimulus was presented. This was to ensure the least possible interference of the impulse on probe processing (i.e., rapid onset and offset of the white circles immediately before probe presentation could deteriorate probe visibility), as well as keeping the different SOA conditions as similar as possible (longer SOA would include an additional offset). Participants completed 280 trials (approximately 30 minutes). See Fig. 4.8A for a trial schematic.

Data collection and analyses were not performed blind to the conditions of the experiments.

Due to the within-subject design in all three experiments, randomization of conditions between subjects was not applicable.

EEG Acquisition

The EEG signal was acquired from 61 Ag/AgCl sintered electrodes (EasyCap, Herrsching, Germany) laid out according to the extended international 10-20 system. Data was recorded at 1000 Hz using NeuroScan SynAmps RT amplifier and Scan 4.5 software in Experiment 1 and Curry 7 software in Experiment 2 (Compumedics NeuroScan, Charlotte, NC). The anterior midline frontal electrodes (AFz) served as the ground. Bipolar electrooculography (EOG) was recorded from electrodes placed above and below the right eye, and from electrodes placed to the left of the left eye and to the

right of the right eye. The impedances of all electrodes were kept below 5 kΩ. Online, the EEG was referenced to the right mastoid and filtered using a 200 Hz low-pass filter.

EEG pre-processing

Offline, the data was re-referenced to the average of both mastoids, down-sampled to 500 Hz and band-pass filtered (0.1 Hz high-pass and 40 Hz low-pass) using EEGLAB (Delorme & Makeig, 2004). The data was then epoched to the onset of the memory items and the impulses. In Experiment 1, the memory item epoch was from -200 ms to 1050 ms, relative to onset, and in Experiment 2 from -200 ms to 1200 ms. The impulse epochs were from -200 ms to 500 ms relative to onset in both experiments. Additionally, for the purpose of artefact rejection, which included the rejection of trials containing saccadic eye-movement prior to the time of interest (see below), the cue segment in Experiment 1 was also epoched (-200 ms to 1100 ms).

Subsequent artefact detection and trial rejection focused exclusively on the 17 posterior channels that were included in the analyses (P7, P5, P3, P1, Pz, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, O2) and the EOGs. Each trial of each epoch was individually visually inspected for blinks, saccades and non-stereotyped artefacts. Trials from individual epochs were rejected from analyses involving that epoch if it contained any of the above-mentioned artefacts. Furthermore, impulse-epoch trials were also excluded from corresponding analyses if the EOG signal suggested that saccades occurred during any of the previous epochs of that trial. In Experiment 1 this exclusion procedure was applied to the cue-epoch as well. In Experiment 2, late impulse trials were also excluded if no response was registered for the preceding probe. For the decoding analyses, each epoch was baselined using the average signal from -200 ms to 0 ms before stimulus onset. The multivariate data were also demeaned at each time-point by subtracting the average voltage for all posterior channels included in the analyses.

Time-frequency decomposition and lateralization analysis

In order to explore alpha power (8-12 Hz) lateralization (Schneider, Mertes, & Wascher, 2016; Worden, Foxe, Wang, & Simpson, 2000), the spectral power from 6 to 16 Hz (in steps of 0.5 Hz) of the EEG signal was computed using Hanning tapers with time-windows of 5 cycles per frequency (in steps of 10 ms) using the MATLAB toolbox FieldTrip (Oostenveld et al., 2010). We included the whole experimental trial, ranging from 1000 ms before memory item onset until 1500 ms after (second) probe onset (-1000 to 4150 ms relative to memory items in Experiment 1, and -1000 to 5800 ms relative to memory items in Experiment 2). The power was log transformed, and lateralization was computed by subtracting the average power of the ipsilateral posterior electrodes from the average power of the contralateral posterior electrodes in relation to the cued memory item in Experiment 1 and to the early-tested item in Experiment 2 (P7, P5, P3, P1, PO7, PO3, O1 versus P8, P5, P6, P4, P2, PO8, PO4, O2).

Significant clusters of lateralization were determined using a cluster-corrected non-parametric sign-permutation test (Maris & Oostenveld, 2007). In both experiments, the

whole trial was included in this analysis (-100 to 3150 ms relative to memory items onset in Experiment 1, and -100 to 4800 ms in Experiment 2).

Orientation decoding

To test whether the activity pattern of the posterior EEG channels of interest contained orientation-specific activity, we used the Mahalanobis distance (De Maesschalck et al., 2000) to compute the trial-wise distances between the full range of possible orientations, and quantify to what extent the computed distances adhere to the parametric circular space of the orientations (Sprague et al., 2016). This approach is an extension of the pairwise distance approach we used before (Wolff et al., 2015/Chapter 3) and is conceptually similar to the population tuning curve model (Saproo & Serences, 2010).

The left and right presented items were decoded separately and independently within each participant and experimental session. All 17 posterior channels (see above) were used for all decoding analyses. The procedure followed a leave-one-trial-out cross-validation approach to compute the trial-wise decodability of the orientation of interest. The activity pattern of a single test-trial at a particular time-point was compared to the pattern of all other trials at the same time-point. These were averaged into 12 orientation bins relative to the orientation of the test-trial, each containing trials with orientations within a range of 30° and centred around -75°, -60°, -45°, -30°, -15°, 0°, 15°, 30°, 45°, 60°, 75°, and 90°. The Mahalanobis distances between the test-trial and each orientation bin was computed using the covariance estimated from all trials excluding the test-trial using a shrinkage estimator (Ledoit & Wolf, 2004). To simplify visualization and interpretation, the 12 resulting distances were mean centred and the sign was reversed, resulting in a visual representation of a tuning curve. Higher values correspond to greater relative similarity between the test-trial and the averaged train-trials within a particular orientation bin, and lower values correspond to greater dissimilarity.

Next, the vector means of the tuning curves were computed (Sprague et al., 2016). First, the cosine of the centre of each orientation bin (θ) was rescaled to the range -180 to 180. It was then multiplied with the corresponding sign-reversed distances (d(θ)) before the mean of the resulting 12 values was taken, which made up the decoding accuracy (da).

Equation 1: $da = mean(d\ (\theta) \cos\ (2\ \theta))$

A high value reflects evidence for orientation tuning: the difference between the test-trial and train-trials with a similar orientation is smaller than between the test-trial and train-trials with different orientations. This procedure was repeated for all trials and all time-points. See Supplementary Information for the custom Matlab function used to decode orientations using Mahalanobis distance.

The decoding values were averaged over all trials, and smoothed over time with a Gaussian smoothing kernel (SD = 16 ms) for visualization and time-resolved significance testing.

Cluster-corrected sign-permutation significance tests were carried out within the memory items epoch (0 to 1050 ms in Experiment 1, 0 to 1200 ms in Experiment 2) and impulse epochs separately (0 to 500 ms in both experiments), in order to explore the significant decoding time-course. Additionally, to assess the overall decodability within an epoch, the decoding values were averaged over time (from 100 ms after stimulus onset until the end of the epoch) and then submitted to a two-sided permutation test.

Relationship between behaviour and decoding

The trial-wise average decoding scores after memory items presentation (100 to 1050 ms in Experiment 1, and 100 to 1200 ms in Experiment 2) and impulse presentation (100 ms to 500 ms) was median split. Non-response trials (to the early probe in Experiment 2) were excluded from this analysis. The average behavioural accuracies of high and low decoding trials were statistically compared using a two-sided permutation test.

Behavioural modelling

To further explore the relationship between WM task performance and trial-wise decoding, we modelled the behavioural performance as a function the difference in degrees between the orientation of the memory item and the probe using the following model that was fit to each participant separately (Murray, Nobre, & Stokes, 2011).

Equation 2: $y = \lambda + \frac{(1-2\lambda)}{2} \times erfc(\frac{-\beta}{\sqrt{2}}\ (x - \alpha))$

where *erfc* is the complementary Gaussian error function, λ is the asymptote, β is the slope and α is the threshold/bias parameter. The modelling fitting was performed using the Palamedes Matlab toolbox (Prins & Kingdom, 2009). The asymptote represents the guess rate, where a higher value reflects a higher probability that no information about the probed item is maintained in WM, resulting in a higher probability for mistakes even when the angular difference between the probe and the memory item is large. The slope is interpreted as the memory precision, where a high precision reflects a relatively high proportion of correct responses at small degree rotations between the probe and memory item. The asymptote and slope parameters were both unconstrained across the high and low decoding conditions. A single bias parameter was used, which was included (instead of fixing it at 0) because cumulative-likelihood tests (Claessens & Wagemans, 2008) showed better model fits for all cases (Experiment 1: $n = 30$, $\chi^2(30) = 135.978$, $p < 0.001$; Experiment 2, $n = 19$, early accuracy: $\chi^2(19) = 215.351$, $p < 0.001$; late accuracy: $\chi^2(19) = 33.69$, $p = 0.02$).

The unconstrained model parameters (slope and asymptote) were subsequently compared between high and low decoding trials. Since the behavioural modelling was carried out as a direct follow up to the average accuracy effects observed in both experiments (two-sided tests), we had clear expectations about the directionality of the effects. For the positive relationship between decoding and accuracy observed for the cued item in Experiment 1 and both tests in Experiment 2, we expected that decoding should have a negative relationship with guess-rate (i.e., lower guess-rate for high

decoding) and/or a positive relationship with precision (higher precision for higher decoding), and vice versa for the negative accuracy effect of the uncued item in Experiment 1. Therefore, all tests of model parameter comparisons between high and low decoding trials were one-sided.

Cross-temporal decoding

We also explored the cross-temporal dynamics of stimulus processing and maintenance as a function of item priority in Experiment 2, and cross-generalization between impulse and memory presentation epochs in both experiments. The decoding approach was the same as described above, except classifiers trained at each time point were tested at every other time point, resulting in 2-dimensional cross-temporal decoding matrices (King & Dehaene, 2014).

If the decoding patterns are stationary, it should not matter whether train/test is performed using the same time points. In contrast, decoding often appears dynamic: training and testing on the same time-points results in higher decoding scores than training and testing on different time-points (i.e., minimal cross-temporal generalization). We tested for this hallmark feature of dynamic coding using a non-parametric test used previously (Myers et al., 2015). The decodability at each cross-temporal time-point $t_{x,y}$ was compared to the pair of decodabilities at the corresponding within time-points ($t_{x,x}$ and $t_{y,y}$) with two separate permutation tests. A significant difference in both was taken as evidence for dynamic coding. Time-points of significant dynamic coding were corrected for multiple comparisons using a two-dimensional cluster-based permutation test.

Significance testing

To determine statistical significance, we used the non-parametric sign-permutation test (Maris & Oostenveld, 2007) (with one exception, see ANOVA below), which does not make assumptions about the underlying distribution. Since the null hypotheses of all tests corresponded to no effect (i.e. no difference in power lateralization, no difference in decodability, etc.), the sign of the data of each participant was randomly flipped with a probability of 50% 50,000 times. The resulting distribution was used to derive at the *p*-value of the null-hypothesis that the mean effect was equal to 0. All tests were two-sided, unless otherwise stated.

For time-series and frequency data, the above procedure was repeated for each time-point and frequency (when applicable). To correct for multiple comparisons over time and/or frequencies, a cluster-based permutation test was subsequently used using 50,000 permutations (5,000 for cross-temporal decoding, due to computer memory limitations), with a cluster-forming threshold and cluster significance threshold of $p < 0.05$. Tests concerning the average of specific time-windows (which includes decoding-behaviour relationships) were performed to test unique and independent hypotheses, therefore no correction applied. The sample size for all tests in Experiment 1 was $n = 30$, $n = 19$ in Experiment 2, and $n = 20$ in Experiment 3.

The 95 % confidence intervals of the error-bars were determined by bootstrapping from the corresponding data 50,000 times.

The boxplots used in our figures follow the standard conventions. The middle line represents the median, the box the first and third quartile, and the whiskers all data within 1.5 * interquartile range of the lower and upper quartile. Where appropriate, data points outside this range are displayed individually (small crosses).

A repeated measures ANOVA was used to analyse the behavioural data of Experiment 3. The normality and equal variances assumptions were tested with the Shapiro-Wilk test of normality and Mauchly's test of sphericity, respectively. Neither test provided evidence for assumption violations of the data.

Data availability

The data that support the finding of this study are publically available at http://datasharedrive.blogspot.co.uk/2017/03/dynamic-hidden-states-underlying.html. All necessary task/condition information has been provided within a self-contained format, as specified in the OECD Principles and Guidelines for Access to Research Data from Public Funding (Pilat & Fukasaku, 2007).

Code availability

All complete custom Matlab routines used to generate the figures of this paper are available at http://datasharedrive.blogspot.co.uk/2017/03/dynamic-hidden-states-underlying.html

Results

Experiment 1

In Experiment 1, 30 human participants performed a visual WM task while EEG was recorded. At the beginning of each trial (see Fig. 4.1A), two memory items were presented, but a retrospective cue (retro-cue) presented during the delay instructed participants which item would actually be probed (Griffin & Nobre, 2003; Landman et al., 2003). The other item could be simply forgotten. The retro-cue in this design is essential to differentiate WM from basic stimulation history (Harrison & Tong, 2009). During a subsequent memory delay, we then presented a high contrast "impulse" stimulus. Memory performance for the cued item was tested after the impulse by a centrally presented memory probe (Fig. 4.1B). Time-frequency decomposition of lateralised activity in posterior sensors (Fig. 4.1C) shows significant lateralization in the alpha range (8-12 Hz) after the presentation of the cue (permutation test, $n = 30$, $p < 0.001$, corrected, cluster-forming threshold $p < 0.05$). This pattern is consistent with a

shift in spatial attention (Worden et al., 2000) according to the retro-cue, which confirms that the cue manipulation was effective.

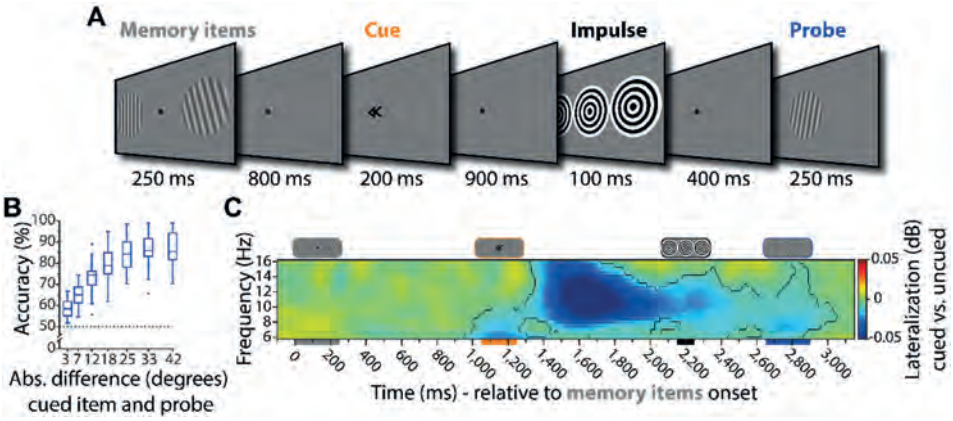


Figure 4.1. Experiment 1 task structure, behavioural performance and attention-related alpha band activity. **(A)** Trial schematic. Two memory items were presented (randomly oriented grating stimuli), and participants were instructed to memorize both orientations. A retro-cue then indicated which item would actually be tested at the end of the current trial (100% valid). The impulse stimulus (high contrast, task-irrelevant visual input) was then presented during the subsequent delay while participants should have only the cued item in WM. At the end of the trial, a forced-choice probe was presented at the centre of the screen. Participants indicated whether the probe was rotated clockwise or anti-clockwise relative to the orientation of the cued item. **(B)** Boxplots show WM accuracy as a function of the absolute angular difference (in degrees) between the memory item and the probe. Data points outside of the 1.5 * interquartile range are shown separately (small crosses). **(C)** Time-frequency representation of the difference between the contra- and ipsilateral posterior electrodes relative to the cued hemifield. The highlighted cluster in the alpha frequency band (8-12 Hz) indicates significant contralateral desynchronization (permutation test, $n = 30$, cluster-forming threshold $p < 0.05$, corrected significance level $p < 0.05$). The coloured bars under the x-axis represent the timings of the corresponding stimuli illustrated on top.

Decoding parametric memory items

To decode the memory items used in this experiment, we developed a parametric variant of distance-based discrimination (see Methods, Fig. 4.2A-D). As shown in Fig. 4.2A, this capitalises on the parametric structure of the stimulus space (Saproo & Serences, 2010),

whilst maintaining the statistical advantages of the Mahalanobis distance metric used in previous EEG/MEG decoding studies (Myers et al., 2015; Wolff et al., 2015/Chapter 3). To summarise briefly here: for a given trial, we compare the activity pattern across electrodes to the corresponding activity pattern observed in the remaining trials, averaged by orientation-difference to the test trial (at a bin width of 30 degrees). This procedure is repeated for all trials and all time-points. If the pattern of activity contains information about item orientation, we expect greater pattern dissimilarity (i.e., Mahalanobis distance) at larger angular differences. Fig. 4.2B shows distance as a function of reference angle and time after the presentation of the left and right item separately (upper/lower respectively). Distance values were then converted into a decoding accuracy score (Fig. 4.2C) and averaged across both items at each time-point (Fig. 4.2D). Item orientation could be decoded from 56 ms until 1026 ms after onset (permutation test, $n = 30$, $p < 0.001$ (corrected), cluster-forming threshold $p < 0.05$). This is consistent with previous empirical evidence that EEG is sufficiently sensitive to detect subtle differences in scalp-level activity patterns associated with different stimulus orientation (Wolff et al., 2015/Chapter 3). The current decoding results further validate the utility of multivariate pattern analysis for two simultaneously presented orientation gratings.

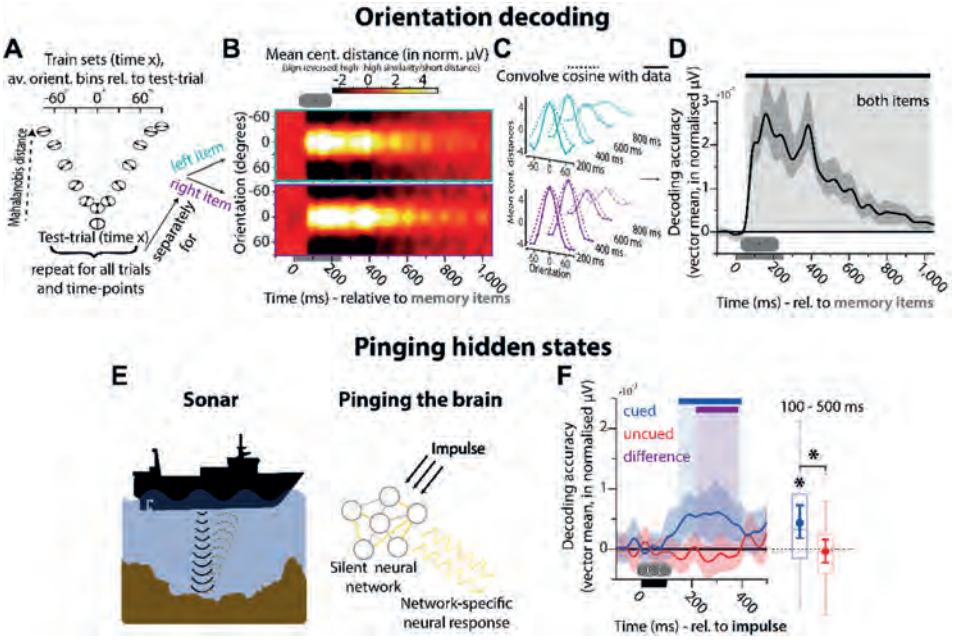


Figure 4.2. Orientation decoding in EEG and pinging hidden states of WM. **A-D.** Decoding procedure. **(A)** The dissimilarity in the neural pattern between a single trial and all other trials is computed as a function of orientation difference (binned: 30 degrees). **(B)** Average distance to template of all trials for each time-point during and after memory item presentation, plotted separately for the left and the right memory item (upper/lower respectively). Distances are mean centred and sign reversed (high = small distance/high similarity) for visualization. **(C)** A cosine is convolved with the data. **(D)** The vector mean of the convolved tuning curves (i.e., decoding accuracy) over time, averaged over left and right items. The black bar indicates significant decoding (permutation test, n = 30, cluster-forming threshold $p < 0.05$, corrected significance level $p < 0.05$). Error shading is the 95 % C.I. of the mean. **(E)** Pinging hidden states. Analogy to active sonar: differences in hidden state are inferred from differences in the measured response to a well-characterised impulse. **(F)** Decoding results in the impulse epoch. The blue bar indicates significant decoding of the cued item. The purple bar indicates significant difference in decodability between the cued and uncued item (permutation test, n = 30, cluster-forming threshold $p < 0.05$, corrected significance level $p < 0.05$). Error shading is the 95 % C.I. of the mean. The boxplots and superimposed circles with error-bars (mean and 95 % C.I. of the mean) represent average decoding from 100 to 500 ms after impulse onset. Data points outside of the 1.5 * interquartile range are shown separately (small crosses). Significant average decoding and significant difference in average decodability between the cued and uncued item are marked by asterisks (permutation test, n = 30, $p < 0.05$).

Pinging hidden states

According to the dynamic coding framework, we hypothesised that the input/output mapping of neural circuits maintaining information in WM should systematically reflect the memory content (Stokes, 2015). We tested this using an impulse stimulus to 'ping' potentially hidden neural states (Fig. 4.2E). As predicted, the impulse-specific response clearly differentiated the content of WM (Fig. 4.2F), even though the driving input ('ping') was held constant on each trial. The decodability of the cued item showed a significant cluster from 148 to 398 ms after impulse stimulus onset (permutation test, $n = 30$, $p = 0.002$, corrected, cluster-forming threshold $p < 0.05$). Average decodability from 100 to 500 ms was also significant ($p = 0.004$). Cued item decoding was also higher than task-irrelevant (uncued) item decoding (cluster: 216 to 386 ms, $p = 0.009$, corrected; average: $p = 0.028$). Indeed, the uncued item showed no evidence for decoding (no corrected clusters; average: $p = 0.687$), suggesting that content can be rapidly purged from WM when instructed, leaving effectively no trace in the neural state.

To test whether the impulse response reflects a literal 'reactivation' of item-specific activity observed during encoding, we also examined whether a classifier trained on the activity elicited by the memory stimuli during encoding could be used to decode the memory item during the impulse epoch (and vice versa). However, we found no evidence for significant cross-generalization between discriminative activity patterns during encoding and discriminative activity driven by the impulse (corrected clusters, $p > 0.347$). We propose that the impulse stimulus simply acts as a functional ping to recover hidden states, rather than a literal 'reactivation' of a latent representation (Wolff et al., 2015/Chapter 3).

Trial-wise variability in decoding the impulse response also predicted variability in WM performance. Higher decoding trials of the cued item were accompanied by higher performance than low decoding trials (permutation test, $n = 30$, $p = 0.043$; Fig. 4.3A, left). There was also a complementary cost for decoding the uncued item (i.e., a high decoding score for the uncued item led to a decrease in accuracy on the cued item; $p = 0.002$; Fig. 4.3B, left), suggesting that participants might have failed to discard the uncued item (or simply did not use the cue properly) on some trials, contributing to error in performance. Finally, the difference between the accuracy effect of the cued and the uncued item was also significant (permutation test, $n = 30$, $p < 0.001$).

In principle, the relationship between trial-wise decoding and WM performance may rest on an increase in guess-rate (i.e., due to forgetting or failure to encode), or a reduction in precision, or both (Bays & Husain, 2008; Zhang & Luck, 2008). To separate these possible contributions, we modelled the behavioural profile over degrees of angular rotation between the memory item and the probe stimulus (Methods; Murray, Nobre, & Stokes, 2011; Prins & Kingdom, 2009). We found that the link to behaviour is most likely driven by a decrease in precision (the slope parameter of the model) for weakly encoded hidden states of WM (permutation test, $n = 30$, $p = 0.023$, one-tailed; Fig. 4.3A, right), while no evidence for an effect in guess rate (the asymptote parameter) was found ($p = 0.867$, one-tailed). Modelling the observed uncued item accuracy effect

was inconclusive (Fig. 4.3B, right), with no evidence for either a precision or guess rate effect (p = 0.443 and p = 0.184 respectively, one-tailed). Finally, we found no evidence that trial-wise item decoding during the initial presentation of the memory stimuli relates to memory performance, further suggesting that the relationship between accuracy and decoding triggered by the impulse is not due to a failure to encode the memory item (p > 0.3).

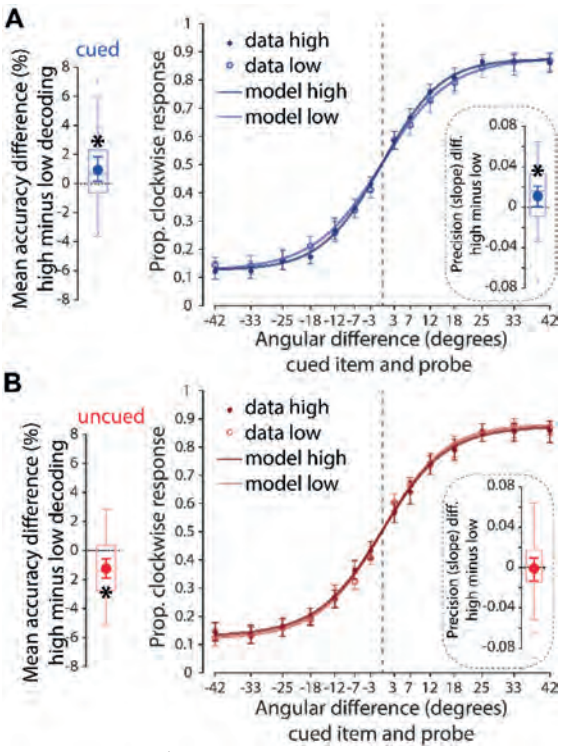


Figure 4.3. Relationship between item-specific impulse decoding and WM accuracy. **(A)** Difference in overall WM task performance between high and low cued item decoding trials (left). Proportion clockwise response for high and low decoding trials as a function of the angular difference between the memory item and the probe (right). Inset shows the difference in the slope parameter (a measure of memory precision) between high and low decoding trials. Data points outside of the 1.5 * interquartile range are shown separately in the boxplots (small crosses). Superimposed circles and error-bars are the mean and 95% C. I. of the mean. **(B)** The same convention as in a. but for the decoding of the uncued item. Significant differences in accuracy/precision between high and low decoding trials are highlighted by asterisks (permutation test, n = 30, p < 0.05).

Experiment 2

Recently, it has been proposed that information in WM can be represented in qualitatively different states (Larocque et al., 2014; Olivers et al., 2011; Souza & Oberauer, 2016), with attended items encoded in activity states measurable with standard recordings of delay activity, whereas activity-silent states could underlie the representation of currently unattended information in WM. In Experiment 2 (n = 19) we test whether unattended but nevertheless remembered information in WM can still be decoded from the impulse response. Again, two memory items were presented at the start of the trial, however both were ultimately relevant as they would both be probed.

Priority was manipulated by blocking the order in which items would be probed (Fig. 4.4A), and instructing participants accordingly. Because there was no other clue as to which item was being probed first or second, non-random responses already indicate that participants used this blocked information (Fig. 4.4B). This was further supported by lateralised changes in alpha power (Fig. 4.4C). During and shortly after the initial presentation of the memory stimuli, there was a relative decrease in power at sensors contralateral to the initially prioritised item, consistent with selective allocation of attention (permutation test, $n = 19$, $p = 0.023$, corrected, cluster-forming threshold $p < 0.05$). Moreover, this pattern reversed after the response to the first item ($p = 0.009$, corrected), consistent with the assumption that participants then shift the originally de-prioritised item into the focus of attention in WM in preparation for the second probe (Ede, Niklaus, & Nobre, 2016).

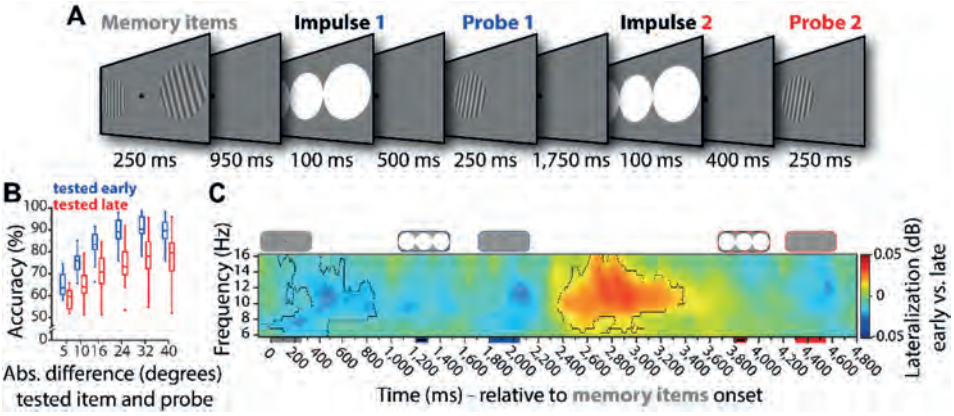


Figure 4.4. Experiment 2 task structure, behavioural performance and attention-related alpha band activity. **(A)** Trial schematic. Two memory items were presented. Participants were instructed to maintain both items and were told at the start of each block in which order the items would be tested. The first impulse was presented within the first memory delay (maintain both items, but attend the prioritised item), after which the prioritised item was probed. The second impulse was presented during the subsequent memory delay (maintain and attend only the now-prioritised item), after which the remaining item was probed. **(B)** Boxplots show the accuracy of the early and late tested item as a function of the absolute angular difference (in degrees) between the memory item and the probe. Data points outside of the 1.5 * interquartile range are shown separately in the boxplots (small crosses). **(C)** Time-frequency representation of the difference between the contra- and ipsilateral posterior electrodes relative to the presentation side of the early tested memory items. Highlighted areas indicate significant difference (permutation test, $n = 19$, cluster-forming threshold $p < 0.05$, corrected significance level $p < 0.05$).

Decoding during stimulus presentation

We first analysed decoding during the initial processing of the memory stimuli. The results are plotted separately as a function of test-time (early or late in the trial) as this could be meaningfully classified from the beginning of the trial (Fig. 4.5A). As expected, decoding the prioritised item (cluster: 74 to 1,200 ms, $p < 0.001$, corrected, cluster-forming threshold $p < 0.05$; average: $p < 0.001$), relative to the de-prioritised item (cluster: 82 to 542 ms, corrected, $p < 0.001$, corrected; average: $p < 0.001$) was more robust (average: $p = 0.013$). While decoding of the unattended item drops to chance relatively quickly after item presentation, the attended item shows significant decoding until the end of the epoch, replicating previous evidence showing that maintenance of only attended WM items is represented in the recorded brain activity patterns (LaRocque et al., 2012; Lewis-Peacock et al., 2011; Sprague et al., 2016).

The difference between attended and unattended item-maintenance in WM was even more apparent when comparing their cross-temporal decoding matrices. Minimal cross-temporal generalization during and shortly after memory item presentation suggested highly dynamic item encoding: orientation discriminative patterns change over time. This was supported by significant dynamic coding clusters during item encoding for both the early and late tested item, where off-diagonal time-points show significantly lower decodability than both corresponding on-diagonal time-points (permutation test, $n = 19$, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$; see Methods; Fig. 4.5B, left and middle). However, the attended item clearly showed a more time-invariant decoding pattern at the end of the epoch than the unattended item, apparent by both significantly higher decodability on same time-point as well as cross time-point decoding ($n = 19$, $p = 0.023$, corrected, cluster-forming threshold $p < 0.05$; Fig. 4.5B right). This further suggests that while the attended item also has a corresponding WM maintenance signature in stable activity patterns, the unattended item does not.

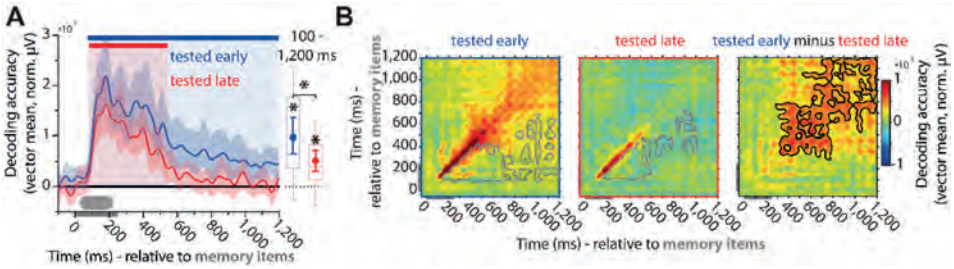


Figure 4.5. Priority-dependent encoding and maintenance in WM. **(A)** Decodability of the item that is tested early (blue) and the item that is tested late (red) during memory item presentation. Blue and red bars indicate significant decoding clusters for the early- and late-tested item, respectively (permutation test, n = 19, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$). Error shading is 95% C.I. of the mean. Boxplots and superimposed circles with error bars (mean and 95 % C.I. of the mean) represent average decodability from 100 ms after stimulus onset until the end of the epoch. Significant average decoding and average difference between the decodability of the early and late item are marked by an asterisk (permutation test, n = 19, $p < 0.05$). **(B)** Cross-temporal decoding matrices of the early (left) and late-tested (middle) item derived from training and testing on all time-point combinations, and the difference between the decoding of the early and late tested item (right). The grey outline indicates time-points of significantly lower decoding relative to both equivalent time-points along the diagonal, which is taken as evidence for dynamic coding (permutation test, n = 19, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$). The black outline (right) indicates significantly higher decodability of the early compared to the late tested item (permutation test, n = 19, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$).

Decoding of the impulse responses

Critically, we found that both the attended (clusters: 80 to 308 ms, $p = 0.004$, and 332 ms to 434 ms, $p = 0.031$, corrected; average: $p < 0.001$) and unattended items (cluster: 172 to 306 ms, $p = 0.011$, corrected; average: $p = 0.045$) were decodable in the first impulse response (Fig. 4.6A). This contrasts with the clear cueing differences observed in Experiment 1, and suggests multiple items can be encoded in hidden states and revealed by the impulse, even if only one item is in the focus of attention. It is worth noting, however, that the decodability of the attended item was significantly higher than that of the unattended item (average: $p = 0.031$), consistent with the behavioural evidence for relatively better memory for the initially prioritised item.

We found no evidence for a relationship between trial-wise differences in alpha lateralization and WM item decodability of the impulse response for either the attended or unattended item (Suppl. fig. 4.1). This further suggests that the item-specific impulse response does not even vary with trial-wise differences in the focus of attention.

We also found that the remaining relevant and initially unattended item could also be decoded in the second impulse response (cluster: 196 to 326 ms, $p = 0.016$, corrected; average: $p = 0.012$), while decoding the initially prioritised item failed to reach significance in this epoch (clusters: $p > 0.109$, corrected; average: $p = 0.112$; Fig 6B). The now-deprioritised item was presumably cleared from the hidden state because it was no longer relevant, similar to forgetting observed after the retro-cue from Experiment 1.

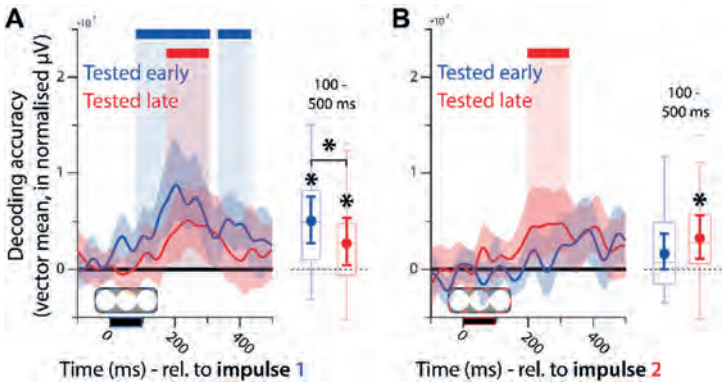


Figure 4.6. Attended and unattended WM items in early and late epochs. **(A)** Item decoding of the early (blue) and late tested item (red) during the first impulse epoch. Coloured bars on top indicate significant decoding clusters of the corresponding items (permutation test, n = 19, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$). Error shading is 95% C.I. of the mean. Boxplots and superimposed circles with error bars (mean and 95% C.I. of the mean) represent average decodability from 100 ms after stimulus onset until the end of the epoch. Significant average decoding and average difference between the decodability of the early and late item are marked by an asterisk (permutation test, n = 19, $p < 0.05$). **(B)** Item decoding during the second impulse epoch, same conventions as **A**.

Again, we also tested for cross-generalization between the decodable patterns of the memory items epoch (Fig. 4.5A) and the impulse-epochs (Fig. 4.6A, B). However, like in Experiment 1, we found no evidence that the impulse literally 'reactivates' activity patterns associated with initial encoding for either item (all corrected clusters: $p > 0.32$).

There was also a positive relationship between trial-wise decoding of the attended items at the first and at the second impulse with WM performance (early: $p = 0.038$, Fig. 4.7A; late: $p = 0.04$; Fig. 4.7B), replicating and extending the findings of Experiment 1. As in Experiment 1, we modelled the behavioural profile to test if the positive relationship between decoding and task performance is due to an increase in precision and/or a decrease in guess-rate. While the modelling results were inconclusive for the early-tested item (precision: $p = 0.399$, one-tailed; guess-rate: $p = 0.329$, one-tailed; Fig. 4.7A), there was evidence for an effect in precision of working memory for the late item

(precision: $p = 0.006$, one-tailed; guess-rate: $p = 0.942$, one-tailed; Fig. 4.7A), replicating the precision effect of Experiment 1. Note that there was again no relationship between accuracy and item decoding during the encoding phase ($p > 0.8$).

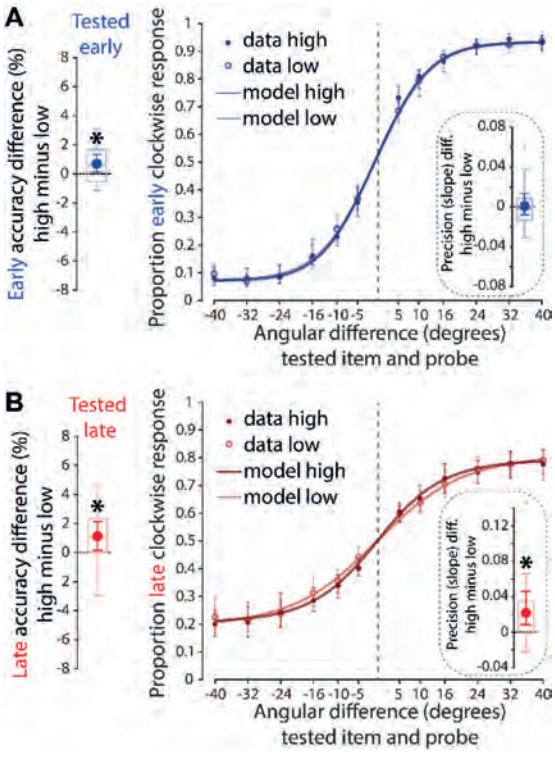


Figure 4.7. Relationship between decoding of relevant item at corresponding and WM accuracy. **(A)** Boxplot and superimposed circles and error-bars represent the difference in overall WM task performance between high and low early-tested item decoding trials during the first impulse (left). Proportion of clockwise responses for high and low decoding trials as a function of the angular difference between the memory item and the probe (right). Inset shows the boxplot and error-bar of the difference in the slope parameter (a measure of memory precision) between high and low decoding trials. **(B)** The same convention as in **A** but for the decoding of the late-tested item during the late impulse. Significant differences in accuracy/precision between high and low decoding trials are highlighted by asterisks (permutation test, $n = 19$, $p < 0.05$, two-sided and one-sided for accuracy and precision tests, respectively).

Experiment 3

We developed the impulse perturbation approach to reveal otherwise hidden neural states, without necessarily transforming the mnemonic representation (Stokes, 2015; Wolff et al., 2015/Chapter 3). This contrasts with other studies using retro-cues (Larocque et al., 2014; Lewis-Peacock et al., 2011; Sprague et al., 2016) or TMS (Rose et al., 2016) to 'reactivate' a latent item in working memory. However, to test whether our impulse stimulus actually did result in a behaviourally relevant transformation of the memory item (i.e., from a functionally latent to active state), we conducted an additional behavioural experiment ($n = 20$). Adapting the design of Experiment 1, we now varied

the presentation of the stimulus-onset asynchrony (SOA) between impulse and probe onset in Experiment 3 (SOA from 0 to 500ms; Fig. 4.8A). If the increase in impulse-specific decodability observed in both EEG experiments reflects a functional "reactivation" of an otherwise latent memory item, there should be a corresponding benefit to behaviour.

A repeated measures ANOVA provided no evidence for an effect of SOA ($F(4, 76) = 1.184$, $p = 0.325$). Uncorrected paired comparisons between the no-impulse condition (SOA 0 ms) and all other SOAs provided no evidence for an impulse-specific effect on accuracy for any SOA either (permutation test, $n = 20$, all $p > 0.12$; Fig. 4.8B). This suggests that our impulse stimulus is effective for 'pinging' activity silent neural states, without resulting in any behaviourally relevant transformation of the mnemonic representation.

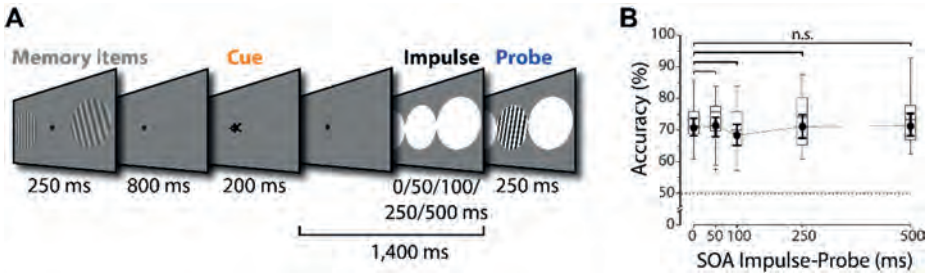


Figure 4.8. Task schematic and results of behavioural experiment. **(A)** Two memory items were presented, and participants were instructed to memorize both. A retro-cue indicated which item would be tested at the end of the current trial. The impulse stimulus was presented at varying delays (or not at all) and stayed on screen until the probe was presented. Participants indicated whether the probe was rotated clockwise or anti-clockwise relative to the orientation of the cued item. **(B)** Behavioural performance as a function of impulse-probe SOA. None of the uncorrected paired comparisons between the no-impulse condition (SOA 0 ms) and the other SOA conditions reached significance (permutation test, $n = 20$). Circles and error bars superimposed on the boxplots represent mean and 95% C.I. of the mean. Data points outside the 1.5 * interquartile range are marked as crosses in the boxplots.

Discussion

Recent theoretical models of WM predict a key role for activity-silent neural states in maintaining item-specific information (Lundqvist et al., 2016; Mongillo et al., 2008; Stokes, 2015). This raises a particular challenge for contemporary neuroscience that is

dominated by measurement and analysis of neural activation states. Here, we address this challenge using a perturbation approach to reveal hidden neural states that code the contents of WM. We show that the response to an impulse stimulus faithfully reflects item-specific information in WM. We further demonstrate that the impulse response reflects both attended and unattended items in WM, yet recently forgotten information leaves no detectable trace in the hidden state. Behavioural modelling further suggests that the hidden-state coding determines the quality of information in WM.

Previous evidence from non-human primates showed that a neutral visual stimulus presented during the WM delay period can elicit distinct patterns of neural activity that depend on recent visual input (Stokes et al., 2013). Although the previous work could not deconfound previous sensory stimulation and WM proper, the observed effect helped motivate a dynamic coding model for WM (Stokes, 2015). According to this framework, distinct memoranda are associated with distinct changes in neural response profile, which would be readable to downstream systems from the state-dependent response to a retrieval probe (Mongillo et al., 2008; Stokes et al., 2013). Crucially, WM depends on the maintenance of the item-specific neural response profile, rather than an explicit representation of an item in a persistent activity state. We now provide direct evidence for a WM-dependent impulse response decoupled from previous stimulation history, and further demonstrate that this WM state is highly flexible and coupled to behavioural performance. The hidden state for a specific item can be rapidly cleared if it is no longer relevant to the task, providing a striking neural correlate of directed forgetting in WM.

Recent retro-cuing evidence suggests that prioritising one WM item relative to other task-relevant items improves neural decoding of the cued item, whereas decoding of unattended items drops to chance levels even though the unattended information is still ultimately task relevant and retrievable at the end of the trial (Lewis-Peacock et al., 2011). Item-specific delay activity therefore seems to reflect the focus of attention, rather than WM per se (Larocque et al., 2014). The impulse response reported here clearly differs from the typical profile observed for decoding delay activity patterns. In Experiment 2, both attended and unattended items could be decoded from the impulse response of the hidden state as long as they are both still ultimately required for task performance. This suggests that if the information is successfully maintained in WM, there is a corresponding trace in the hidden state, irrespective of attentional priority. These results highlight the flexibility of WM, independently of switching attention between specific items in WM. Activity states appear to track the focus of attention (Larocque et al., 2014; Lewis-Peacock et al., 2011; Sprague et al., 2016), whereas hidden states, as revealed by the impulse response, more closely track the actual contents of WM.

Exactly how the proposed hidden state can be used for WM-guided behaviour remains an important open question. Computationally, supervised learning could determine the mapping between the memory-dependent probe response and the correct behavioural response (Mante et al., 2013), however such a learning strategy seems

implausible for real-world behaviour. Trial and error learning of arbitrary patterns does not seem a realistic model for WM, at least for humans. Instead, the inherent dynamics could establish a history-dependent match filter (Sugase-Miyamoto et al., 2008), which would be capable of transforming probe input to a common decision signal (i.e., match/no-match, or in our case clockwise/counter-clockwise). In Myers et al. (Myers et al., 2015), such a mechanism was shown to generate two distinct decision-related signals in an orientation detection task: a signed (i.e., directional) and unsigned difference signal, even though the signed difference was actually irrelevant to behaviour in that task. A similar process could underpin WM encoding in hidden states. The hidden state could establish a flexible, task dependent circuit for WM-dependent decision-making (Martínez-García, Rolls, Deco, & Romo, 2011). When the probe stimulus is presented, the hidden state transforms the input to decision-relevant output: e.g., direction of angular rotation. However, because the impulse stimulus used in these experiments does not contain decision-relevant features, the impulse response reflects an input-output transformation of the arbitrary input.

It may be noted that although the response to an arbitrary input is sufficient to 'read-out' the hidden state, it is unlikely to constitute an explicit 'reactivation' of the memory representation. In contrast, retro-cueing can convert an unattended item to a prioritised state in preparation for the recall (Griffin & Nobre, 2003). Similarly, a recent transcranial magnetic stimulation study suggests that stimulation of the visual cortex can also render an item active from its latent state (Rose et al., 2016). We find no evidence that our impulse stimulus reactivates the same pattern associated with stimulus processing. Moreover, a further behavioural experiment designed to test the possible behavioural consequences of our impulse stimulus provides no evidence that it interacts with the mnemonic representation. Rather, we argue that the impulse response simply 'echoes' the representational structure of the hidden state, but does not drive an explicit transformation of latent memories to a prioritized state.

It has long been assumed that WM maintenance depends on persistent neural activity (Curtis & D'Esposito, 2003). Instead, we propose that activity-silent neural states are sufficient to bridge memory delays. Activity-dependent transformations in hidden states determine the temporary coding properties of memory networks: i.e., dynamic coding (Stokes, 2015; Stokes et al., 2013). WM decisions are made by the state-dependent response to subsequent input. However, WM is also classically associated with active manipulation of content in short-term memory (Baddeley, 2003). We argue that such transformations are activity dependent, but the results of the transformation can be maintained in short term memory via latent network states. This alternative account does not ignore previous evidence for decodable activity during mnemonic delays, but rather attributes such evidence to focused attention (Rose et al., 2016), periodic (Mongillo et al., 2008) or stochastic (Lundqvist et al., 2016) updating, and/or response preparation (Barak et al., 2010). Interestingly, our current results also show that cue-directed forgetting can rapidly wipe the mnemonic representation from the hidden state. Rapid construction and dissolution of hidden states places important constraints on the basic mechanisms of hidden-state coding.

Although the present study addressed a specific model of WM, it is worth noting that the general impulse response approach for inferring otherwise silent neural states could also be particularly fruitful for exploring other tonic cognitive states, such as task set, attention and expectation. It is becoming increasingly apparent that we need to look beyond simple measures of neural activity, and consider a richer diversity of neural states that underpin context-dependent behaviour. Here we focus on perturbation to illuminate hidden states, but future work will also profit from more direct measures of functionally relevant hidden states (e.g., synaptic efficacy, membrane potentials, extra-cellular transmitter concentrations). This will require more sophisticated measurements in awake behaving animals, coupled with non-invasive approaches like described here for human studies.

Acknowledgements

We would like to thank the Biotechnology & Biological Sciences Research Council (BB/M010732/1 to M.G.S.), and the National Institute for Health Research Oxford Biomedical Research Centre Programme based at the Oxford University Hospitals Trust, Oxford University. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. We would also like to thank E. Spaak, A. Cravo, and N. Myers for helpful comments, and all our volunteers for their participation.

Chapter 5

Unimodal and bimodal access to sensory working memories by auditory and visual impulses

A slightly revised version of this chapter is published as:

Wolff, M. J., Kandemir, G., Stokes, M. G., & Akyürek, E. G. (2020). Unimodal and bimodal access to sensory working memories by auditory and visual impulses. *Journal of Neuroscience*, *40*(3), 671-681.

Data and code available at osf.io/u7k3q

Abstract

It is unclear to what extent sensory processing areas are involved in the maintenance of sensory information in working memory (WM). Previous studies have thus far relied on finding neural activity in the corresponding sensory cortices, neglecting potential activity-silent mechanisms such as connectivity-dependent encoding. It has recently been found that visual stimulation during visual WM maintenance reveals WM-dependent changes through a bottom-up neural response. Here, we test whether this impulse response is uniquely visual and sensory-specific. Human participants (both sexes) completed visual and auditory WM tasks while electroencephalography was recorded. During the maintenance period, the WM network was perturbed serially with fixed and task-neutral auditory and visual stimuli. We show that a neutral auditory impulse-stimulus presented during the maintenance of a pure tone resulted in a WM-dependent neural response, providing evidence for the auditory counterpart to the visual WM findings reported previously. Interestingly, visual stimulation also resulted in an auditory WM-dependent impulse response, implicating the visual cortex in the maintenance of auditory information, either directly, or indirectly as a pathway to the neural auditory WM representations elsewhere. In contrast, during visual WM maintenance only the impulse response to visual stimulation was content-specific, suggesting that visual information is maintained in a sensory-specific neural network, separated from auditory processing areas.

Introduction

Working memory (WM) is necessary to maintain information without sensory input, which is vital to adaptive behaviour. In spite of its important role, it is not yet fully clear how WM content is represented in the brain, or whether sensory information is maintained within a sensory-specific neural network. Previous research has relied on testing whether sensory cortices exhibit content-specific neural activity during maintenance. While this has indeed been shown for visual memories in occipital areas (e.g., Harrison & Tong, 2009) and, more recently, for auditory memories in the auditory cortex (Huang et al., 2016; Kumar et al., 2016; Uluç et al., 2018), WM-specific activity in the sensory cortex is not always present (Bettencourt & Xu, 2016), fuelling an ongoing debate over whether sensory cortices are necessary for WM maintenance (Scimeca et al., 2018; Xu, 2017). However, the neural WM network may not be solely based on measurable neural activity, and it has been proposed that information in WM may be maintained in an "activity-silent" network (Stokes, 2015) – for example, through changes in short-term connectivity (Mongillo et al., 2008). Potentially silent WM states should also be taken into account to better investigate the sensory-specificity account of WM.

Silent network theories predict that its neural "impulse" response to external stimulation can be used to infer its current state (Buonomano & Maass, 2009; Stokes, 2015). This has been shown in visual WM experiments, in which the evoked neural response from a fixed, neutral and task-irrelevant visual stimulus presented during the maintenance period of a visual WM task contained information about the contents of visual WM (Wolff et al., 2015/Chapter 3; Wolff, Jochim, Akyürek, & Stokes, 2017/Chapter 4). This not only suggests that otherwise hidden processes can be illuminated, but also implicates the involvement of the visual cortex in the maintenance of visual information, even when no ongoing activity can be detected. It has been suggested that this WM-dependent response profile might be not merely a by-product of connectivity-dependent WM, but a fundamental mechanism that affords efficient and automatic readout of WM content through external stimulation (Myers et al., 2015).

It remains an open question, however, whether information from other modalities in WM is similarly organized. If auditory WM depends on content-specific connectivity changes that include the sensory cortex, we would expect a network-specific neural response to external auditory stimulation. Furthermore, it may be hypothesized that sensory information need not necessarily be maintained in a network that is detached from other sensory processing areas. Direct connectivity (Eckert et al., 2008) and interplay (Iurilli et al., 2012; Martuzzi et al., 2007) between the auditory and visual cortices, or areas where information from different modalities converges, such as the parietal and pre-frontal cortices (Driver & Spence, 1998; Stokes et al., 2013), raise the possibility that WM could exploit these connections even during maintenance of unimodal information. Content-specific impulse responses might be observed not only during sensory-specific but also sensory non-specific stimulation.

In the present study, we tested whether WM-dependent impulse responses can be observed in visual and auditory WM, and whether that response is sensory specific. We measured electroencephalography (EEG) while participants performed visual and auditory WM tasks. We show that the evoked neural response of an auditory impulse stimulus reflects relevant auditory information maintained in WM. Visual perturbation also resulted in an auditory WM-dependent neural response, implicating both the auditory and visual cortices in auditory WM. By contrast, visual WM content could only be decoded after visual, but not auditory perturbation, suggesting that visual information is maintained in a sensory-specific visual WM network with no evidence for a WM-related interplay with the auditory cortex.

Methods

Participants

Thirty healthy adults (12 female, mean age 21 years, range 18-31 years) were included in the main analyses of the auditory WM experiment and 28 healthy adults (11 female, mean age 21 years, range 19-31 years) of the visual WM experiment. Three additional participants in the auditory WM experiment and 8 additional participants in the visual WM experiment were excluded during pre-processing due to excessive eye movements (more than 30% of impulse epochs contaminated). The exclusion criterion and resulting minimum number of trials for the multivariate pattern analysis were similar to our previous study (Wolff et al., 2017). Participants received either course credits or monetary compensation (8€ an hour) for participation and gave written informed consent. Both experiments were approved by the Departmental Ethical Committee of the University of Groningen (approval number: 16109-S-NE).

Apparatus and Stimuli

Stimuli were controlled by Psychtoolbox, a freely available toolbox for Matlab. Visual stimuli were generated with Psychtoolbox and presented on a 17-inch (43.18 cm) CRT screen running at 100 Hz refresh rate and a resolution of 1280 by 1024 pixels. Auditory stimuli were generated with the freely available software Audacity and were presented with stereo Logitech computer speakers. The intensity of all tones was adjusted to 70 dB SPL at a fixed distance of 60 cm between speakers and participants in both experiments. All tones had 10 ms ramp up and ramp down time. Responses were collected with a custom two-button response box, connected via a USB interface.

The memory items used in the auditory WM experiment were 8 pure tones, ranging from 270 Hz to 3055 Hz in steps of half an octave. The probes in the auditory experiment were 16 pure tones that were one-third of an octave higher or lower than the corresponding auditory memory items.

The memory items used in the visual WM experiment were 8 sine-wave gratings with orientations of 11.25° to 168.75° in steps of 22.5°. The visual probes were 16 sine-

wave gratings that were rotated 20° clockwise or counter-clockwise relative to the corresponding visual memory items. All gratings were presented at 20% contrast, with a diameter of 6.5° (at 60 cm distance) and a spatial frequency of 1 cycle per degree. The phase of each grating was randomized within and across trials.

The remaining stimuli were the same in both experiments. The retro-cue was a number ("1" or "2") that subtended 0.7°. The visual impulse stimulus was a white circle with a diameter of 12°. The auditory impulse was a complex tone consisting of the combination of all pure tones used as memory items in the auditory task. A grey background (RGB = 128, 128, 128) and a black fixation dot with a white outline (0.25°) were maintained throughout the trials. All visual stimuli were presented in the centre of the screen.

Procedure

The trial structure was the same in both experiments, as shown in Figure 5.1 (panels A and C). In both cases, participants completed a retro-cue WM task. Only the memory items and probes differed between experiments. Memory items and probes were pure tones in the auditory WM task and sine-wave gratings in the visual WM task. Each trial began with the presentation of a fixation dot, which stayed on the screen throughout the trial. After 1,000 ms the first memory item was presented for 200 ms. After a 700 ms delay the second memory item in the same modality as the first item was presented for 200 ms. Each memory item was selected randomly without replacement from a uniform distribution of 8 different tonal frequencies or grating orientations (see above) for the auditory and visual experiment, respectively. After another delay of 700 ms, the retro-cue was presented for 200 ms, indicating to participants whether the first or second memory item would be tested at the end of the trial. After a delay of 1,000 ms the impulse stimuli (the visual circle and the complex tone) were presented serially for 100 ms each with a delay of 900 ms in-between.

The order of the impulses was fixed for each participant but counter-balanced between participants. Impulse order was fixed within participants for two reasons: First, it removed the effect of surprise by making the order of events within trials perfectly consistent and predictable (Wessel & Aron, 2017), ensuring minimal intrusion by the impulse stimuli during the maintenance period. Second, random impulse order might have resulted in qualitatively different neural responses of each impulse, depending on when it was presented, due to different trial histories and elapsed maintenance duration at the time of impulse onset (Buonomano & Maass, 2009). This would have necessitated splitting the neural data by impulse order for the decoding analyses, resulting in reduced power.

The probe stimulus followed 900 ms after the second impulse offset and was presented for 200 ms. In the auditory WM experiment the probe was a pure tone and the participant's task was to indicate via button press on the response box whether the probe's frequency was lower (left button) or higher (right button) than the cued memory item. In the visual task, the probe was another visual grating, and the participants

indicated whether it was rotated counter-clockwise (left button) or clockwise (right button) relative to the cued memory item. The direction of the tone or tilt was selected randomly without replacement from a uniform distribution. After each response, a smiley face was shown for 200 ms, which indicated whether the response was correct or incorrect. The next trial began automatically after a randomized, variable delay of 700-1,000 ms after response input. Each experiment consisted of 768 trials in total and lasted approximately 2 hours.

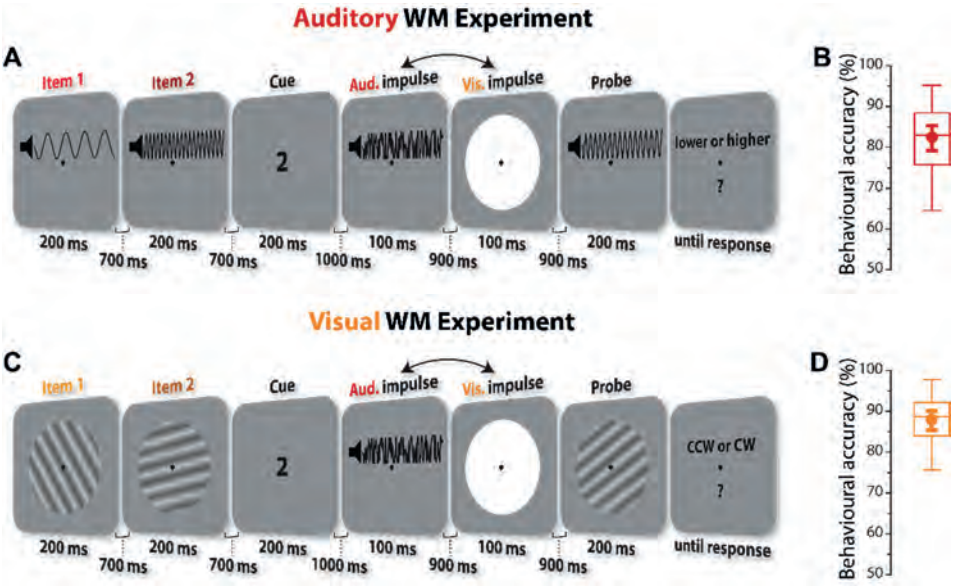


Figure 5.1. Task structure and behavioural performance. **(A)** Trial schematic of auditory task. Two randomly selected pure tones (270 Hz to 3055 Hz) were serially presented and a retro-cue indicated which of those tones would be tested at the end of the trial. In the subsequent delay, two irrelevant impulse stimuli (a complex tone and a white circle) were serially presented. At the end of each trial another pure tone was presented (the probe), and participants were instructed to indicate whether the frequency of the previously cued tone was higher or lower than the probe's frequency. **(B)** The boxplot shows auditory task accuracy. Centre line indicates the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range. The superimposed circle and error bars indicate the mean and its 95% C.I., respectively. **(C)** Trial schematic of visual task. The trial structure was the same as in the auditory task. Instead of pure tones, memory items were randomly orientated gratings. The probe was another orientation grating, and participants were instructed to indicate whether the cued item's orientation was rotated clockwise or counter-clockwise relative to the probe's orientation. **(D)** Visual task performance.

EEG acquisition and pre-processing

The EEG signal was acquired from 62 Ag/AgCls sintered electrodes laid out according to the extended international 10-20 system. An analogue-to-digital TMSI Refa 8-64/72 amplifier and Brainvision recorder software were used to record the data at 1000 Hz using an online average reference. An electrode placed just above the sternum was used as the ground. Bipolar electrooculography (EOG) was recorded by electrodes placed above and below the right eye, and to the left and right of the left and right eye, respectively. The impedances of all electrodes were kept below 10 kΩ.

Offline the data was down-sampled to 500 Hz and bandpass filtered (0.1 Hz high-pass and 40 Hz low-pass) using EEGLAB (Delorme & Makeig, 2004). The data was epoched relative to the onsets of the memory items (-150 ms to 900 ms) and to the onsets of the auditory and visual impulse stimuli (-150 to 500 ms). The signal's variance across channels and trials was visually inspected using a visualization tool provided by the Matlab extension Fieldtrip (Oostenveld et al., 2010), and especially noisy channels were removed and replaced through spherical interpolation. This led to the interpolation of 1 channel in 3 participants and 2 channels in 1 participant in the auditory WM task, and 1 channel in 5 participants and 5 channels in 1 participant in the visual WM task. Noisy epochs were removed from all subsequent electrophysiological analyses. Epochs containing any artefacts related to eye movements were identified by visually inspecting the EOG signals and also removed from analyses. The following percentage of trials were removed for each epoch in the auditory WM experiment: item 1 epoch (M = 13.39%, SD = 6.08%), item 2 epoch (M = 9.28%, SD = 4.42%), auditory impulse epoch (M = 11.53%, SD = 7.03%), visual impulse epoch (M = 9.81%, SD = 5.44%). The following percentage of trials were removed for each epoch in the visual WM experiment: item 1 epoch (M = 19.81%, SD = 5.91%), item 2 epoch (M = 20.69%, SD = 5.88%), auditory impulse epoch (M = 18.51%, SD = 5.73%), visual impulse epoch (M = 19.33%, SD = 4.94%).

Multivariate pattern analysis of neural dynamics

We wanted to test if the electrophysiological activity evoked by the memory-stimuli and impulse-stimuli contained item-specific information. Since event-related potentials are highly dynamic, we used an approach that is sensitive to such changing neural activity within pre-defined time-windows, by pooling relative voltage fluctuations over space (i.e., electrodes) and time. This approach has two key benefits: First, pooling information over time (in addition to space) multivariately can boost decoding accuracy (Grootswagers, Wardle, & Carlson, 2017; Nemrodov, Niemeier, Patel, & Nestor, 2018). Secondly, by removing the mean-activity level within each time-window, the voltage fluctuations are normalized. This is similar to taking a neutral pre-stimulus baseline common in ERP analysis. Notably, this also removes stable activity traces that do not change within the chosen time-window, making this approach ideal to decode transient, stimulus-evoked activation patterns, while disregarding more stationary neural processes.

The following details of the analyses were the same for each experiment, unless explicitly stated.

For the time-course analysis, we used a sliding window approach that takes into account the relative voltage changes within a 100 ms window. The time-points within 100 ms of each channel and trial were first down-sampled by taking the average every 10 ms, resulting in 10 voltage values for each channel. Next, the mean activity within that time-window of each channel was subtracted from each individual voltage value. All 10 voltage values per channel were then used as features for the 8-fold cross-validation decoding approach.

We used Mahalanobis distance (De Maesschalck et al., 2000) to take advantage of the potentially parametric neural activity underlying the processing and maintenance of orientations and tones. The distances between each of the left-out test-trials and the averaged, condition-specific patterns of the train-trials (tones and orientations in the auditory and visual experiment, respectively), were computed, with the covariance matrix estimated from the train-trials using a shrinkage estimator (Ledoit & Wolf, 2004). To acquire reliable distance estimates, this process was repeated 50 times, where the data was randomly partitioned into 8 folds using stratified sampling each time. The number of trials of each condition (orientation/tone frequency) of the 7 train-folds were equalized by randomly subsampling the minimum number of condition-specific trials to ensure an unbiased training set. The average was then taken of these repetitions. For each trial the 8 distances (one of each stimulus condition) were sign-reversed for interpretation purposes, so that higher values reflect higher pattern-similarity between test and train-trials. For visualization, the sign-reversed distances were furthermore mean-centred by subtracting the mean distance of all distances of a given trial and ordered as a function of tone difference, in 1 octave steps by averaging over adjacent half octave differences, and orientation difference.

To summarize the expected positive relationship between tone-similarity and neural activation similarity (indicative of tone-specific information in the recorded signal) into a single value in the auditory WM experiment, the absolute tonal differences were linearly regressed against the corresponding pattern similarity values for each trials. The obtained beta values of the slopes were then averaged across all trials to represent "decoding accuracy", where high values suggest a strong positive effect of tone similarity on neural pattern similarity.

To summarize the tuning curves in the visual WM experiment, we computed the cosine vector means (Wolff et al., 2017/Chapter 4), where high values suggest evidence for orientation decoding.

The approach described above was repeated in steps of 8 ms across time (-52 to 900 ms relative to item 1 and 2 onset, and -52 to 500 ms relative to auditory and visual onset). The decoding values were averaged over trials, and the decoding time-course was smoothed with a Gaussian smoothing kernel (s.d. = 16 ms). Within the time-window, information was pooled from -100 to 0 ms relative to a specific time-point. By only including data-points from before the time-point of interest, it is ensured that decoding

onsets can be more easily interpreted, whereas decoding offsets should be interpreted with caution (Grootswagers et al., 2017). In addition to the sliding window approach, we also pooled information multivariately across the whole time-window of interest (Nemrodov et al., 2018). As before, the data was first down-sampled by taking the average every 10 ms, and the mean activity from 100 to 400 ms relative to impulse onset was subtracted. The resulting 30 values per channel were then provided to the multivariate decoding approach in the same way as above, resulting in a single decoding value per participant. The time-window of interest was based on previous findings showing that the WM-dependent impulse-response is largely confined within that window (Wolff et al., 2017/Chapter 4). Additionally, items in the item-presentation epochs were also decoded using each channel separately, using the data from 100-400 ms relative to onset. Decoding topographies were visualized using fieldtrip (Oostenveld et al., 2010).

Cross-epoch generalization analysis

We also tested if WM-related decoding in the impulse epochs generalized to the memory presentation. Instead of using the same epoch (100-400 ms) for training and testing, as described above, the classifier was trained on the memory item epoch and tested on the impulse epoch that contained significant item decoding (and vice versa). In the auditory task, we also tested if the different impulse epochs cross-generalized by training on the visual and testing on the auditory impulse (and vice versa).

Representational similarity analysis

While the decoding approach outlined above takes into account the potentially parametric relationship of pitch/orientation difference, it is not an explicit test for the presence of a parametric relationship. Indeed, decodability could theoretically be solely driven by high within stimulus-condition pattern similarity, and equally low pattern similarities of all between stimulus-condition comparisons. To explicitly test for a linear/circular relationship between stimuli, and explore additional stimulus coding schemes, we used representational similarity analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008).

The RSA was based on the mahalanobis distances between all stimulus conditions (unique orientations and frequencies) in both experiments using the same time-window of interest as in the decoding approach described above (100-400 ms relative to stimulus onset). For each participant, the number of trials of each stimulus condition were equalized by randomly subsampling the minimum number of trials of a condition before taking the average across all same stimulus condition trials and computing all pairwise mahalanobis distances. This procedure was repeated 50 times, with random subsamples each time, before averaging them all into a single representation dissimilarity matrix (RDM). The covariance matrix was computed from all trials using the shrinkage estimator (Ledoit & Wolf, 2004). Since each experiment contained 8 unique memory items, this resulted in an 8 x 8 RDM for each participant and epoch of interest.

For the RSA in the auditory WM experiment we considered two models; a positive linear relationship between absolute pitch height difference (i.e., the more dissimilar pitch frequency, the more dissimilar the brain activity patterns), and a positive relationship of pitch chroma (i.e. higher similarity between brain activity patterns of the same pitch chromas). Note that the tone frequencies used in the experiment increased in half octave steps. Every other tone had thus the same pitch chroma (i.e. the same note in a different octave). The model RDMs are shown for illustration in Figure 5.4A. The model RDMs were z-scored to make the corresponding model fits between them more comparable, before entering both of them into a multiple regression analysis with the data RDM.

In the visual WM experiment we also considered two models. The first model was designed to capture the circular relationship between absolute orientation difference (i.e., the more dissimilar the orientation, the more dissimilar the brain activity patterns). The second model was designed to capture the specialization of cardinal orientations (i.e., horizontal and vertical) that could reflect the "oblique effect", where orientations close to the cardinal axes are discriminated and recalled more accurately than more oblique orientations (Appelle, 1972; Pratte, Park, Rademaker, & Tong, 2017). The model assumed the extreme case, where orientations are clustered into one of three categories depending on their circular distance to vertical, horizontal, or oblique angles. This captures the relatively higher dissimilarity and distinctiveness of the cardinal axes (vertical and horizontal) compared to the oblique axes (-45 degrees and +45 degrees) and reflects neurophysiological findings of an increased number of neurons tuned to the cardinal axes (Shen, Tao, Zhang, Smith, & Chino, 2014). The model RDMs are shown for illustration in Figure 5.4D. The model RDMs were also z-scored and then both included into a multiple regression with the data RDM

Statistical significance testing

All statistical tests were the same between experiments. Sample sizes of all analyses were n=30 and n=28 in the auditory and visual tasks, respectively. Sample size of the event-related potential (ERP) analyses as a function of impulse modality and task was n=16, as it only included participants who participated in both WM tasks. To determine if the decoding values (see above) or model fits of the RSA are higher than 0 or different between items, or if the evoked potentials were different between tasks, we used a non-parametric sign-permutation test (Maris & Oostenveld, 2007). The sign of the decoding value, model fit value, or voltage difference of each participant was randomly flipped 100.000 times with a probability of 50%. The p value was derived from the resulting null distribution. The above procedure was repeated for each time-point for time-series results. A cluster based permutation test (100.000 permutations) was used to correct for multiple comparisons over time using a cluster forming and cluster significance threshold of $p < 0.05$. Complementary Bayes factors to test for decoding evidence for the cued and uncued items within each impulse epoch separately were also computed.

We were also interested if there were differential effects on the decoding results between cueing (cued/uncued) and impulse modality (auditory/visual) during WM maintenance. To test this, we computed the Bayes factors of models with and without each of these predictors versus the null model that only included subjects as a predictor (Bayesian equivalent of repeated measures ANOVA). The freely available software package JASP (JASP Team, 2018) was used to compute Bayes factors.

Code and data availability

All data and custom Matlab scripts used to generate the results and figures of this manuscript will be made available upon peer-reviewed publication.

Results

Behavioural results

Behavioural task performance was $M = 82.322\%$, $SD = 8.841\%$ in the auditory WM task (Fig. 5.1B), and $M = 87.908\%$, $SD = 6.374\%$ in the visual WM task (Fig. 5.1D). Note that while task performance seemed to be slightly better in the visual WM task, participants performed well above chance in both, suggesting that the relevant sensory features were reliably remembered and recalled in both tasks.

Decoding visual and auditory stimuli

Auditory WM task

The neural dynamics of auditory stimulus processing suggest a parametric effect, with a positive relationship between tone and pattern similarity (Fig. 5.2A) for both memory items. The neural dynamics showed significant item-specific decoding clusters during, and shortly after, corresponding item presentation for item 1 (44 to 708 ms relative to item 1 onset, $p < 0.001$, one-sided, corrected) and item 2 (28 to 572 ms relative to item 2 onset, $p < 0.001$, one-sided, corrected; Fig. 5.2B). The topographies of channel-wise item-decoding for each item using the neural data from 100-400 ms after item-onset, revealed strong decoding for frontal-central and lateral electrodes (Fig. 5.2C), suggesting that the tone-specific neural activity is most likely generated by the auditory cortex (Chang, Bosnyak, & Trainor, 2016). These results provide evidence that stimulus-evoked neural activity fluctuations contain information about presented tones that can be decoded from EEG.

Visual WM task

Processing of visual orientations also showed a parametric effect (Fig. 5.2D), replicating previous findings (Saproo & Serences, 2010). The item-specific decoding time-courses of the dynamic activity showed significant decoding clusters during and shortly after item presentations (item 1: 84-724 ms, $p < 0.001$; item 2: 84-636 ms, $p < 0.001$, one-sided, corrected; Fig. 5.2E). As expected, the topographies of channel-wise item-decoding

showed strong effects in posterior channels that are associated with the visual cortex (Fig. 5.2F).

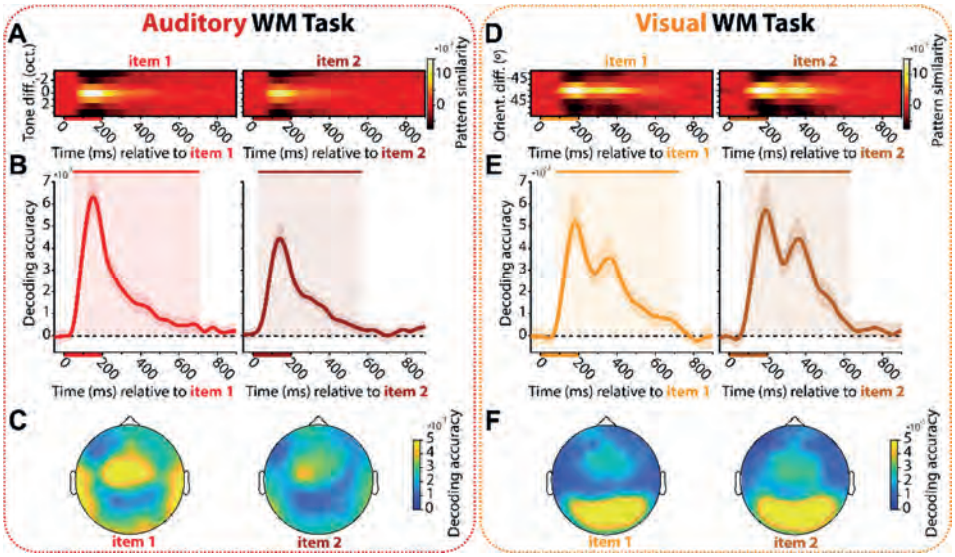


Figure 5.2. Decoding during item encoding. **(A-C)** Auditory WM task. **(D-F)** Visual WM task. **(A & D)** Normalized average pattern similarity (mean-centred, sign-reversed mahalanobis distance) of the neural dynamics for each time-point between trials as a function of tone similarity in A and orientation similarity in D, separately for item 1 and item 2, in item 1 and item 2 epochs, respectively. Bars on the horizontal axes indicate item presentations. **(B & E)** Beta values in B and cosine vector-means in E of pattern similarities for item 1 and 2. Upper bars and corresponding shading indicate significant values. Error shading indicates 95% C. I. of the mean. **(C & F)** Topographies of each item of channel-wise decoding (100-400 ms relative to item onset).

Content-specific impulse responses

Auditory WM task

In the auditory impulse epoch, the neural dynamics time-course revealed significant cued-item decoding (180-308 ms, $p = 0.004$, one-sided, corrected), while no clusters were present for the uncued item (Fig. 5.3A & B, left). Similarly, the cued item was decodable in the visual impulse epoch (204-372 ms, $p = 0.009$, one-sided, corrected), while the uncued item was not (Fig. 5.3A & B, right).

The time-of-interest (100-400 ms relative to impulse onset) analysis provided similar results. The cued item showed strong decoding in both impulse epochs (auditory impulse: Bayes factor = 11462.607, $p < 0.001$; visual impulse: Bayes factor = 85.843, $p < 0.001$, one-sided), but the uncued item did not (auditory impulse: Bayes factor = 0.968, $p = 0.075$; visual impulse: Bayes factor = 0.204, $p = 0.476$, one-sided; Fig. 5.3C). A model only including the cueing predictor yielded the highest Bayes factor of 8.123 (± 0.996 %) compared to the null model. A model including impulse modality as a predictor resulted in a Bayes factor of 0.848 (± 1.075 %). Including both predictors (impulse modality and cueing) in the model resulted in a Bayes factor of 7.553 (± 0.991 %) that was slightly lower than only including cueing.

Taken together, these results provided strong evidence that both impulse stimuli elicit neural responses that contain information about the cued item in auditory WM, but none about the uncued item.

Visual WM task

No significant time clusters were present in the auditory impulse epoch of the visual WM experiment for either the cued or the uncued item task (Fig. 5.3D & E, left). The decoding time-course of the visual impulse epoch revealed a significant decoding cluster of the cued item (108-396 ms, $p < 0.001$, one-sided, corrected) but not for the uncued item (Fig. 5.3D & E, right), replicating previous findings (Wolff et al., 2017/Chapter 4).

The analysis on the time-of-interest interval (100-400 ms) showed the same pattern of results; neither the cued, nor uncued item in the auditory impulse epoch showed above 0 decoding (cued: Bayes factor = 0.236, $p = 0.417$; uncued: Bayes factor = 0.119, $p = 0.787$, one-sided). In the visual impulse epoch the cued item showed strong decodability (Bayes factor = 1695.823, $p < 0.001$, one-sided) but the uncued item did not (Bayes factor = 0.236, $p = 0.421$, one-sided; Fig 5.3F). A model including both predictors (cueing and impulse modality) as well as their interaction resulted in the highest Bayes factor compared to the null model (Bayes factor = 56.284 (± 1.557 %)). Models with each predictor alone resulted in notably smaller Bayes factors (cueing: Bayes factor = 6.26 (± 0.398 %); impulse modality: Bayes factor = 5.877 (± 0.686 %)). The Bayes factor of the model including both predictors without interaction (46.728 (± 0.886 %)) was only 1.205 times smaller than the model that also included the interaction, highlighting that while there was strong evidence in favour of both impulse modality and cueing, there was only weak evidence in favour of an interaction.

Overall, these results provided evidence that while a visual impulse clearly evokes a neural response that contains information about the cued visual WM item, replicating previous findings (Wolff et al., 2017/Chapter 4), an auditory impulse does not.

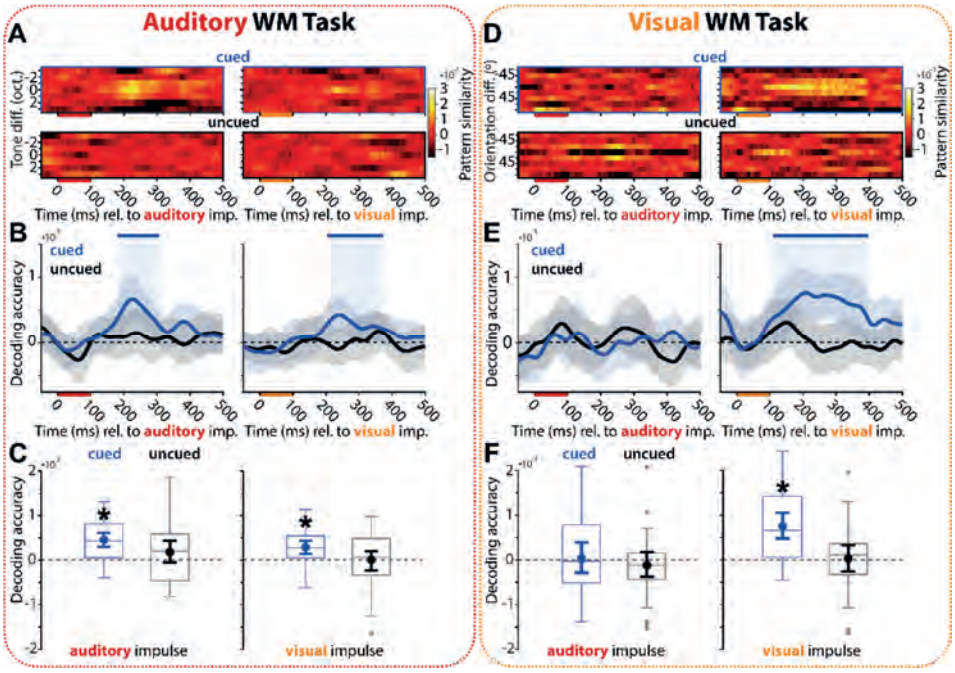


Figure 5.3. Decoding auditory and visual WM content from the impulse response. **(A-C)** Auditory WM task. **(D-F)** Visual WM task. **(A & D)** Normalized average pattern similarity (mean-centred, sign-reversed mahalanobis distance) of the neural dynamics for each time-point between trials as a function of tone similarity in A and orientation similarity in D. Top row: cued item. Bottom row: uncued item. Left column: auditory impulse. Right column: visual impulse. **(B & E)** Decoding accuracy time-course: Beta values in B and cosine vector-means in E of pattern similarities for cued (blue) and uncued item (black). Upper bars and shading indicate significant values of the corresponding item. Error shading indicate 95% C. I. of the mean. **(C & F)** Boxplots show the overall decoding accuracies for the cued (blue) and uncued (black) item, using the whole time-window of interest (100-400 ms relative to onset) from the auditory (left) and visual (right) impulse epoch. Centre lines indicate the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range. Extreme values are shown separately (dots). Superimposed circles and error bars indicate mean and its 95% C.I., respectively. Asterisks indicate significant decoding accuracies ($p < 0.05$, one-sided).

Parametric encoding and maintenance of auditory pitch and visual orientation

As indicated, RSA was performed to explicitly test and explore for specific stimulus coding relationships in both experiments (Fig. 5.4A & D).

Auditory WM task

The RDMs of each epoch of interest are shown in Fig. 5,4B. There was strong evidence in favour of the pitch height difference model during item encoding (item 1 and item 2 presentation epochs; Bayes factor > 100.000, $p < 0.001$, one-sided) while evidence against the pitch chroma model was evident (Bayes factor = 0.177, $p = 0.523$, one-sided; Fig. 5,4B & C, left). Moderate evidence in favour of the pitch height model was also evident for the cued item in the auditory impulse epoch (Bayes factor = 4.016, $p = 0.0113$, one sided), while there was weak evidence against the pitch chroma model (Bayes factor = 0.838, $p = 0.079$, one sided; Fig. 5,4B & C, middle). The visual impulse epoch also suggested a pitch height coding model of the cued auditory item, though the evidence was weak (Bayes factor = 1.346, $p = 0.049$, one sided), and there was again evidence against the pitch chroma model of the cued item (Bayes factor = 0.123, $p = 0.736$, one-sided; Fig. 5,4B & C, right).

Overall, these RSA results provide evidence that both the encoding and maintenance of pure tones are coded parametrically according to pitch height (Uluç et al., 2018), but not pitch chroma.

Visual WM task

The RDMs of the averaged encoding epochs (item 1 and item 2) and the visual impulse epoch are shown in Fig. 5.4E. There was strong evidence in favour for a circular orientation difference code (Bayes factor > 100.000, $p < 0.001$, one-sided), as well as an additional "cardinal specialization" code (Bayes factor > 100.000, $p < 0.001$, one-sided) during item encoding (Fig. 5.4E & F, left). The evoked neural response by the visual impulse also provided strong evidence for a circular orientation difference code for the maintenance of the cued item (Bayes factor = 362.672, $p < 0.001$, one-sided). No evidence in favour of an additional "cardinal specialization" code during maintenance was found, however (Bayes factor = 0.252, $p = 0.318$, one-sided; Fig. 5.4E & F, right).

These results provide evidence that orientations are encoded and maintained in a parametric orientation selective code (e.g. Ringach, Shapley, & Hawken, 2002; Saproo & Serences, 2010). We additionally considered the "cardinal specialization" coding model, which captures the expected increased neural distinctiveness of horizontal and vertical orientations compared to tilted orientations, based on the superior visual discrimination of cardinal orientations (Appelle, 1972) as well as previous neurophysiological reports of cardinal specialization (Li, Peterson, & Freeman, 2003; Shen et al., 2014). Evidence for this model was only found during orientation encoding, but not maintenance.

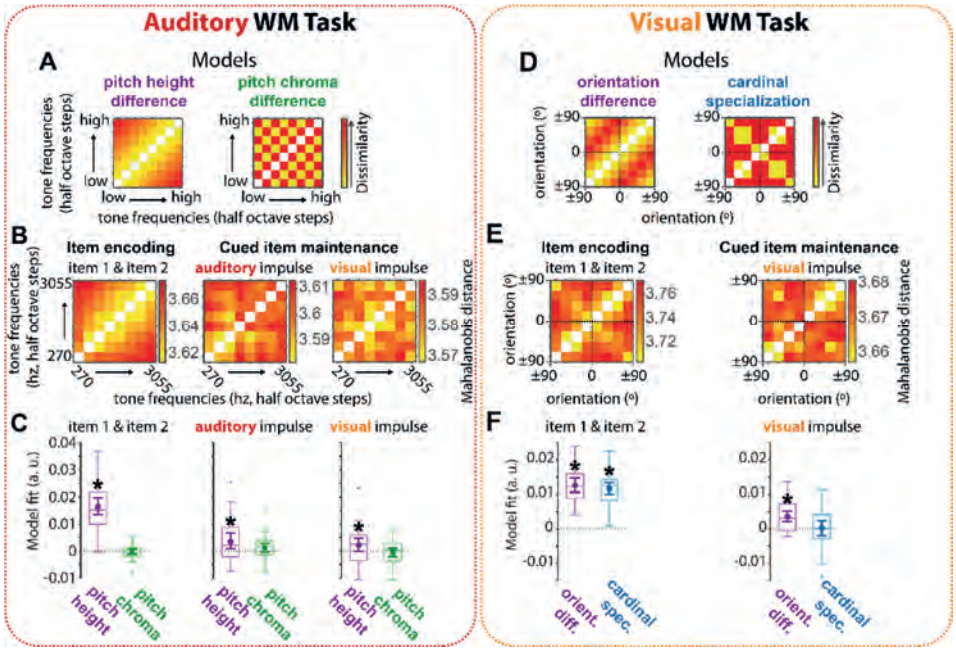


Figure 5.4. Stimulus coding relationship during encoding and maintenance. **(A-C)** Auditory WM task. **(D-F)** Visual WM task. **(A & D)** Model RDMs of pitch (**A**) and orientation (**D**). **(B & E)** Data RDMs. **(C & F)** Model fits of model RDMs on data RDMs. Centre lines indicate the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range. Extreme values are shown separately (dots). Superimposed circles and error bars indicate mean its 95% C.I, respectively. Asterisks indicate significant model fits ($p < 0.05$, one-sided).

No WM-specific cross-generalization between impulse and WM-item presentation

It has been shown previously that the visual WM-dependent impulse-response does not cross-generalize with visual item processing (Wolff et al., 2015/Chapter 3). Here we tested if this is also the case for auditory WM, and additionally explored the cross-generalizability between impulses.

Auditory WM task

The representation of the cued item did neither cross-generalise between item presentation and either of the impulse epochs (auditory impulse: Bayes factor = 0.225, p = 0.58; visual impulse: Bayes factor = 0.356, p = 0.26, two-sided), nor between impulse epochs (Bayes factor = 0.267, p = 0.417, two-sided; Fig. 5.5A).

Visual WM task

Replicating previous reports (Wolff et al., 2015/Chapter 3, 2017/Chapter 4), the visual impulse response of the cued visual item did not cross-generalize with item processing during item presentation (Bayes factor = 0.491, p = 0.168; Fig. 5.5B).

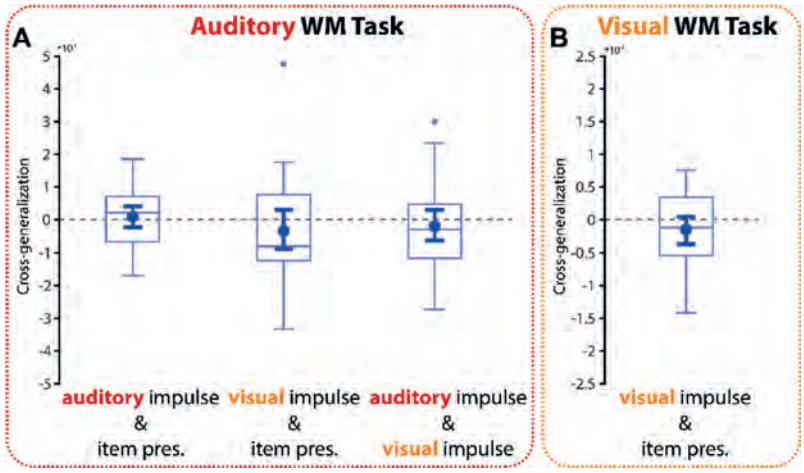


Figure 5.5. Cross-generalization between epochs. **(A)** Cross-generalization of the cued item between the memory item epoch and impulse epochs in the auditory WM task. **(B)** Cross-generalization between visual impulse and memory item in the visual WM task. Centre lines indicate the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range. Extreme values are shown separately (dots). Superimposed circles and error bars indicate mean its 95% C.I, respectively.

5

Evoked response magnitudes of impulse stimuli are comparable between tasks

Since the impulse stimuli were always the same across trials, presented at the same relative time within each trial, and were completely task irrelevant, we believe that the WM-specific impulse responses reported here and in previous work rely on low-level interactions of the impulse stimuli with the WM network, which do not depend on higher order cognitive processing of the impulse.

Nevertheless, it could be argued that the impulse stimuli are differentially processed even at an early stage between the WM tasks. Since the auditory impulse was the only auditory stimulus in the visual WM task, it may have been more easily filtered out and ignored compared to the other impulse stimuli. Indeed, it is possible that the neural response to the auditory impulse stimulus was just too "weak" to result in measurable,

WM-specific neural response in the visual WM task. However, given the uniqueness of the auditory impulse in the visual WM task, the opposite could be argued as well.

To test for potential differences of attentional filtering of impulse stimuli between tasks, we examined the event-related potentials (ERPs) to the impulse stimuli in both tasks. If there is indeed a difference in early processing, this should be visible in associated early evoked responses within 250 ms of stimulus presentation (Boutros, Korzyukov, Jansen, Feingold, & Bell, 2004; Luck, Woodman, & Vogel, 2000). Due to large individual differences in ERPs, only participants who participated in both tasks (n = 16) are included in this analysis.

Auditory ERPs

The average auditory ERP (Fz, FCz, Cz) evoked from the auditory impulse stimulus within each task is shown in fig. 5.6A. The P50, N100, and P200 components, all of which have been shown to be reduced when irrelevant auditory stimuli are filtered out (sensory gating; e.g., Boutros et al., 2004; Cromwell, Mears, Wan, & Boutros, 2008; Kisley, Noecker, & Guinther, 2004), can clearly be identified in both tasks. One time-cluster of the difference between tasks was significant within the time-window of interest (148 to 184 ms, $p = 0.048$, two-sided, corrected). Visual inspection of the ERPs suggests that while there is no difference in P50 and N100 amplitude between tasks, P200 amplitude is larger in the visual than in the auditory task. Note that this difference goes in the opposite direction as would be expected if the auditory impulse stimulus was somehow more easily filtered out and ignored in the visual than in the auditory task.

Visual ERPs

The visual impulse ERP recorded from occipital electrodes (O1, Oz, O2) is shown in fig. 5.6B. Early components of interest (C1, P1, N1), which have been shown to be modulated by attentional processes (e.g. Di Russo, Martínez, & Hillyard, 2003; Luck et al., 2000; Rauss, Pourtois, Vuilleumier, & Schwartz, 2009), have been marked. Visual inspection suggests that there are no discernible differences in these visual components between tasks. Indeed, no significant time-clusters were found ($p > 0.19$, two-sided, corrected), suggesting that the visual impulse stimulus was processed similarly between tasks.

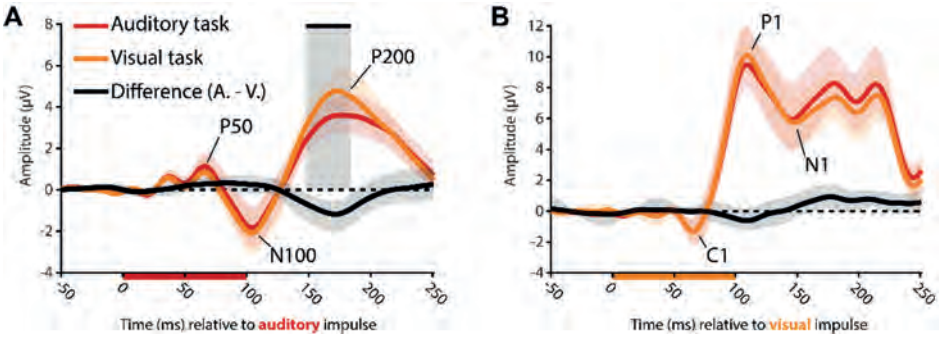


Figure 5.6. Evoked responses to impulse stimuli as a function of task for participants who participated in both tasks (n=16). **(A)** Average voltages (electrodes Fz, FCz, Cz) evoked by auditory impulse in the auditory task (red) and visual task (orange). Difference voltage (auditory task minus visual task) is plotted in black. Individual ERP components of interest are labelled. Error shadings are 95% C. I. of the mean. The significant time-cluster of difference is indicated by the black bar ($p < 0.05$, corrected, two-sided). **(B)** Average voltages (electrodes O1, Oz, O2) evoked by the visual impulse. Same convention as in A.

Discussion

It has previously been shown that the bottom-up neural response to a visual impulse presented during the delay of a visual WM task contains information about relevant visual WM content (Wolff et al., 2015/Chapter 3, 2017/Chapter 4), which is consistent with a key prediction of WM theories that assume information is maintained in activity-silent brain states (Stokes, 2015). We used this approach to investigate whether sensory information is maintained within sensory-specific neural networks, shielded from other sensory processing areas. We show that the neural impulse response to sensory-specific stimulation is WM content-specific not only in visual WM, but also in auditory WM, demonstrating the feasibility and generalisability of the approach in the auditory domain. Furthermore, for auditory WM, a content-specific response was obtained not only during auditory, but also during visual stimulation, suggesting a sensory modality-unspecific path to access the auditory WM network. In contrast, only visual, but not auditory, stimulation evoked a neural response containing relevant visual WM content. This pattern of impulse responsivity supports the idea that visual pathways may be more dominant in WM maintenance.

Recent studies have shown that delay activity in the auditory cortex reflects the content of auditory WM (Huang et al., 2016; Kumar et al., 2016; Uluç et al., 2018). Thus, similar to visual WM maintenance, which has been found to result in content-specific delay-activity in the visual cortex (Harrison & Tong, 2009), auditory WM content is also maintained in a network that recruits the same brain area responsible for sensory

processing. However, numerous visual WM studies have shown that content-specific delay activity may in fact reflect the focus of attention (Lewis-Peacock et al., 2011; Sprague et al., 2016; K. Watanabe & Funahashi, 2014). The memoranda themselves may instead be represented within connectivity patterns that generate a distinct neural response profile to internal or external neural stimulation (Lundqvist et al., 2016; Rose et al., 2016; Wolff et al., 2017/Chapter 4). While previous research has focused on visual WM, we now provide evidence for a neural impulse response that reflects parametric auditory WM content, suggesting a similar neural mechanism for auditory WM.

The neural response to a visual impulse stimulus also contained information about the behaviourally relevant pitch. It has been shown that visual stimulation can result in neural activity in the auditory cortex (Martuzzi et al., 2007; Morrill & Hasenstaub, 2018). Thus, direct connectivity between visual and auditory areas (Eckert et al., 2008) might be such that visual stimulation activates auditory WM representations in auditory cortex, providing an alternate access pathway. Alternatively, visual cortex itself might retain auditory information. It has previously been shown that natural sounds can be decoded from the activity in the visual cortex, during both processing and imagination (Vetter, Smith, & Muckli, 2014). Even though pure tones were used in the present study, it is nevertheless possible that they have been visualised, for example by imagining the pitch as a location in space. Tones may have also resulted in semantic representations, by categorising them into arbitrary sets of low, medium, and high tones. The decodable signal from the impulse-response might thus not necessarily originate from the sensory-processing areas, but rather from higher brain regions such as the prefrontal cortex (Stokes et al., 2013). Future studies that employ imaging tools with high spatial resolution might be able to arbitrate the neural origin of the cross-modal impulse response in WM.

While the neural impulse response to visual stimulus contained information about the relevant visual WM item, replicating previous results (Wolff et al., 2017/Chapter 4), the neural response to external auditory stimulation did not. This suggests that, in contrast to auditory information, visual information is maintained in a sensory-specific neural network with no evidence of content-specific connectivity with the auditory system, possibly reflecting the visual dominance of the human brain (Posner, Nissen, & Klein, 1976). Indeed, while it has been found that auditory stimulation results in neural activity in the visual cortex, it is notably weaker than the other way around (Martuzzi et al., 2007), which corresponds with our asymmetric findings of sensory specific and sensory non-specific impulse responses of visual and auditory WM between visual and auditory cortices.

It could be argued that the asymmetric findings reported here are the result of the asymmetry between the visual and auditory experiments; whereas the auditory impulse was the only non-visual stimulus in the visual task, the auditory task contained several non-auditory stimuli (cue, fixation cross, visual impulse). The auditory impulse may have thus been more easily ignored and filtered out in the visual task, causing a neural response that is too "weak" to interact with the neural WM network. However, we found no evidence for this alternative explanation. None of the early auditory ERPs, which have

been shown to be reduced by attentional filtering (e.g. Boutros et al., 2004; Kisley et al., 2004), were smaller in amplitude in the visual task compared to the auditory task. Indeed, the auditory P200 was larger in the visual task, the opposite direction as would be expected if the auditory impulse was more easily ignored in the visual task. Given the predictability and irrelevance of the impulse stimuli in both tasks (regardless of modality), we believe that the results reported here and in our previous research (Wolff et al., 2015/Chapter 3, 2017/Chapter 4) depend on low-level interactions of the bottom-up neural responses with the WM network, as proposed previously (Buonomano & Maass, 2009; Mongillo et al., 2008).

We found that both the processing and maintenance (as revealed by the impulse stimuli) of pure tones was coded parametrically according to the height of the pitch, similar to previous reports of parametric auditory WM (Spitzer & Blankenburg, 2012; Uluç et al., 2018). On the other hand, a neural code for pitch chroma, the cyclical similarity of the same notes across different octaves, was not found during either perception or maintenance. It has previously been found that complex tones (similar to musical instruments) may be more likely to result in a neural representation of pitch chroma than pure tones (as were used in this study) during perception (Briley, Breakey, & Krumbholz, 2013).

Visual orientations were clearly coded parametrically during encoding and maintenance, replicating previous findings (e.g. Saproo & Serences, 2010). Interestingly, we also found evidence for a neural coding scheme that reflects the specialization of orientations close to the cardinal axes (horizontal and vertical) compared to the oblique orientations during the encoding of orientations. This coding scheme is related to the previously reported "oblique effect" (higher discrimination and report accuracy of cardinal compared to oblique orientations; Appelle, 1972), and neural evidence for specialized neural structures in cat and macaque visual cortices for cardinal orientations (Li et al., 2003; Shen et al., 2014). The visual impulse response did not reveal such a coding scheme during maintenance, however, which could reflect a genuinely different coding scheme, but could also be due to the generally weaker orientation code during maintenance

It has previously been reported that the WM-related neural pattern evoked by the impulse response does not cross-generalize with the neural activity evoked by the memory stimulus itself (Wolff et al., 2015/Chapter 3), suggesting that the neural activation patterns are qualitatively different. In the present study, we also found no cross-generalization between item processing and the impulse response, neither in the visual nor in the auditory WM task. The neural representation of WM content may thus not be an exact copy of stimulation history, literally reflecting the activity pattern during information processing and encoding, but rather a reconfigured code that is optimized for future behavioural demands (Myers, Stokes, & Nobre, 2017). Similarly, no generalizability was found between auditory and visual impulse responses in the auditory task. This could suggest that distinct neural networks are perturbed by the different impulse modalities, or, as alluded to above, that it reflects the unique interaction between

impulses and the perturbed neural network. Future research should employ neural imaging tools with high spatial resolution to investigate the neural populations involved in the WM-dependent impulse-response.

The present results provide a novel approach to the ongoing debate on the extent to which sensory processing areas are essential for the maintenance of information in WM (Gayet et al., 2018; Scimeca et al., 2018; Xu, 2018). This is usually investigated by looking for the presence of WM-specific delay activity in the visual cortex in visual WM tasks (Bettencourt & Xu, 2016; Harrison & Tong, 2009), where null-results are interpreted as evidence against the involvement of specific brain regions, which is inherently problematic (Ester, Rademaker, & Sprague, 2016), and by which non-active WM states are not considered. In the present study, we found that sensory-specific stimulation, and both sensory specific and non-specific stimulation, resulted in WM-specific neural responses during the maintenance of visual and auditory information, respectively. Sensory cortices were thus linked to WM maintenance not by relying on ambient delay-activity, but rather by perturbing the underlying, connectivity-dependent, representational WM network via a bottom-up neural response.

Acknowledgments

This research was in part funded by a Biotechnology and Biological Sciences Research Council (BB/M010732/1) and James S. McDonnell Foundation Scholar Award (220020405) to MGS, and by the NIHR Oxford Health Biomedical Research Centre. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health. We would like to thank P. Albronda for providing technical support and M. Rietdijk for helping with data collection.

Chapter 6

Drifting codes within a stable coding scheme for working memory

A slightly revised version of this chapter is published as:

Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2020). Drifting codes within a stable coding scheme for working memory. *PLoS biology*, *18*(3), e3000625.

Data and code available at osf.io/cn8zf

Abstract

Working memory (WM) is important to maintain information over short time periods to provide some stability in a constantly changing environment. However, brain activity is inherently dynamic, raising an important challenge for maintaining stable mental states. To investigate the relationship between WM stability and neural dynamics, we used electroencephalography to measure the neural response to impulse stimuli during a WM delay. Multivariate pattern analysis revealed a clear difference in neural states between time-specific impulse responses, the coding scheme for memorized orientations was remarkably stable. This suggests that a stable subcomponent in WM that enables stable maintenance within a dynamic system. A stable coding scheme simplifies readout for WM-guided behaviour, whereas the low-dimensional dynamic component could provide additional temporal information. Despite this elegant coding scheme, WM is clearly not perfect – memory performance still degrades over time. Indeed, we find that even within the stable coding scheme, specific memories drift during maintenance. When averaged across trials, such drift contributes to the width of the error distribution, providing an alternative explanation for decreasing precision over time.

Introduction

Neural activity is highly dynamic, yet often we need to hold information in mind in a stable state to guide ongoing behaviour. Working memory is a core cognitive function that provides a stable platform for guiding behaviour according to time extended goals; however, it remains unclear how such stable cognitive states emerge from a dynamic neural system.

At one extreme, WM could effectively pause the inherent dynamics by falling into a stable attractor (e.g., Compte, Brunel, Goldman-Rakic, & Wang, 2000; Wang, 2001). This solution has been well-studied, and provides a simple readout of memory content irrespective of time (i.e., memory delay). However, more dynamic models have also been suggested. For example, in a recent hybrid model, stable attractor dynamic coexist with a low-dimensional, time varying component (Bouchacourt & Buschman, 2019; J. D. Murray et al., 2017); see Fig. 6.1A for model schematics). This permits some dynamic activity, whilst also maintaining a fixed coding relationship of WM content over time (Spaak, Watanabe, Funahashi, & Stokes, 2017). As in the original stable attractor model, the coding scheme is stable over time, permitting easy and unambiguous WM read out by downstream systems, regardless of maintenance duration (Cueva et al., 2019). Finally, it is also possible to maintain stable information in a richer dynamical system (e.g., Barak, Sussillo, Romo, Tsodyks, & Abbott, 2013). Although the relationship between activity pattern and memory content changes over time, the representational geometry could remain relatively constant (Spaak et al., 2017). Such dynamics emerge naturally in a recurrent network, and provide rich information about the previous input, and elapsed time (Romo, Brody, Hernández, & Lemus, 1999), but necessarily entail a more complex readout strategy (time-specific decoders or a high-dimensional classifier that finds a high-dimensional hyperplane that separates memory condition for all time points - (Druckmann & Chklovskii, 2012)).

Although all models seek to account for stable WM representation, it is also important to note that maintenance in WM is far from perfect. In particular, WM performance decreases over time. (Rademaker et al., 2018), which could be ascribed to two different mechanisms (Fig. 6.1B). On the one hand, the neural representation could simply degrade over time, either due to an overall decrease in WM specific neural activity, or through a general broadening of the neural representation (Barrouillet & Camos, 2001). In this framework, the distribution of recall error reflects sampling from a broad underlying distribution. On the other hand, the neural representation of WM content might gradually drift along the feature dimension as a result of the accumulating effect of random shifts due to noise (Kinchla & Smyzer, 1967). Even if the underlying neural representation remains sharp, variance in the mean over trials results in a relative broad distribution of errors over trials.

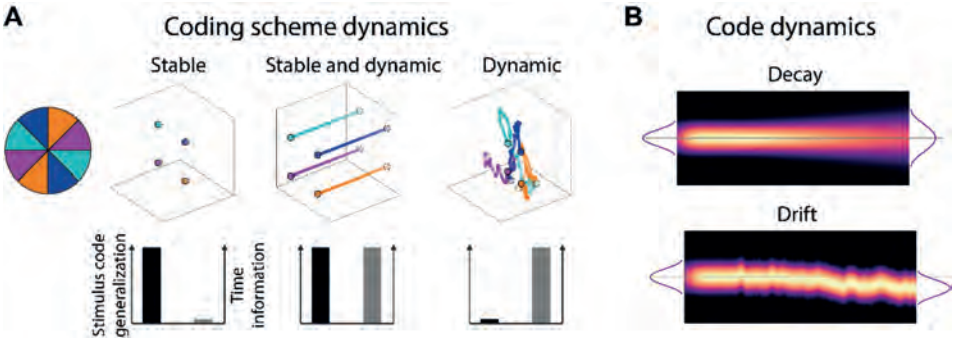


Figure 6.1. Model predictions. **(A)** The relationship between the neural coding scheme of orientations (colours) in WM over time. Left: A stable coding scheme within a stable neural population. Middle: A stable coding scheme within a dynamic neural population. Right: A dynamically changing coding scheme. **(B)** The fidelity of the population code in WM over time. Top: The code decays and becomes less specific over time, leading to random errors during read-out. Bottom: The code drifts along the feature dimension, leading to a still sharp, but shifted code during read-out.

Computational modelling based on behavioural recall errors from WM tasks with varying set-sizes and maintenance periods predict a drift for colours and orientations maintained in WM (Panichello, DePasquale, Pillow, & Buschman, in press; Schneegans & Bays, 2018). At the neural level, evidence for drift has been found in the neural population code in monkey prefrontal cortex during a spatial WM task (Wimmer et al., 2014), where trial-wise shifts in the neural tuning profile predicted if recall error was clockwise or counter-clockwise relative to the correct location. Recently, a human fMRI study has found that delay activity reflected the probe stimulus more when participants erroneously concluded that it matched the memory item (Lim, Ward, Vickery, & Johnson, 2019), which is consistent with the drift account.

Tracking these neural dynamics of non-spatial neural representations, which are not related to spatial attention or motor planning, is not trivial in humans. Previously we found that the presentation of a simple impulse stimulus (task-relevant visual input) presented during the maintenance period of visual information in WM results in a neural response that reflects non-spatial WM content (Wolff et al., 2015, 2017). Here we extend this approach to track WM dynamics. In the current study we developed a paradigm to test the stability (and/or dynamics) of WM neural states and the consequence for readout by "pinging" the neural representation of orientations at specific time-points during maintenance.

We found that the coding scheme remained stable during the maintenance period, even-though maintenance time was coded in an additional low-dimensional axis. We furthermore found that the neural representation of orientations drifts in WM. This was

reflected in a shift of the reconstructed orientation towards the end of the maintenance period that predicted behaviour.

Methods

Participants

Twenty-six healthy adults (17 female, mean age 25.8 years, range 20-42 years) were included in all analyses. Four additional participants were excluded during preprocessing due to excessive eye-movements (more than 30% of trials contaminated). Participants received monetary compensation (£10 an hour) for participation and gave written informed consent. The experiment was approved by the Central University Research Ethics Committee of the University of Oxford.

Apparatus and stimuli

The experimental stimuli were generated and controlled by Psychtoolbox (Kleiner, 2010), a freely available Matlab extension. Visual stimuli were presented on a 23-inch (58.42 cm) screen running at 100 Hz and a resolution of 1,920 by 1,080. Viewing distance was set at 64 cm. A Microsoft Xbox 360 controller was used for response input by the participants.

A grey background (RGB = 128, 128, 128; 20.5 cd/m^2) was maintained throughout the experiment. A black fixation dot with a white outline (0.242°) was presented in the centre of the screen throughout all trials. Memory items and the probe were sine-wave gratings presented at 20% contrast, with a diameter of 8.51° and spatial frequency of 0.65 cycles per degree, with randomised phase within and across trials. Memory items were presented at 6.08° eccentricity. The rotation of memory items and probe were randomized individually for each trial. The impulse stimulus was a single white circle, with a diameter of 20.67°, presented at the centre of the screen. The retro-cue was two arrowheads pointing right (>>) or left (<<), and was 1.58° wide. A coloured circle (3.4°) was used for feedback. Its colour depended dynamically on the precision of recall, ranging from red (more than 90 degrees error) to green (0 degrees error). A pure tone also provided feedback on recall accuracy after each response, ranging from 200 Hz (more than 90 degrees error) to 1,100 Hz (0 degrees error).

Procedure

Participants participated in a free-recall, retro-cue visual WM task. Each trial began with the fixation dot. After 1,000 ms the memory array was presented for 200 ms. After a 400 ms delay, the retro-cue was presented for 100 ms, indicating which of the previously two items would be tested, rendering the other item irrelevant. The first impulse stimulus was presented for 100 ms, 900 ms after the offset of the retro-cue. After a delay of 700 ms, the second impulse stimulus was presented for 100 ms. After another delay of 700 ms the probe was presented. Participants used the left joystick on the controller with the

left thumb to rotate the orientation of the probe until it best reflected the memorized orientation, and confirmed their answer by pressing the "x" button on the controller with the right thumb. Note that one complete rotation of the joystick corresponded to 0.58 of a rotation of the probe. In conjunction with the fact that the probe was randomly orientated on each trial, it was impossible for participants to plan the rotation beforehand or memorize the direction of the joystick instead of the orientation of the memory item. Accuracy feedback was given immediately after the response where both the coloured circle and tone were presented simultaneously. Each participant completed 1,100 trials in total, over a course of approximately 135 minutes, including breaks. See Figure 6.2A for a trial schematic.

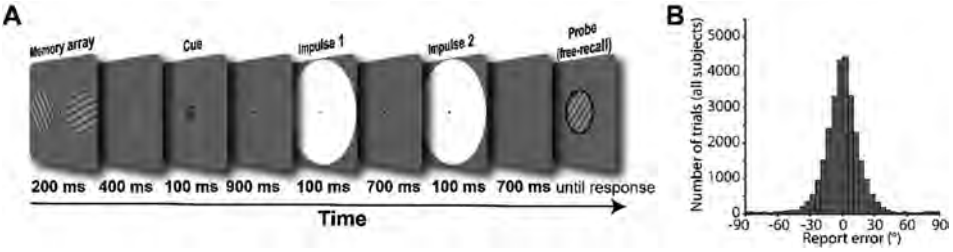


Figure 6.2. Trial schematic and behavioural results (A) Two randomly orientated grating stimuli were presented laterally. A retro-cue then indicated which of those two would be tested at the end of the trial. Two impulses (white circles) were serially presented in the subsequent delay period. At the end of the trial a randomly oriented probe grating was presented in the centre of the screen, and participants were instructed to rotate this probe until it reflected the cued orientation. (B) Report errors of all trials across all subjects.

EEG acquisition

EEG was acquired with 61 Ag/AgCl sintered electrodes (EasyCap, Herrsching, Germany) laid out according to the extended international 10–20 system and recorded at 1,000 Hz using Curry 7 software (Compumedics NeuroScan, Charlotte, NC). The anterior midline frontal electrodes (AFz) was used as the ground. Bipolar electrooculography (EOG) was recorded from electrodes placed above and below the right eye and the temples. The impedances were kept below 5 kΩ. The EEG was referenced to the right mastoid during acquisition.

EEG preprocessing

Offline, the EEG signal was re-referenced to the average of both mastoids, down-sampled to 500 Hz, and bandpass filtered (0.1 Hz high-pass and 40 Hz low-pass) using

EEGLAB (Delorme & Makeig, 2004). The continuous data was epoched relative to the memory array onset (-500 ms to 3,600 ms) before independent component analysis (Hyvarinen, 1999) was applied. Components related to eye-blinks were subsequently removed. The data was then epoched relative to memory array onset and the two impulse onsets (0 ms to 400 ms), and trials were individually inspected. Trials with saccadic eye movements, visually identified from the electrooculography, and trials with non-archetypical artefacts, visually identified from the EEG, in the memory array epoch and in either impulse epoch were removed from all subsequent analyses. Furthermore, trials where the report error was 3 circular standard deviations from the participant's mean response error were also excluded from EEG analyses to remove trials that likely represent complete guesses (Fritsche, Mostert, & de Lange, 2017). This lead to the removal of M = 2.3% (SD = 1.2%) trials due to inaccurate report trials, in addition to the M = 3.52 % (SD = 4.21%) and M = 5% (SD = 5.2%) of trials removed due to eye-movements and non-archetypical EEG artefacts from the memory array and impulse epochs, respectively.

While MVPA on electrophysiological data is usually performed on each time-point separately, taking advantage of the highly dynamic waveform of evoked responses in EEG by pooling information multivariately over electrodes as well as time can improve decoding accuracy, at the expense of temporal resolution (Grootswagers et al., 2017; Nemrodov et al., 2018). Since the previously reported WM-dependent impulse response reflects the interaction of the WM state at the time of stimulation and does not reflect continuous delay activity, we treat the impulse responses as discrete events in the current study. Thus, the whole time window of interest relative to impulse onsets (100 to 400 ms) from the 17 posterior channels was included in the analysis. The time window was based on previous, time-resolved findings, which showed that the WM-dependent neural response from a 100 ms impulse (as used in the current study) is largely confined to this window (Wolff et al., 2017). In the current study, instead of decoding at each time-point separately, information was pooled across the whole time-window. The mean activity level within each time window of each channel was first removed, thus normalizing the voltage fluctuations and isolating the dynamic, impulse-evoked neural signal from more stable brain states. The time-window was then down-sampled by taking the average every 10 ms, thus resulting in 50 values per channel, each of which was treated as a separate dimension in the subsequent multivariate analysis (850 in total). This data format was used on all subsequent MVPA analyses, unless explicitly mentioned otherwise. The same approach over the same time window of interest was used in our previous study (Wolff, Kandemir, Stokes, & Akyurek, 2019).

Orientation reconstruction

We computed the mahalanobis distances as a function of orientation difference to reconstruct grating orientations (Wolff et al., 2017). The following procedure was performed separately for items that were presented on the left and right side. Since the grating orientations were determined randomly on a trial-by-trial basis and the resulting

orientation distribution across trials was unbalanced, we used a k-fold procedure with subsampling to ensure unbiased decoding. Trials were first assigned the closest of 16 orientations (variable, see below) which were then randomly split into 8 folds using stratified sampling. Using cross-validation, the train trials in 7 folds were used to compute the covariance matrix using a shrinkage estimator (Ledoit & Wolf, 2004). The number of trials of each orientation bin were equalized by randomly subsampling the minimum number of trials in any bin. The subsampled trials of each angle bin were then averaged. To pool information across similar orientations, the average bins were convolved with a half cosine basis set raised to the 15th power (Brouwer & Heeger, 2009; Myers et al., 2015; Serences & Saproo, 2012). The mahalanobis distances between each trial of the left-out test fold and the averaged and basis-weighted angle bins were computed and mean-centred across the 16 distances to normalize. This was repeated for all test and train fold combinations. To get reliable estimates, the above procedure was repeated 100 times (random folds and subsamples each time), separately for eight orientation spaces (0° to 168.75°, 1.40625° to 170.1563°, 2.8125° to 171.5625°, 4.2188° to 172.9688°, 5.625° to 174.375°, 7.0313° to 175.7813°, 8.4375° to 177.1875°, 9.8438° to 178.5938°, each in steps of 11.25°). For each trial we thus obtained 800 samples for each of the 16 mahalanobis distances. The distances were averaged across the samples of each trial and ordered as a function of orientation difference. The resulting "tuning curve" was summarized into a single value (i.e., "decoding accuracy") by computing the cosine vector mean of the tuning curve (Wolff et al., 2017), where a positive value suggests a higher pattern similarity between similar orientations than between dissimilar orientations. The approach was the same for the reanalysis of (Wolff et al., 2015).

We also repeated the above analysis iteratively for a subset of electrodes in a searchlight analysis across all 61 electrodes. In each iteration, the "current" as well as the closest two neighbouring electrodes were included in the analysis (similar as in Ede, Chekroud, Stokes, & Nobre, 2019) The freely available MATLAB extension fieldtrip (Oostenveld et al., 2010) was used to visualise the decoding topographies. Note that the topographies were flipped, such that the left represents the ipsilateral and the right the contralateral side relative to stimulus presentation side.

Orientation code generalization

To test cross-generalization between impulses, instead of training and testing within the same time-window, the train folds were taken from impulse 1, and the test fold from impulse 2, and vice versa. The analysis was otherwise exactly as described above.

To test cross-generalization between presented locations, the classifier was similarly trained on trials where the item was presented on the left, and tested on the right, and vice versa. Since left and right trials were independent trial sets, cross-validation does not apply. However, to ensure a balanced training set, the number of trials of each orientation bin were nevertheless equalized by subsampling (as described above), and this approach was repeated 100 times.

The cross-generalization of the orientation code between impulse onsets in (Wolff et al., 2015) was tested with the same analyses as the location cross-generalization described in the paragraph above: The classifier was trained on the early onset condition, and tested on the late-onset condition, and vice versa, while making sure that the training set is balanced through random subsampling.

Impulse/time and location decoding

To decode the difference of the evoked neural responses between impulses, we used a leave-one-out approach. The mahalanobis distances between the signals from a single trial from both impulse epochs and the average signal of all other trials of each impulse epoch were computed. The covariance matrix was computed by concatenating the trials of each impulse (excluding the left-out trial). The average difference of same impulse distances were subsequently subtracted from different impulse distances, such that a positive distance difference indicates more similarity between same than different impulses. To convert the distance difference into trial wise decoding accuracy, positive distance difference were simply converted into "hits" (1) and negative into "misses" (0). The percentage of correctly classified impulses were subsequently compared to chance performance (50%).

The presentation side and impulse onset (in Wolff et al., 2015) was decoded using 8-fold cross-validation, where the distance difference between different and same location/onset was computed for each trial, which were then converted to "hits" and "misses".

Visualization of the spatial, temporal, and orientation code

To explore and visualize the relationship between the location or impulse/time code and the orientation code in state space (see Fig. 6.1A for different predictions), we used classical multidimensional scaling (MDS) of the mahalanobis distances between the average signal of trials belonging to one of four orientation bins (0° to 45°, 45° to 90°, 90° to 135°, 135° to 180°) and location (left/time) or time (impulse 1/impulse2).

For the visualization of the code across impulse/time, distances were computed separately for left and right trials, before taking the average. Within each orientation bin, the data of half of the trials were taken from impulse 1, and the data of the other half from impulse 2 (determined randomly). The number of trials within each orientation of each impulse were equalized through random subsampling before averaging. The mahalanobis distances between both orientation and impulses were then computed using the covariance matrix estimated from all trials of both impulses. This was repeated 50 times (for each side), randomly subsampling and splitting trials between impulses each time and then taking the average across all iterations.

For the visualization of the code across space, the data of each trial were first averaged across impulses. The number of trials of orientation bins (same as above) of each location were equalized through random subsampling. The mahalanobis distances

of the average of each bin within each location condition were computed using covariance estimated from all left and right trials. This was repeated 50 times, before taking the average across all iterations.

For the code across impulse onset/time visualization of the data from Wolff et al. (2015), the same procedure as in the paragraph above was used, but instead of visualizing the stimulus code between locations, it was visualized between impulse onsets (-30 ms, +30 ms).

Relationship between behaviour and the neural representation of the WM item

We were interested if imprecise reports that are too clockwise (cw) or counter-clockwise (ccw) relative to the actual orientation are accompanied by a corresponding shift of the neural representation in WM (see Fig. 6.1B for model schematics). We used two approaches to test for such a shift (Fig. 6.5A & 6.6A).

First, the trial-wise pattern similarities as a function of orientation differences (as obtained from the orientation-reconstruction approach described above) were averaged separately for all cw and ccw responses (Fig. 6.5A). Note that cw and ccw responses were defined relative to the median response error within each orientation bin. This ensures a balanced proportion of all orientations in cw and ccw trials, which is necessary to obtain meaningful orientation reconstructions. It furthermore removes the report bias away from cardinal angles in the current experiment (Suppl. fig. 6.1), similar to previous reports of orientation response biases (Pratte et al., 2017), and thus isolates random from systematic report errors.

We used another approach that exaggerates the potential difference between cw and ccw trials and thus might be more sensitive to detect a shift. The data was first divided into cw and ccw trials using the same within orientation bin approach as described above. The classifier was then trained on cw trials, and tested on ccw trials, and vice versa (Fig. 6.6A). The orientation bins in the training set were balanced through random subsampling, and the procedure was repeated 50 times. Given an actual shift in the neural representation, the shift magnitude of the resulting orientation reconstruction of this method should be doubled, since both the testing data and the training data (the reference point) are shifted, but in opposite directions.

To improve orientation reconstruction from the impulse epochs, the classifier was trained on the averaged trials of both impulses, but tested separately on each impulse epoch individually. While training on both impulses improved orientation reconstruction, in particular for the second approach where only half of the trials are used for training, the shifts in orientation representations as a function of cw/ccw reports are qualitatively the same when training and testing within each impulse epoch separately (Fig. 6.5, 6.6, & Suppl. fig. 6.3).

Statistical significance testing

To test for statistical significance of decoding accuracies, the sign of the data of each participant was randomly flipped with a probability of 50% 100.000 times, and the resulting null-distribution was used to calculate the p value of the null hypothesis (no difference, chance decoding). Note that tests of within condition decoding (within presentation location, impulse/onset) were one-sided, since only positive decoding is plausible in those cases, whereas tests of cross-generalization between conditions were two-sided, since negative decoding is theoretically plausible in those cases. Comparisons of decodability between conditions/items were also two-sided.

The possible shift in representation towards the response was quantified and tested for statistical significance at the group level. The circular mean of the shifted average tuning curve (summarized such that a positive shift reflects a shift towards the response) was tested against 0. The tuning curve of each subject was flipped left to right with 0.5 probability, such that a subject's positively shifted tuning curve would then be negatively shifted, before computing the circular mean of the resulting tuning curve averaged over all subjects 100.000 times. The resulting null distribution was used to obtain the p-value by calculating the proportion of permuted tuning curves with circular means more positive than the actual group-level circular mean. The test obtained p-value was one-sided, since we expected the shift of the neural representation of the orientation to be towards the response.

Code and data availability

All data and custom Matlab scripts used to generate the results and figures of this manuscript will be made available upon peer-reviewed publication.

Results

Item and WM content-specific evoked responses during encoding and maintenance

The neural response elicited by the memory array contained parametric information about the presented orientations ($p < 0.001$, one-sided; Fig. 6.3, left).

The first impulse response contained statistically significant information about the cued item ($p = 0.008$, one sided), but not the uncued item, which failed to reach the statistical significance threshold ($p = 0.057$, one-sided). The difference between cued and uncued item decoding was not significant ($p = 0.694$, two-sided; Fig. 6.3, middle).

The decodability of the cued item was also significant at the second impulse response ($p < 0.001$, one-sided), while it was not of the uncued item ($p = 0.919$, one-sided). Notably, the decodability of the cued item was significantly higher than that of the uncued item ($p = 0.002$, two-sided; Fig. 6.3, right).

Overall, these results reflect previous findings (Wolff et al., 2017) in that the impulse response reflects relevant information in WM, and that no longer relevant information leave no detectable trace in the WM network.

The decoding topographies highlight that most of the decodable signal came from posterior electrodes during both encoding and maintenance, and is therefore likely generated by the visual cortex. Notably, while contralateral electrodes showed unsurprisingly higher item decoding during encoding, this was not the case during maintenance in either impulse response (Fig. 6.3 bottom row).

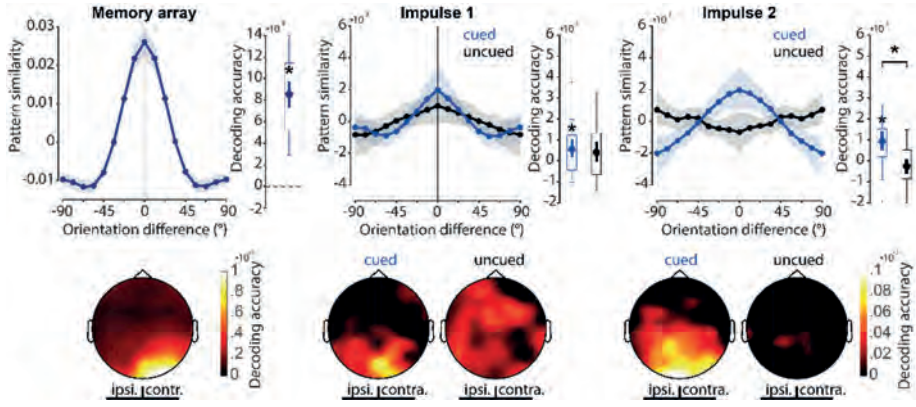


Figure 6.3. Decoding results. Top row: Normalized average pattern similarity (mean-centred, sign-reversed mahalanobis distance) of the evoked neural responses (100 to 400 ms relative to stimulus onset) as a function of orientation similarity, and decoding accuracy (cosine vector means of pattern similarities). Error shadings and error bars are 95 % C.I. of the mean. Centre lines of boxplots indicate the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range. Extreme values are shown separately (dots). Asterisks indicate significant decoding accuracies (p < 0.05, one-sided) or differences (p < 0.05, two-sided). Bottom row: Decoding topographies of the searchlight analysis.

Stable WM coding scheme in time

The relationship between orientations and impulses/time is visualized in state-space through MDS (Fig. 6.4A). While the first dimension clearly differentiates between impulses, the second and third dimensions code the circular geometry of orientations in both impulses, suggesting that while the impulse responses are different between impulses, the orientation coding schemes revealed by the impulse are the same. This is

corroborated by significant decoding accuracy of the impulse ($p < 0.001$, one-sided; Fig. 6.4B) on the one hand, but also significant cross-generalization of the orientation code between impulses ($p < 0.001$, two-sided), which was not significantly different from same-impulse orientation decoding ($p = 0.581$, two-sided; Fig. 6.4C).

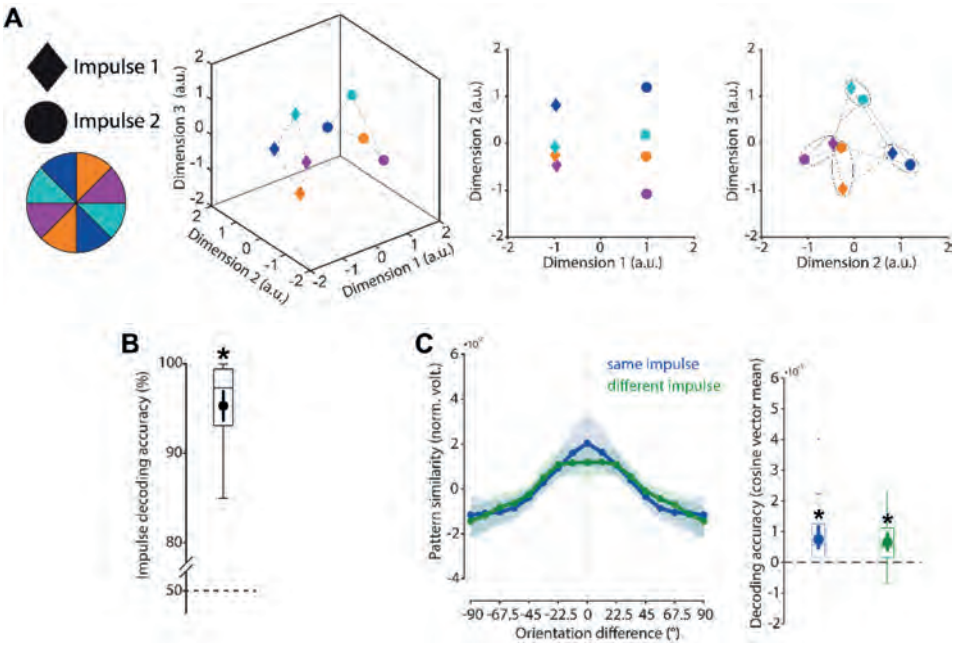


Figure 6.4. Cross-generalization of coding scheme between impulses. **(A)** Visualization of orientation and impulse code in state-space. The first dimension discriminates between impulses. The second and third dimensions code the orientation space in both impulses. **(B)** Trial-wise accuracy (%) of impulse decoding. **(C)** Orientation decoding within each impulse (blue) and orientation code cross-generalization between impulses (green). Error shadings and error bars are 95 % C.I. of the mean. Centre lines of boxplots indicate the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range. Extreme values are shown separately (dots). Asterisks indicate significant decoding accuracies or cross-generalization ($p < 0.05$).

It is not possible to conclude whether the difference between impulses is due to a neural network that changes during the maintenance period over time, due to different stimulation histories at the time of perturbation (i.e., the first impulse always preceded

the second impulse), or due to different WM operations at each impulse event (e.g. item selection at impulse 1, response preparation at impulse 2).

To rule out that the difference in impulse response reported above is not only due to difference in stimulation history and changing WM operations, but also due to temporal coding in the WM network, we reanalysed previously published data where a single impulse stimulus was presented either 1,170 or 1,230 ms after the presentation of a single memory item (Wolff et al., 2015). The findings largely replicate the results reported above: State-space visualization of impulse-onset and orientations shows the same circular geometry of the orientations at each impulse onset, while also highlighting a separation of impulse onsets in state-space (Suppl. fig. 6.2A). Decoding impulse-onset was significantly than from chance ($p = 0.005$, one-sided; Suppl. fig. 6.2B). Cross-generalization of the orientation code between impulse-onsets was significant ($p < 0.001$, two-sided), and did not significantly differ from decoding the memorized orientation within the same impulse-onset ($p = 0.244$, two-sided; Suppl. fig. 6.2C).

Overall, the results of the current study, as well as the reanalyses of (Wolff et al., 2015) provide evidence for a low-dimensional change over time, that can be revealed by perturbing the WM network at different time-points (as predicted in (Buonomano & Maass, 2009)), while at the same time providing evidence for a temporally stable coding scheme of WM content (Bouchacourt & Buschman, 2019; J. D. Murray et al., 2017).

Specific WM coding scheme in space

As a counterpart to the stable coding scheme in time reported above, we explicitly tested if the coding scheme is location specific (i.e., dependent on the previous presentation location of the cued orientation). State-space visualization of cued item location and orientations shows a clear separation between locations and no overlap in orientation coding between locations (Fig. 6.5A). The cued location was significantly decodable from the impulse responses ($p < 0.001$, one-sided; Fig. 6.5B). Cross-generalization of the orientation coding scheme between cued item locations was not significant ($p = 0.403$, two-sided), and significantly lower than same side orientation decoding ($p = 0.009$, two-sided; Fig. 6.5C). These results reflect previous reports of spatially specific WM codes, even when location is no longer relevant (Pratte & Tong, 2014).

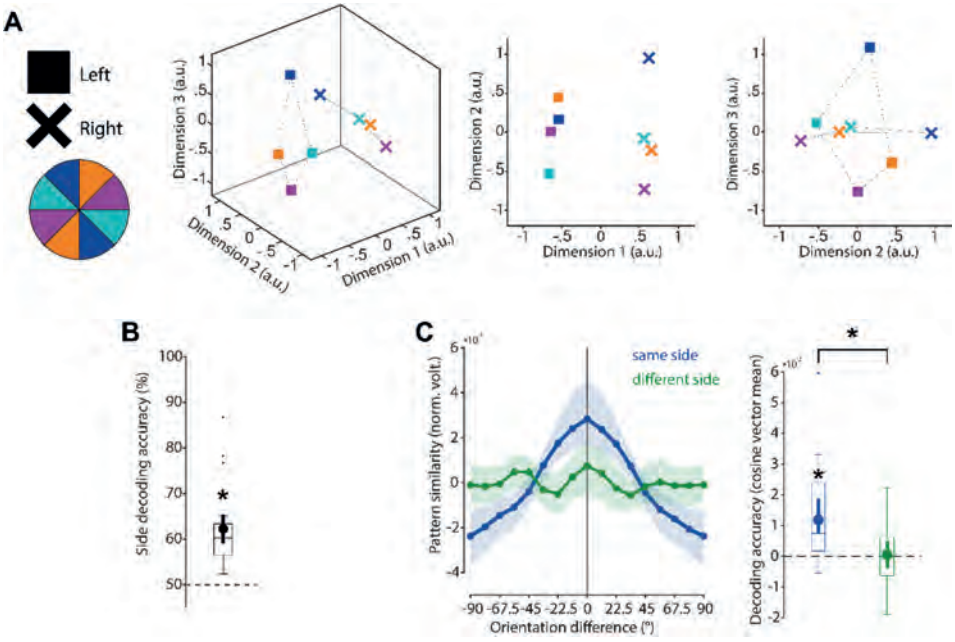


Figure 6.5. No cross-generalization of coding scheme between cued item locations during impulse responses **(A)** Visualization of orientation and item location code in state-space. The first dimension discriminates between item locations. The first and second dimensions code the orientation space, separately for WM items previously presented on the left or right side. **(B)** Trial-wise accuracy (%) of item location decoding. **(C)** Orientation decoding within each item location (blue) and orientation code cross-generalizing between different item locations (green). Error shadings and error bars are 95 % C.I. of the mean. Centre lines of boxplots indicate the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range. Extreme values are shown separately (dots). Asterisks indicate significant decoding accuracies and differences ($p < 0.05$).

Drifting WM code

The first approach to test for a possible shift of the neural representation towards the response averaged the trial-wise orientation tuning curves obtained from the cross-validated orientation reconstruction on all trials (see Methods and Fig. 6.6A).

No significant shift towards the response was evident during encoding/memory array presentation ($p = 0.117$, one-sided; Fig. 6.6B & C, left). No evidence for such a shift was found at impulse 1/early maintenance either ($p = 0.07$, one-sided; Fig. 6.6B & C, middle). However, the orientation tuning curve was significantly shifted towards the response at impulse 2/late maintenance ($p < 0.001$, one-sided; Fig. 6.6B & C, right).

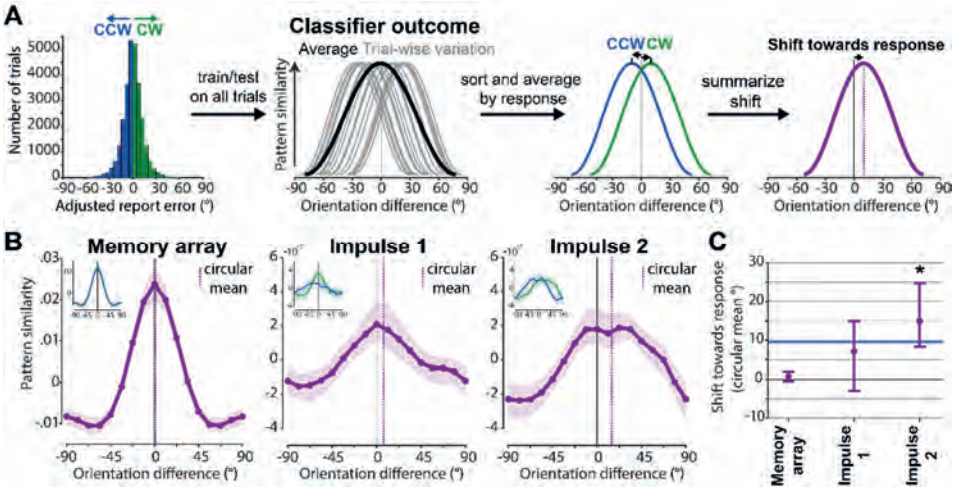


Figure 6.6. Response-dependent averaging of trial-wise tuning curves demonstrates drift. Schematic and results. **(A)** Testing for shift towards response by averaging trial-wise tuning curves by ccw/cw responses. **(B)** Results of schematised approach in A. Orientation tuning curves averaged by response such that a right-ward shift reflects a shift towards the response (purple) at each event. Purple vertical lines show circular means of the tuning curves. Insets show orientation tuning curves for ccw (blue) and cw (green) responses separately. Error shadings are 95 % C. I. of the mean. **(C)** Group-level shifts towards the response (circular mean) of each response-dependent tuning curve. Error-bars are 95 % C. I. of the mean.

The second approach to test for a possible shift of the neural representation towards the response may be more sensitive since it trains the orientation classifier only on ccw trials, and tests it on cw trials, and vice versa (see Methods and Fig. 6.7A), thus exaggerating any response related shift by a factor of two.

This approach yielded similar results as the previous approach, though the shift magnitudes are indeed larger. Neither the memory array presentation/encoding, nor impulse 1/early maintenance showed a significant shift towards the response ($p = 0.121$, $p = 0.104$, respectively, one-sided; Fig. 6.7, left & middle), while impulse 2/late maintenance did ($p < 0.001$, one-sided; Fig. 6.7, right).

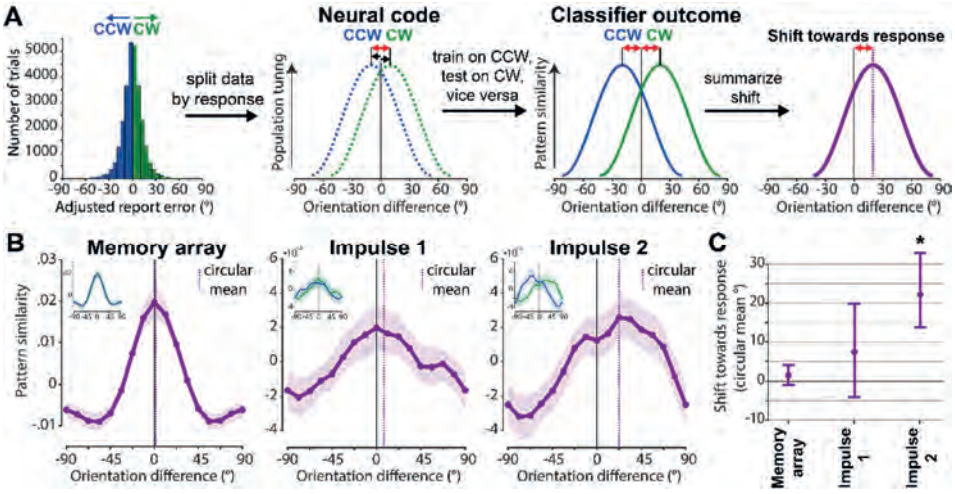


Figure 6.7. Response-dependent training and testing demonstrates drift. Schematic and results. **(A)** Testing for shift towards response by first splitting the neuroimaging data into cw and ccw data sets, and training on cw trials and testing on ccw trials, and vice versa. Given an actual shift, the shift of the resulting orientation reconstruction will be doubled, since training and testing data are shifted in opposite directions. **(B)** Results of schematised approach in A. Average orientation tuning curves such that a right-ward shift reflects a shift towards the response (purple) at each event. Purple vertical lines show circular means of the tuning curves. Insets show orientation tuning curves for ccw (blue) and cw (green) responses separately. Error shadings are 95 % C. I. of the mean. **(C)** Group-level shifts towards the response (circular mean) of each response-dependent tuning curve. Error-bars are 95 % C. I. of the mean.

Note the reported results of shifts during impulse presentations were obtained by training the classifier on both impulses, but testing it on each impulse separately. This was done to improve power (as explained in Methods). This improved orientation reconstruction particularly for the latter shift-analysis where the classifier is trained on only half the trials (cw trials only or ccw trials only). However, the same analyses based on training (and testing) within each impulse epoch separately yielded qualitatively similar results (no significant shifts at impulse 1 in either approach, significant shifts at impulse 2 in both approaches; Suppl. fig. 6.3).

Discussion

In the present study, we investigated the neural dynamics of WM maintenance by probing both the coding scheme of WM content, as well as the code itself at specific time-points during the maintenance period. The neural impulses to external visual stimulation enabled us to show that the coding scheme of orientations in WM remained

stable during maintenance, while the orientation code itself drifted and predicted behaviour.

The significant cross-generalization of the orientation-coding scheme between impulses presented at different time-points (Fig. 6.4) is consistent with previous reports of stable WM maintenance (Murray et al., 2017; Spaak et al., 2017), providing evidence for a time-invariant coding scheme for orientations maintained in WM. However, non-stable schemes have also been reported. It has been found that attentional (de-) prioritization of WM content changes its format (van Loon, Olmos-Solis, Fahrenfort, & Olivers, 2018), and that coding schemes morph after the presentation of interfering distractors (Parthasarathy et al., 2017). It seems that the maintenance of simple features is relatively stable (Murray et al., 2017; Spaak et al., 2017), as was the case in the present study, whereas complex objects and more demanding WM tasks are more likely to elicit dynamically changing coding schemes (Meyers, 2018; Stokes et al., 2013). Given the fact that the first impulse always preceded the second in the present study, our results suggest that neither time, nor the passive viewing of irrelevant stimuli change the coding scheme of an orientation maintained in WM.

While there was no cost of cross-generalizing the orientation code between impulses, there was nevertheless a clear difference in the neural pattern between them (Fig. 6.4), suggesting that a separate, dynamic neural pattern codes the passage of time that is orthogonal to the stable WM content-code (Murray et al., 2017). However, the large neural difference between impulses (~95% decoding accuracy) in the present study can not only be attributed to the passage of time, but also to changes to the network by the first impulse, which preceded the second impulse in every trial. Additionally, the possibly still ongoing dropping of the irrelevant orientation at impulse 1 presentation and response preparation at impulse 2 towards the end of the maintenance period were likely contributors to this difference. We therefore reanalysed the data of a previously published study where the impulse onset was randomly jittered by +-30 ms at 1,000 ms after the offset of a single memory item (Wolff et al., 2015). The significant decodability of impulse onset shows that the WM network changes during the maintenance even within 60 ms, resulting in distinct neural impulse responses at different time-points providing evidence for a neural time-code (Suppl. fig. 6.2).

The stable WM-content coding scheme could be achieved by low-level activity states that self-sustain a stable code through recurrent connections, a key feature of attractor models of WM (Chaudhuri & Fiete, 2016; Compte et al., 2000), while dynamic activity patterns are coded in an orthogonal subspace that represents time (Cueva et al., 2019; Murray et al., 2017; Tiganj, Cromer, Roy, Miller, & Howard, 2018). Though we did not explicitly consider persistent delay activity, the dynamic impulse responses analysed in the present study could reflect non-linear interactions with low-level, persistent activity states that are otherwise difficult to measure with EEG. Alternatively, silent WM states that do not depend on persistent activity to maintain WM content (Mongillo et al., 2008; Stokes, 2015), which could be revealed by the impulse responses (Wolff et al., 2017), might be also be a plausible mechanism. Here, the activity state during encoding

leaves behind a neural trace in the WM network through short-term synaptic plasticity resulting in a stable code for maintenance, whereas the time-dimension could be represented in its gradual fading (Buonomano & Maass, 2009; Nikolić, Häusler, Singer, & Maass, 2009; Nikolić et al., 2007; Zucker & Regehr, 2002).

We also found evidence that the orientation code itself drifts along the orientation dimension, predicting recall errors (Fig. 6.6, Fig. 6.7, Suppl. fig. 6.3). While there was no shift in the neural orientation representation at either encoding or early maintenance, the second impulse towards the end of the maintenance period revealed a code that was shifted towards the direction of response error. This pattern of results is consistent with the drift account of WM, where the encoding of information into WM might be perfect, but neural noise leads to an accumulation of error during maintenance, resulting in a still sharp, but shifted (i.e., slightly wrong) neural representation of the maintained information (Compte et al., 2000; Schneegans & Bays, 2018). While previous neurophysiological recordings from monkey PFC found evidence for drift for spatial information (Wimmer et al., 2014), by using lateralized orientations in the present study, we could demonstrate a shifting representation that more faithfully represents actual, non-spatial WM content that is unrelated to sustained spatial attention or motor preparation.

Bump attractors have been proposed as an ideal neural mechanism for the maintenance of continuous representations (i.e. space, orientation, colour), where a specific feature is represented by the persistent activity "bump" of the neural population at the feature's location along the network's continuous feature space. Neural noise randomly shifts this bump along the feature dimension, while inhibitory and excitatory connections maintain the same overall level of activity and shape of the neural network (Amari, 1977; Brody, Romo, & Kepecs, 2003). Random walk along the feature dimension is thus a fundamental property of bump attractors, and has been found to explain neurophysiological findings (Wimmer et al., 2014).

The drift and resulting error do not necessarily have to be random, however. Modelling of report errors in a free recall colour WM task suggests that an increase of report errors over time may be due to separable attractor dynamics, with a systematic drift towards stable colour representations, resulting in a clustering of reports around specific colour values, in addition to random drift elicited by neural noise (Panichello et al., in press). The report bias of oblique orientations seen in the present study could be explained by a similar drift towards specific orientations, which would predict an increase of report bias for longer retention periods. However, clear behavioural evidence for such an increase in systemic report errors of orientations is lacking (Rademaker et al., 2018). In the present study we isolated random from systematic errors, both as a methodological necessity, but also to be able to conclude that any observed shift is due to random errors. Thus, while a systematic drift towards specific orientations might be possible, the shift in representation reported here is unrelated to it.

Serial dependence, the systematic attraction of remembered features towards previously presented feature values (Fischer & Whitney, 2014), may be another non-

random contributor to drift in WM. It seems to be a robust phenomenon (Kiyonaga, Scimeca, Bliss, & Whitney, 2017), and was also present in the present study (not shown). While we did not explicitly control for it in our analyses, we did not find evidence that the observed shift at the second impulse was driven by serial dependence (not shown). However, given the previous report of increase of serial dependence magnitude for longer WM maintenance periods (Bliss, Sun, & D'Esposito, 2017), it is nevertheless plausible that current WM representations may drift towards previous perception, but that its magnitude is too small to be detectable with EEG.

How can a drifting WM code be reconciled by "silent" WM accounts (Miller et al., 2018; Stokes, 2015), where the WM-content specific code is maintained in transient connectivity changes in the WM network that decay over time and are periodically refreshed by short bursts of activity? While it is not theoretically impossible that the state-dependent neural impulse response of such a decaying synaptic WM network results in a shifted WM-code (as observed in the present study), a more intuitive prediction of a decaying WM network would be a broader WM code, that does not predict the direction of report error (Barrouillet & Camos, 2001). Conversely, while the temporary connectivity changes of the memorized WM item may indeed slowly dissolve and become coarser, periodic activity bursts (Lundqvist et al., 2016) may keep this to a minimum, by periodically reinstating a sharp representation. However, since this refreshing depends on the read-out of a coarse representation, the resulting representation may be slightly wrong and thus shifted. This interplay between decaying silent WM-states that are readout and refreshed by active WM-states thus also predicts a drifting WM code, without depending on an unbroken chain of persistent neural activity.

Acknowledgments

This research was in part funded by a James S. McDonnell Foundation Scholar Award (220020405) and an ESRC grant (ES/S015477/1) to MGS, and by the NIHR Oxford Health Biomedical Research Centre. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health. EGA is in part funded by an Open Research Area grant (464.18.114). We would like to thank N.E. Myers and D. Trübutschek for helpful comments.

Chapter 7

General Discussion

The short-term maintenance and manipulation of information to guide behaviour without its perceptual counterpart in the environment is a fundamental ability of the (human) brain. Finding its neural correlate has been a major challenge in cognitive neuroscience ever since the first WM modulated neurons have been recorded in monkey PFC almost 50 years ago (Fuster & Alexander, 1971). Until now, the common approach to research the neural correlate of WM has been to look for neural activity that spans the maintenance period in WM tasks that in some way reflects WM task conditions or WM content. The assumption has been that persistent neural activity enables the short-term maintenance of information, given numerous confirmatory observations, but also due to an implicit necessity of neuroimaging: measurable neural activity is required for research that employs neuroimaging. Finding no neural activity of WM is essentially a null result and difficult to interpret.

More recently however, the idea that WM might not be solely reliant on an unbroken chain of neural activity has gained traction. Numerous studies have reported that unprioritized WM content elicits no or little measurable neural activity (Larocque et al., 2014; Lewis-Peacock et al., 2011; Watanabe & Funahashi, 2014), suggesting WM delay activity reflects the focus of attention within WM, but not the content per se. Furthermore, it has been suggested that persistent WM delay activity is an artefact of trial averaging and that maintenance is actually mediated by short-lived activity bursts interleaved by activity-silent periods (Lundqvist et al., 2016), which are bridged via short-term connectivity-changes (Mongillo et al., 2008; Stokes, 2015).

For this thesis, it was hypothesized that the neural response to external stimulation is an interaction between the input and the current state of the network, analogous to echolocation where the echo reflects the stimulation as well as the hidden structures. This thesis explicitly tested this hypothesis and exploited its functionality to study WM states. Across several experiments it is shown that instead of relying on measurable WM delay activity, the state-dependent impulse response to external stimulation can be used to infer (hidden) WM states by taking advantage of the spatially rich EEG signal.

Chapter-wise summary of main findings

Chapter 2 highlights the implication of a study demonstrating that MEG actually contains rich spatial information origination from orientation columns in V1 (Cichy et al., 2015). This implies that MEG is a powerful neuroimaging tools for multivariate-pattern analysis, an analysis that takes advantage of systematic activity patterns, and until relatively recently almost exclusively used in fMRI research. While the spatial ambiguity of MEG remains, the signals are nevertheless spatially rich in that even neural activity patterns originating from small brain areas result in unique activity patterns across the scalp. It has been shown that the same may apply to EEG, which is in an almost equally powerful neuroimaging tool for multivariate analyses as MEG (Cichy & Pantazis, 2017). All following experimental chapters take advantage of the rich EEG signal, and

demonstrate that even objects within the same category (i.e. orientations) can be decoded with high temporal resolution.

Chapter 3 not only demonstrates for the first time that EEG is sensitive enough to decode randomly orientated visual gratings but is also the first proof of principle of the impulse approach. Using time-point-by-time-point decoding analysis of the EEG signal from posterior channels, it is shown that the recorded EEG signal distinguishes between visually presented, randomly orientated gratings. Furthermore, the presentation of a neutral "impulse" stimulus about 1 second later evokes a neural signal that also distinguishes between orientations, providing evidence for the hypothesis that that the neural impulse response reflects that current state of the network, in this case the imprint of the previously presented grating.

It must be noted that these outcomes could not establish if the orientation-specific impulse response is indeed a reflection of WM maintenance, or a neural afterimage of stimulation history. Indeed, the original proposition suggests that any neural activity leaves behind a synaptic trace in the connectivity pattern (Buonomano & Maass, 2009) and makes no WM-specific predictions. This was based on findings in cat visual cortex where the presentation of visual information resulted in a neural activity that reflected both current and previous visual stimulation (Nikolic, Haeusler, Singer, & Maass, 2007). Since the cats used in that study were anaesthetised during neural recording, one can hardly speak of a WM-dependent neural response.

Chapter 4 further explores and extends the findings of chapter 3, and shows with a retro-cue design that the impulse response is in fact specific to WM content, and does not reflect stimulation history in general. This suggests that irrelevant information can be removed from the WM state, leaving no detectable trace in the impulse response. In an additional experiment, chapter 4 also shows that the impulse response reflects both attended and unattended WM content, providing evidence that the WM states codes all relevant information, dissociating it from attentional processes, and thus showing that the WM-dependent impulse is not dependent on measurable delay activity mediated by attention. In both experiments trial-wise decodability of the relevant WM item from the impulse response predicted the quality of behavioural recall.

Chapter 5 shows that the impulse approach, demonstrated in the previous chapters using visual stimuli, is also applicable in the auditory domain. Specifically, a neutral auditory stimulus presented during the delay period of an auditory WM tasks resulted in an auditory WM-dependent neural response. This indirectly implicates the auditory cortex in the maintenance of auditory tones in the same way the previously observed visual impulse response implicates the visual cortex, and providing evidence for the involvement of sensory processing areas in the maintenance of information in WM (Christophel et al., 2017; Kumar et al., 2016; Serences, 2016). It is furthermore shown that the cross-modal (i.e., auditory) impulse response during visual WM maintenance is content-unspecific, suggesting that the visual WM network is separated from auditory processing areas and that a bottom-up auditory response does not perturb it. However, the neural impulse response to visual stimulation during auditory WM maintenance is

content-specific, suggesting some involvement of the visual cortex during auditory maintenance, either directly (through visualization), or indirectly (through content-specific connectivity with the auditory WM network).

Chapter 6 uses the impulse approach as a tool to investigate the relationship between the neural representation of WM content and WM recall. More specifically, it provides neural imaging evidence to the proposal that the neural representations of continuous WM items (for example orientations or locations) randomly drift along the item representation over time (Schneegans & Bays, 2018). The impulse response revealed a shift of the neural representation of an orientation maintained in WM towards the end of the trial that predicted small behavioural variations during recall. Additionally it was found that the neural representation of orientations in WM remained stable across the delay period, as evidenced by robust cross-generalization between impulses presented at different time-points during the maintenance period, even though the neural responses to the impulses were highly different. This suggests that the coding scheme remains stable, despite time-varying neural dynamics (Murray et al., 2017).

Neural impulse-response is WM-dependent

The experimental chapters of this thesis (chapters 3 to 6) provide evidence that the neural impulse response to irrelevant, external sensory stimuli contains information about WM content. Chapter 3 provides evidence that this effect is time-locked to the onset of the impulse stimulus, and thus cannot be explained by the temporal expectation of the impulse, which could have resulted in a pre-activation of the WM item (Nobre et al., 2007; Stokes et al., 2009). In other words, the reported results suggest that it is indeed the neural response to exogenous stimulation, and not an endogenous reactivation of the WM network. Chapter 4 furthermore shows that this WM-dependent impulse response is not reliant on measurable WM-related delay activity, by demonstrating that even unattended WM content, previously shown not to be measurable with conventional neuroimaging techniques (LaRocque et al., 2012; Lewis-Peacock et al., 2011; Lewis-Peacock & Postle, 2012), is contained in the impulse response. This thesis thus provides evidence for a key prediction made by the connectivity-dependent WM account (Mongillo et al., 2008), namely that information in WM can be maintained in a hidden, activity-silent neural state (Stokes, 2015) that may be mediated by short-term synaptic plasticity (Zucker & Regehr, 2002), and which can be read out from the state-dependent impulse-response (Buonomano & Maass, 2009). By presenting a fixed, task-irrelevant impulse stimulus in the delay periods of WM tasks, this thesis confirms that the resulting neural impulse-response contains information about the current state of the WM network (Fig. 7.1).

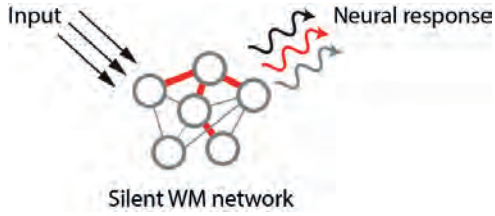


Figure 7.1. The neural response to external stimulation to the silent WM-network is an interaction between the input properties and the current state of the network.

Within the activity-silent WM framework, item-specific delay activity only reflects the focus of attention towards specific information in WM, and not WM content per se. To maintain multiple items in WM, activity states need to periodically refresh the transient connectivity patterns of each item in the WM network (Bahramisharif, Jensen, Jacobs, & Lisman, 2018; Lundqvist et al., 2016). Only a single item may be maintained in a measurable, active neural state at any one time and unattended information could be maintained in an activity-silent state that does not elicit neural delay activity (Chapter 4). The neural impulse response during WM maintenance may thus provide a more complete picture of the current state of the WM network, than measuring endogenous delay activity alone.

It has been found that retro-cues can refocus attention to a previously unattended item in WM, transforming it from a silent to an active state (Lewis-Peacock & Postle, 2012; Sprague et al., 2016). Similarly, it was found that the neural representation of unattended WM content can seemingly be transformed to an active neural state through a targeted TMS pulse with a corresponding behavioural effect (Rose et al., 2016; Zokaei, Manohar, Husain, & Feredoes, 2014). In contrast to those findings, no evidence of either a beneficial or a detrimental effect on behaviour was currently found, nor did the impulse result in a clear internal attentional shift towards the unattended item (Chapter 4). A recent WM model that incorporates rapid plasticity demonstrated that while strong network stimulation disrupts attention-related neural activity, which affects WM recall, weak stimulation does not, though it nevertheless results in a WM-dependent impulse response (Manohar, Zokaei, Fallon, Vogels, & Husain, 2019). The passively viewed sensory impulse stimuli used in this thesis might not have been strong enough to disrupt or change the WM network and should thus not be interpreted as an actual transformation of WM states, but rather as the automatic neural "echo", reflecting both input and neural structure (Buonomano & Maass, 2009; Sugase-Miyamoto et al., 2008).

Orientations (Chapters 3-6) and the frequencies of pure tones (Chapter 5) were decodable during both perception and from the impulse responses during maintenance. And while both encoding and maintenance of orientations and tonal frequencies seem to be parametric (Chapter 4-6), providing evidence for a continuous population code (Ester, Serences, & Awh, 2009; Hubel & Wiesel, 1962; Uluç et al., 2018), no evidence for an overlap of the population code during perception and maintenance (as revealed by the impulse response) was found (Chapters 3-5). Information encoding and

maintenance are thus seemingly accomplished by qualitatively different coding schemes, suggesting that information maintenance is not a literal preservation of stimulation history, but rather a reformatted code that best reflects the current task demands and affords easy readout during recall (Myers et al., 2017). Stimulus encoding is highly dynamic itself (Chapter 3 & Chapter 4) (Cichy, Pantazis, & Oliva, 2014), possibly reflecting the rapid reformatting of the stimulus into a mnemonic code. However, the nature of the impulse stimulus might itself (and likely does) influence the neural impulse response, resulting in a non-linear interaction between the network and the impulse properties (Fig. 7.1.). The impulse response is thus not a linear increase of activity across the network, making comparisons between neural responses elicited by different stimuli difficult to compare. Indeed, while not explicitly tested in this thesis, it is likely that the impulse approach would not have worked if impulse stimuli had varied randomly across trials.

Stable WM coding scheme in a dynamic WM network

A fundamental property of the neural dynamics in a neural network with short-term synaptic plasticity is hysteresis. This implies that each neural activity state leaves behind a neural trace in the connectivity patterns that in turn influences the activity patterns of subsequent activity states (Buonomano & Maass, 2009; Stokes, 2015), resulting in highly dynamic activation patterns that change over time. In Chapter 6 it was found that while the WM coding scheme remained stable over time, the time-specific impulse responses were nevertheless highly different, suggesting that a dynamic sub-component in the network changed over time, while maintaining a stable coding scheme. Indeed, these low-dimensional dynamics were even detectible in a reanalysis of Chapter 3 (presented in Chapter 6) at an extremely short time-scale. That is, the WM network changed within a time-span of 60 ms during the delay period, resulting in distinctly different impulse responses. Thus, it seems that the proposed network dynamics, mediated by the interplay of short-term synaptic plasticity and neural activity, do not necessarily affect the WM coding scheme. A stable coding scheme simplifies readout by downstream systems, rendering a time-specific decoder nonessential (Druckmann & Chklovskii, 2012). The time-changing dynamics seem occur in an independent subspace of the WM network (Murray et al., 2017). While they could simply be an inevitable consequence of the reciprocity between synaptic changes and neural activity, they could also be an efficient way to code the passage of time (Buonomano & Maass, 2009).

While stable WM coding schemes have also been reported elsewhere (Murray et al., 2017; Spaak et al., 2017), WM-specific delay activity has also been found to be highly dynamic (Meyers et al., 2008; Stokes et al., 2013). It has been proposed that more complex WM tasks may be more likely to result in dynamic coding schemes (Meyers, 2018), possibly due to a longer and multifaceted neural transformation from a stimulus code to a more abstract WM code (Myers et al., 2017), whereas the maintenance of simple sensory features (like orientations used in this thesis) can be maintained in a stable

coding scheme, as observed in Chapter 6. Future studies should systematically test the relationship between neural dynamics and item and task complexity.

Multiple WM states

The remarkable retro-cue effect in WM, first reported 16 years ago (Griffin & Nobre, 2003), highlights that WM content is not simply a fixed copy of the encoding stage, but that the internal focus of attention can selectively enhance, protect, transform, and retrieve individual pieces of information in WM, resulting in a seemingly "privileged" state for attended information in WM (Makovski & Pertzov, 2015; Souza & Oberauer, 2016; Souza, Rerko, & Oberauer, 2016). The neural signature of orienting attention towards specific information in WM has been found to be very similar to attentional orienting in external space, as both are accompanied by alpha power modulations that track the locations of attended internal WM representations and external stimuli (Foster, Sutterer, Serences, Vogel, & Awh, 2015; Samaha, Sprague, & Postle, 2016). In Chapter 3 the retro-cue resulted in a marked reduction in alpha power contralateral to the previously presented location of the cued item, providing evidence for an internal orienting and selection of the cued item in WM. While the alpha lateralization was maintained until the end of the trial, it peaked at roughly 500 ms after cue presentation, consistent with other reports of transient alpha lateralization after cue presentation (Wallis, Stokes, Cousijn, Woolrich, & Nobre, 2015), and behavioural studies suggesting that the selecting and transforming a WM item into a privileged attended state takes about 400 ms (Souza, Rerko, & Oberauer, 2014). Since the retro-cue in Chapter 3 was 100% valid, participants had no reason to maintain the uncued item, and could therefore drop it. Indeed, the subsequent impulse response did not contain any information about the uncued item. Considering the short time span between cue and impulse (1 sec), the cue may not only have directed attention towards the cued item, but also resulted in an active forgetting and removal of the uncued item from the hidden state (Oberauer & Lin, 2017).

It has been suggested that alpha lateralization does not necessarily reflect the continuous focus of attention in WM, but rather the transient selection and transformation of an item to an attended state (Souza & Oberauer, 2016), whereas item-related neural activity does reflect the attended but not unattended WM item (Larocque et al., 2014). In the second experiment in Chapter 3, the attentional selection and switching was indeed accompanied by corresponding transient alpha power lateralization modulations, providing neurophysiological evidence that a previously unattended item can be reprioritized in anticipation of the upcoming test probe (Ede et al., 2016), and highlighting the flexibility of moving WM items in and out of distinct WM states depending on context (Zokaei, Ning, Manohar, Feredoes, & Husain, 2014). Notably, these alpha modulations were relatively short-lived, whereas the item-specific delay activity of the early attended item was stable across the delay, thus dissociating attentional selection from active maintenance. Additionally, the impulse response provided evidence that unattended information is nevertheless still present in the hidden state.

As mentioned before, item-specific delay activity may only be present for attended WM content (Larocque et al., 2014; van Loon et al., 2018; K. Watanabe & Funahashi, 2014). This activity may not only enable the a heightened readiness to recall the attended item in anticipation of the test probe, reflected in the ramp up of item-specific activity towards the end of the maintenance delay (K. Watanabe & Funahashi, 2007), but also a fundamental reconfiguration of the WM network to meet changing task demands, and a transformation of the neural representation of WM content into a neural code that can be most efficiently read out and compared to task-specific probes (Myers et al., 2017). A recent study found that a biologically sensible recurrent neural network with short-term synaptic plasticity can maintain items in WM with little to no neural activity, but only when no task related item manipulation is required, which in turn increases WM delay activity as a function of required manipulation (Masse, Yang, Song, Wang, & Freedman, 2019). Unattended information may thus more faithfully represent the sensory stimulation history left behind in the neural network, whereas attended information is transformed into a qualitatively different neural state. Indeed, a reanalysis of experiment 2 in Chapter 4 found no evidence that the neural code of an unattended item cross-generalised with the neural code once it had been attended again (Suppl. fig. 7.1B), suggesting that the attended item was transformed into a fundamentally different neural code. Note that the temporal separation of the impulses perturbing the item when it is unattended and attended in Chapter 4 does make any firm conclusions about the attentional modulation of the WM code problematic, as time itself may result in a dynamically changing WM code (Meyers, 2018). However, evidence against this interpretation comes from Chapter 6, where the coding scheme was found to be stable over time when no attentional modulation was required.

It has recently been reported that attended (currently relevant) and unattended (relevant later) information may not necessarily be coded in different, but opposite schemes, resulting in negative cross-generalization between their neural representations (van Loon et al., 2018). This suggests that the same neural networks are used to store currently and prospectively relevant items, and that the difference arise from opposing neural patterns. No negative cross-generalization was evident in the reanalysis of Chapter 4 (Suppl. fig. 7.1B). This discrepancy could be attributed to different task-evoked neural responses (search array versus irrelevant impulse stimulus) and measurements (BOLD versus EEG). More research is needed to establish differences between the neural representations of attended and unattended information.

The involvement of sensory processing areas in WM maintenance

The impulse approach applied in this thesis assumes that the bottom-up neural response elicited by an irrelevant and predictable sensory stimulus will somehow interact with the very neural network that contains the WM content, which seems to be the case as reported in this thesis. Is this evidence that sensory processing areas are involved in WM

maintenance? According to the sensorimotor-recruitment account, the very brain regions involved in the perception of sensory information are also involved in their short-term maintenance (Postle, 2016; Scimeca et al., 2018; Serences, 2016), and while the evidence of WM specific delay activity in corresponding sensory cortices is abundant for different sensory modalities (Gottlieb, Vaadia, & Abeles, 1989; Harrison & Tong, 2009; Kumar et al., 2016; Zhou & Fuster, 1996) it is nevertheless unclear if, in the case of visual WM, the visual cortex is actually necessary for visual maintenance (Xu, 2018). Indeed, it has been shown that the presence of distractor stimuli can remove the visual code from early visual cortex while maintaining it in IPS (Bettencourt & Xu, 2016). These results can be interpreted as evidence for changing maintenance strategies (visual or not visual when distractors could interfere with the visual code (Pearson & Keogh, 2019). WM-related research with non-human primates is heavily dominated by single unit recordings from the PFC, with abundant evidence of WM-related activity in that area (Sreenivasan et al., 2014). However, these discrepancies also highlight the general problem of solely relying on measurable activity to test for the involvement of specific cortical areas in WM, as the absence of WM-related activity is essentially a null result. The impulse approach my help to address this issue.

While attention may result in measurable WM-dependent neural patterns in some brain regions, silent WM-dependent connectivity changes may present in others. This thesis provides evidence that the bottom up neural signal interacts with the WM network of both visual (Chapters 3-6) and auditory (Chapter 5) information. A searchlight analysis in Chapter 6 (and reanalysis of Chapter 4, see Suppl. fig. 7.1A) highlights that posterior areas contribute most to the decoding of visual WM content, suggesting that the WM network the visual impulse interacts with is indeed in the visual cortex. Note that the decoding topography is less clear and seemingly more distributed for unattended WM content (Suppl. fig. 7.1A), which could reflect that unattended visual information is maintained in a distinct format in parietal cortex and not in the visual cortex (Christophel, Iamshchinina, Yan, Allefeld, & Haynes, 2018). However, the limited spatial resolution of EEG and the inverse problem that is associated with localising EEG signals (Grech et al., 2008) make any firm conclusion about the involved brain regions problematic. Nevertheless, it is still relatively safe to assert that the decoding topography shows no evidence of decoding in frontal electrodes, making any involvement of the PFC in the impulse response effects reported in this thesis unlikely. This is at odds with the one other report of a WM-dependent neural impulse response (Stokes et al., 2013). Here, monkeys performed a delay paired-associate recognition task while activity from PFC was recorded after learning arbitrary associations between specific visual stimuli. Each trial began with the cue stimulus that was associated with a specific target stimulus and the monkeys were required to look at the target stimulus, which was presented after a random number of non-targets. It was found that the presentation of neutral non-targets not only resulted in a marked, but short-lived overall increase in neural activity in PFC, but also in an accompanying transient increase in decoding accuracy of the cue. This is similar to the results in this thesis in that the neural response to a neutral, uninformative visual stimulus contained information about an item held in WM.

So why was no involvement of PFC evident in this thesis? There are several possible reasons for this discrepancy. First, decoding from frontal regions seems generally more difficult with non-invasive neuroimaging in humans, suggested by the abundance of fMRI research finding visual WM codes in visual and parietal cortices but rarely in PFC (Serences, 2016), which can be attributed to the heterogeneity of PFC neurons (Fusi et al., 2016; Manohar et al., 2019). Secondly, the nature of the task in Stokes et al. (2013) required monkeys to encode the neutral non-target stimulus and actively compare it to the internally held target stimulus to be able to decide whether to respond. It is therefore possible that the increase in cue decoding was driven by endogenous neural activity, and not an exogenously driven impulse-response. In contrast, the impulse stimuli in this thesis were always the same, irrelevant, and predictable (with the exception of Chapter 3), and thus did not need to be fully processed. A bottom-up neural response may thus not be enough to interact with the higher order WM network in PFC, and may only do so with low-level WM codes in sensory cortex.

Chapter 5 furthermore provides evidence that visual information may be maintained in a sensory-specific network, since an auditory bottom-up neural response did not contain information about the currently maintained visual WM content. The neural response to auditory stimulation thus does not seem to reach the representation of visual WM content. This suggests that visual information is maintained in a closed system, and while auditory stimulation has been found to excite neurons in the visual cortex (Martuzzi et al., 2007), it might not be enough to result in a measurable, WM-specific response. In contrast, both visual and auditory stimulation was found to result in neural responses that contained information about auditory WM content, providing evidence for cross-sensory involvement in the auditory WM network. On the one hand, this could mean direct involvement of the visual cortex in auditory maintenance; on the other hand, the visual cortex may provide an access path to the neural representation of auditory representations elsewhere.

In sum, the fact that the bottom-up neural response contains information about current WM information, implicates the sensory processing areas in WM maintenance. However, it is unclear if they are directly involved, or merely provide access paths to the WM representation elsewhere, resulting in state-dependent responses. Targeted neural stimulation, in conjunction with high-resolution neuroimaging are needed to further chart the brain areas involved in connectivity-dependent WM maintenance.

Evidence for "silent" WM?

This thesis highlights that WM delay activity may not be the only neural mechanism underlying WM maintenance, and provides evidence for the effectiveness of a relatively simple approach to reveal otherwise hidden WM states, by measuring the WM-dependent neural impulse response. These findings are in line with previous predictions of connectivity-dependent, activity-silent WM (Buonomano & Maass, 2009; Mongillo et

al., 2008; Sugase-Miyamoto et al., 2008), and several recent WM models have implemented transient connectivity-changes as a prominent feature (Manohar et al., 2019; Masse et al., 2019; Miller et al., 2018; Stokes, 2015). Nevertheless, the impulse-approach is by no means direct evidence for "silent" WM maintenance. Reports of no evidence for WM-related delay activity, both in this thesis (Chapter 4) and in other studies (Lewis-Peacock & Postle, 2012; Lundqvist et al., 2016; Rose et al., 2016; Sprague et al., 2016) are essentially null results. Non-invasive recording techniques might not be sensitive enough to detect subtle neural activity states, whereas invasive recordings are limited by the number of neurons and brain areas that can be sampled simultaneously. Alternative possibilities for the here reported impulse-response need to be appreciated. Note that these are not necessarily mutually exclusive.

1. Distinct WM states, which are maintained through low-level, reverberating persistent neural activity, could also elicit a state-dependent neural response given a non-linear interaction with the impulse stimulus. For example, the phase and power of neural oscillations have been found to affect stimulus evoked neural responses (Iemi et al., 2019).
2. The impulse stimulus could result in a decrease in neural background noise (Churchland et al., 2010), exposing the WM-specific delay activity, making it easier to measure.
3. The impulse stimulus could phase-reset WM-dependent neural oscillations (Hanslmayr et al., 2007; Roux & Uhlhaas, 2014; Sauseng et al., 2007) and thus realign oscillations that are difficult to measure to a specific time-point that is more easily detectible in the averaged EEG signal.

Given these possibilities, as well as the dominance of attractor models based on persistent delay in WM, a contentious debate about whether or not WM maintenance depends on persistent neural activity has unfolded (Constantinidis et al., 2018; Lundqvist, Herman, & Miller, 2018). In fact, a recent fMRI study tested a very large number of participants (n=87) to explicitly test whether or not unattended visual WM content can be decoded from the BOLD delay activity (Christophel et al., 2018). Interestingly, and in contrast to previously reported null findings, both the attended *and* the unattended item could be robustly decoded from both the frontal eye-fields and IPS, while only the attended item could be decoded from the visual cortex, providing convincing evidence that unattended visual information are not necessarily completely silent, but simply more difficult to detect due to abstract, non-visual codes in non-visual areas. Note however, that connectivity-dependent WM does not assert WM maintenance to be completely silent, as periodic activity states are necessary to reinstate the decaying short-term synaptic changes, which may only last for ~2s (Mongillo et al., 2008; Zucker & Regehr, 2002). Given the long delay period and analysis window (6s) in Christophel et al. (2018) the unattended item may have had to be refreshed several times, resulting in measurable activity in their high-powered experimental design.

However, it is clear that it is near impossible to find convincing evidence for or against "silent" WM with neuroimaging that rely on neural activity. Measures of short-term connectivity changes need to be employed to provide direct evidence for connectivity-dependent WM (Fujisawa et al., 2008). The impulse-approach is nonetheless a useful tool to explore non-spatial and non-categorical WM states that are otherwise difficult to measure, in particular in EEG.

References

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared Representations for Working Memory and Mental Imagery in Early Visual Cortex. *Current Biology*, *23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87. https://doi.org/10.1007/BF00337259

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278.

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

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

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

Bahramisharif, A., Jensen, O., Jacobs, J., & Lisman, J. (2018). Serial representation of items during working memory maintenance at letter-selective cortical sites. *PLOS Biology*, *16*(8), e2003805. https://doi.org/10.1371/journal.pbio.2003805

Barak, O., Sussillo, D., Romo, R., Tsodyks, M., & Abbott, L. F. (2013). From fixed points to chaos: Three models of delayed discrimination. *Progress in Neurobiology*, *103*, 214–222. https://doi.org/10.1016/j.pneurobio.2013.02.002

Barak, O., & Tsodyks, M. (2014). Working models of working memory. *Current Opinion in Neurobiology*, *25*, 20–24. https://doi.org/10.1016/j.conb.2013.10.008

Barak, O., Tsodyks, M., & Romo, R. (2010). Neuronal Population Coding of Parametric Working Memory. *Journal of Neuroscience*, *30*(28), 9424–9430. https://doi.org/10.1523/JNEUROSCI.1875-10.2010

Barrouillet, P., & Camos, V. (2001). Developmental Increase in Working Memory Span: Resource Sharing or Temporal Decay? *Journal of Memory and Language*, *45*(1), 1–20. https://doi.org/10.1006/jmla.2001.2767

Bastos, A. M., Loonis, R., Kornblith, S., Lundqvist, M., & Miller, E. K. (2018). Laminar recordings in frontal cortex suggest distinct layers for maintenance and control of working memory. *Proceedings of the National Academy of Sciences*, 201710323. https://doi.org/10.1073/pnas.1710323115

Bauer, R. H., & Fuster, J. M. (1976). Delayed-matching and delayed-response deficit from cooling dorsolateral prefrontal cortex in monkeys. *Journal of Comparative and Physiological Psychology*, *90*(3), 293–302. https://doi.org/10.1037/h0087996

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

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

Bliss, D. P., Sun, J. J., & D'Esposito, M. (2017). Serial dependence is absent at the time of perception but increases in visual working memory. *Scientific Reports*, *7*(1), 14739. https://doi.org/10.1038/s41598-017-15199-7

Bortoletto, M., Veniero, D., Thut, G., & Miniussi, C. (2015). The contribution of TMS–EEG coregistration in the exploration of the human cortical connectome. *Neuroscience & Biobehavioral Reviews*, *49*, 114–124. https://doi.org/10.1016/j.neubiorev.2014.12.014

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

Boutros, N. N., Korzyukov, O., Jansen, B., Feingold, A., & Bell, M. (2004). Sensory gating deficits during the mid-latency phase of information processing in medicated schizophrenia patients. *Psychiatry Research*, *126*(3), 203–215. https://doi.org/10.1016/j.psychres.2004.01.007

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Briley, P. M., Breakey, C., & Krumbholz, K. (2013). Evidence for Pitch Chroma Mapping in Human Auditory Cortex. *Cerebral Cortex*, *23*(11), 2601–2610. https://doi.org/10.1093/cercor/bhs242

Brody, C. D., Romo, R., & Kepecs, A. (2003). Basic mechanisms for graded persistent activity: Discrete attractors, continuous attractors, and dynamic representations. *Current Opinion in Neurobiology*, *13*(2), 204–211. https://doi.org/10.1016/S0959-4388(03)00050-3

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(44), 13992–14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, *10*(2), 113–125. https://doi.org/10.1038/nrn2558

Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex. *Neuron*, *76*(4), 838–846. https://doi.org/10.1016/j.neuron.2012.09.029

Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, *108*(27), 11252–11255. https://doi.org/10.1073/pnas.1104666108

Catterall, W. A., Leal, K., & Nanou, E. (2013). Calcium Channels and Short-term Synaptic Plasticity. *The Journal of Biological Chemistry*, *288*(15), 10742–10749. https://doi.org/10.1074/jbc.R112.411645

Chang, A., Bosnyak, D. J., & Trainor, L. J. (2016). Unpredicted Pitch Modulates Beta Oscillatory Power during Rhythmic Entrainment to a Tone Sequence. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00327

Chao, L. L., & Knight, R. T. (1998). Contribution of Human Prefrontal Cortex to Delay Performance. *Journal of Cognitive Neuroscience*, *10*(2), 167–177. https://doi.org/10.1162/089892998562636

Chaudhuri, R., & Fiete, I. (2016). Computational principles of memory. *Nature Neuroscience*, *19*(3), 394–403. https://doi.org/10.1038/nn.4237

Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C., & Haynes, J.-D. (2018). Cortical specialization for attended versus unattended working memory. *Nature Neuroscience*, *21*(4), 494. https://doi.org/10.1038/s41593-018-0094-4

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, *21*(2), 111–124. https://doi.org/10.1016/j.tics.2016.12.007

Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., … Shenoy, K. V. (2010). Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience*, *13*(3), 369–378. https://doi.org/10.1038/nn.2501

Cichy, R. M., Chen, Y., & Haynes, J.-D. (2011). Encoding the identity and location of objects in human LOC. *NeuroImage*, *54*(3), 2297–2307. https://doi.org/10.1016/j.neuroimage.2010.09.044

Cichy, R. M., & Pantazis, D. (2017). Multivariate pattern analysis of MEG and EEG: A comparison of representational structure in time and space. *NeuroImage*, *158*, 441–454. https://doi.org/10.1016/j.neuroimage.2017.07.023

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cichy, R. M., Ramirez, F. M., & Pantazis, D. (2015). Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *NeuroImage*, *121*, 193–204. https://doi.org/10.1016/j.neuroimage.2015.07.011

Claessens, P. M. E., & Wagemans, J. (2008). A Bayesian framework for cue integration in multistable grouping: Proximity, collinearity, and orientation priors in zigzag lattices. *Journal of Vision*, *8*(7), 33–33. https://doi.org/10.1167/8.7.33

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (2000). Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cerebral Cortex*, *10*(9), 910–923. https://doi.org/10.1093/cercor/10.9.910

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X.-L., Wang, M., & Arnsten, A. F. T. (2018). Persistent Spiking Activity Underlies Working Memory. *Journal of Neuroscience*, *38*(32), 7020–7028. https://doi.org/10.1523/JNEUROSCI.2486-17.2018

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552. https://doi.org/10.1016/j.tics.2003.10.005

Cowan, N. (2010). The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Cromwell, H. C., Mears, R. P., Wan, L., & Boutros, N. N. (2008). Sensory Gating: A Translational Effort from Basic to Clinical Science. *Clinical EEG and Neuroscience*, *39*(2), 69–72. https://doi.org/10.1177/155005940803900209

Crowe, D. A., Averbeck, B. B., & Chafee, M. V. (2010). Rapid Sequences of Population Activity Patterns Dynamically Encode Task-Critical Spatial Information in Parietal Cortex. *The Journal of Neuroscience*, *30*(35), 11640–11653. https://doi.org/10.1523/JNEUROSCI.0954-10.2010

Cueva, C. J., Marcos, E., Saez, A., Genovesio, A., Jazayeri, M., Romo, R., … Fusi, S. (2019). Low dimensional dynamics for working memory and time encoding. *BioRxiv*, 504936. https://doi.org/10.1101/504936

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423. https://doi.org/10.1016/S1364-6613(03)00197-9

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Devinsky, O., & D'Esposito, M. (2003). *Neurology of Cognitive and Behavioral Disorders*. Oxford University Press.

Di Russo, F., Martínez, A., & Hillyard, S. A. (2003). Source Analysis of Event-related Cortical Activity during Visuo-spatial Attention. *Cerebral Cortex*, *13*(5), 486–499. https://doi.org/10.1093/cercor/13.5.486

Driver, J., & Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in Cognitive Sciences*, *2*(7), 254–262. https://doi.org/10.1016/S1364-6613(98)01188-7

Druckmann, S., & Chklovskii, D. B. (2012). Neuronal Circuits Underlying Persistent Representations Despite Time Varying Activity. *Current Biology*, *22*(22), 2095–2103. https://doi.org/10.1016/j.cub.2012.08.058

Eckert, M. A., Kamdar, N. V., Chang, C. E., Beckmann, C. F., Greicius, M. D., & Menon, V. (2008). A Cross-Modal System Linking Primary Auditory and Visual Cortices. *Human Brain Mapping*, *29*(7), 848–857. https://doi.org/10.1002/hbm.20560

Ede, F. van, Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2019). Concurrent visual and motor selection during visual working memory guided action. *Nature Neuroscience*, *22*(3), 477. https://doi.org/10.1038/s41593-018-0335-6

Ede, F. van, Niklaus, M., & Nobre, A. C. (2016). Temporal expectations guide dynamic prioritization in visual working memory through attenuated alpha oscillations. *Journal of Neuroscience*, 2272–16. https://doi.org/10.1523/JNEUROSCI.2272-16.2016

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., & Compte, A. (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences*, *106*(16), 6802–6807. https://doi.org/10.1073/pnas.0901894106

Ester, E. F., Rademaker, R. L., & Sprague, T. C. (2016). How Do Visual and Parietal Cortex Contribute to Visual Short-Term Memory? *ENeuro*, *3*(2), ENEURO.0041-16.2016. https://doi.org/10.1523/ENEURO.0041-16.2016

Ester, E. F., Serences, J. T., & Awh, E. (2009). Spatially Global Representations in Human Primary Visual Cortex during Working Memory Maintenance. *Journal of Neuroscience*, *29*(48), 15258–15265. https://doi.org/10.1523/JNEUROSCI.4388-09.2009

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron*, *87*(4), 893–905. https://doi.org/10.1016/j.neuron.2015.07.013

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*(5), 738–743. https://doi.org/10.1038/nn.3689

Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K., & Awh, E. (2015). The topography of alpha-band activity tracks the content of spatial working memory. *Journal of Neurophysiology*, *115*(1), 168–177. https://doi.org/10.1152/jn.00860.2015

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. *Science*, *291*(5502), 312–316. https://doi.org/10.1126/science.291.5502.312

Freeman, J., Heeger, D. J., & Merriam, E. P. (2013). Coarse-Scale Biases for Spirals and Orientation in Human Visual Cortex. *Journal of Neuroscience*, *33*(50), 19695–19703. https://doi.org/10.1523/JNEUROSCI.0889-13.2013

Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite Effects of Recent History on Perception and Decision. *Current Biology*, *27*(4), 590–595. https://doi.org/10.1016/j.cub.2017.01.006

Fujisawa, S., Amarasingham, A., Harrison, M. T., & Buzsáki, G. (2008). Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature Neuroscience*, *11*(7), 823–833. https://doi.org/10.1038/nn.2134

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349. https://doi.org/10.1152/jn.1989.61.2.331

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1993). Dorsolateral prefrontal lesions and oculomotor delayed-response performance: Evidence for mnemonic "scotomas." *Journal of Neuroscience*, *13*(4), 1479–1497. https://doi.org/10.1523/JNEUROSCI.13-04-01479.1993

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Fuster, J. M. (1973). Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *Journal of Neurophysiology*, *36*(1), 61–78. https://doi.org/10.1152/jn.1973.36.1.61

Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654. https://doi.org/10.1126/science.173.3997.652

Gayet, S., Paffen, C. L. E., & Van der Stigchel, S. (2018). Visual Working Memory Storage Recruits Sensory Processing Areas. *Trends in Cognitive Sciences*, *22*(3), 189–190. https://doi.org/10.1016/j.tics.2017.09.011

Gazzaley, A., Cooney, J. W., Rissman, J., & D'Esposito, M. (2005). Top-down suppression deficit underlies working memory impairment in normal aging. *Nature Neuroscience*, *8*(10), 1298–1300. https://doi.org/10.1038/nn1543

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129–135. https://doi.org/10.1016/j.tics.2011.11.014

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485. https://doi.org/10.1016/0896-6273(95)90304-6

Gottlieb, Y., Vaadia, E., & Abeles, M. (1989). Single unit activity in the auditory cortex of a monkey performing a short term memory task. *Experimental Brain Research*, *74*(1), 139–148. https://doi.org/10.1007/BF00248287

Grech, R., Cassar, T., Muscat, J., Camilleri, K. P., Fabri, S. G., Zervakis, M., … Vanrumste, B. (2008). Review on solving the inverse problem in EEG source analysis. *Journal of NeuroEngineering and Rehabilitation*, *5*, 25. https://doi.org/10.1186/1743-0003-5-25

Griffin, I. C., & Nobre, A. C. (2003). Orienting Attention to Locations in Internal Representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Hanslmayr, S., Klimesch, W., Sauseng, P., Gruber, W., Doppelmayr, M., Freunberger, R., … Birbaumer, N. (2007). Alpha Phase Reset Contributes to the Generation of ERPs. *Cerebral Cortex*, *17*(1), 1–8. https://doi.org/10.1093/cercor/bhj129

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Hempel, C. M., Hartman, K. H., Wang, X.-J., Turrigiano, G. G., & Nelson, S. B. (2000). Multiple Forms of Short-Term Plasticity at Excitatory Synapses in Rat Medial Prefrontal Cortex. *Journal of Neurophysiology*, *83*(5), 3031–3041.

Hernández, A., Nácher, V., Luna, R., Zainos, A., Lemus, L., Alvarez, M., … Romo, R. (2010). Decoding a Perceptual Decision Process across Cortex. *Neuron*, *66*(2), 300–314. https://doi.org/10.1016/j.neuron.2010.03.031

Howard, M. W., Rizzuto, D. S., Caplan, J. B., Madsen, J. R., Lisman, J., Aschenbrenner-Scheibe, R., … Kahana, M. J. (2003). Gamma oscillations correlate with working memory load in humans. *Cerebral Cortex (New York, N.Y.: 1991)*, *13*(12), 1369–1374.

Huang, Y., Matysiak, A., Heil, P., König, R., & Brosch, M. (2016). Persistent neural activity in auditory cortex is related to auditory working memory in humans and nonhuman primates. *ELife*, *5*. https://doi.org/10.7554/eLife.15441

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722

Iemi, L., Busch, N. A., Laudini, A., Haegens, S., Samaha, J., Villringer, A., & Nikulin, V. V. (2019). Multiple mechanisms link prestimulus neural oscillations to sensory responses. *ELife*, *8*, e43620. https://doi.org/10.7554/eLife.43620

Inagaki, H. K., Fontolan, L., Romani, S., & Svoboda, K. (2019). Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature*, 1. https://doi.org/10.1038/s41586-019-0919-7

Iurilli, G., Ghezzi, D., Olcese, U., Lassi, G., Nazzaro, C., Tonini, R., … Medini, P. (2012). Sound-Driven Synaptic Inhibition in Primary Visual Cortex. *Neuron*, *73*(4), 814–828. https://doi.org/10.1016/j.neuron.2011.12.026

JASP Team. (2018). *JASP (Version 0.9) [Computer software]*.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671. https://doi.org/10.3758/BF03196323

Kinchla, R. A., & Smyzer, F. (1967). A diffusion model of perceptual memory. *Perception & Psychophysics*, *2*(6), 219–229. https://doi.org/10.3758/BF03212471

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Kisley, M. A., Noecker, T. L., & Guinther, P. M. (2004). Comparison of sensory gating to mismatch negativity and self-reported perceptual phenomena in healthy adults. *Psychophysiology*, *41*(4), 604–612. https://doi.org/10.1111/j.1469-8986.2004.00191.x

Kiyonaga, A., & Egner, T. (2016). Center-Surround Inhibition in Working Memory. *Current Biology*, *26*(1), 64–68. https://doi.org/10.1016/j.cub.2015.11.013

Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial Dependence across Perception, Attention, and Memory. *Trends in Cognitive Sciences*, *21*(7), 493–497. https://doi.org/10.1016/j.tics.2017.04.011

Kleiner, M. (2010). *Visual stimulus timing precision in Psychtoolbox-3: Tests, pitfalls solutions*. Retrieved from http://www.neuroschool-tuebingen-

nena.de/fileadmin/user_upload/Dokumente/neuroscience/AbstractbookNeNa2010u.pdf

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology*, *34*(3), 337–347. https://doi.org/10.1152/jn.1971.34.3.337

Kumar, S., Joseph, S., Gander, P. E., Barascud, N., Halpern, A. R., & Griffiths, T. D. (2016). A Brain System for Auditory Working Memory. *Journal of Neuroscience*, *36*(16), 4492–4505. https://doi.org/10.1523/JNEUROSCI.4341-14.2016

Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, *43*(2), 149–164. https://doi.org/10.1016/S0042-6989(02)00402-9

Lara, A. H., & Wallis, J. D. (2014). Executive control processes underlying multi-item working memory. *Nature Neuroscience*, *17*(6), 876–883. https://doi.org/10.1038/nn.3702

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Decoding Attended Information in Short-term Memory: An EEG Study. *Journal of Cognitive Neuroscience*, *25*(1), 127–142. https://doi.org/10.1162/jocn_a_00305

Larocque, J. J., Lewis-Peacock, J. A., & Postle, B. R. (2014). Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience*, *8*, 5. https://doi.org/10.3389/fnhum.2014.00005

Ledoit, O., & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management*, *30*(4), 110–119. https://doi.org/10.3905/jpm.2004.110

Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nature Neuroscience*, *16*(8), 997–999. https://doi.org/10.1038/nn.3452

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2011). Neural Evidence for a Distinction between Short-term Memory and the Focus of

Attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140

Lewis-Peacock, J. A., & Postle, B. R. (2012). Decoding the internal focus of attention. *Neuropsychologia*, *50*(4), 470–478. https://doi.org/10.1016/j.neuropsychologia.2011.11.006

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique Effect: A Neural Basis in the Visual Cortex. *Journal of Neurophysiology*, *90*(1), 204–217. https://doi.org/10.1152/jn.00954.2002

Lim, P. C., Ward, E. J., Vickery, T. J., & Johnson, M. R. (2019). Not-so-working Memory: Drift in Functional Magnetic Resonance Imaging Pattern Representations during Maintenance Predicts Errors in a Visual Working Memory Task. *Journal of Cognitive Neuroscience*, 1–15. https://doi.org/10.1162/jocn_a_01427

López, J. D., Litvak, V., Espinosa, J. J., Friston, K., & Barnes, G. R. (2014). Algorithmic procedures for Bayesian MEG/EEG source reconstruction in SPM. *NeuroImage*, *84*, 476–487. https://doi.org/10.1016/j.neuroimage.2013.09.002

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Sciences*, *4*(11), 432–440. https://doi.org/10.1016/S1364-6613(00)01545-X

Lundqvist, M., Herman, P., & Lansner, A. (2011). Theta and Gamma Power Increases and Alpha/Beta Power Decreases with Memory Load in an Attractor Network Model. *Journal of Cognitive Neuroscience*, *23*(10), 3008–3020. https://doi.org/10.1162/jocn_a_00029

Lundqvist, M., Herman, P., & Miller, E. K. (2018). Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. *Journal of Neuroscience*, *38*(32), 7013–7019. https://doi.org/10.1523/JNEUROSCI.2485-17.2018

Lundqvist, M., Herman, P., Warden, M. R., Brincat, S. L., & Miller, E. K. (2018). Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature Communications*, *9*(1), 394. https://doi.org/10.1038/s41467-017-02791-8

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., & Miller, E. K. (2016). Gamma and Beta Bursts Underlie Working Memory. *Neuron*, *90*(1), 152–164. https://doi.org/10.1016/j.neuron.2016.02.028

Magnussen, S., Greenlee, M. W., Asplund, R., & Dyrnes, S. (1991). Stimulus-specific mechanisms of visual short-term memory. *Vision Research*, *31*(7), 1213–1219. https://doi.org/10.1016/0042-6989(91)90046-8

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *National Institute of Science of India*.

Makovski, T., & Pertzov, Y. (2015). Attention and memory protection: Interactions between retrospective attention cueing and interference. *Quarterly Journal of Experimental Psychology*, *68*(9), 1735–1743. https://doi.org/10.1080/17470218.2015.1049623

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T., & Husain, M. (2017). A neural model of working memory. *BioRxiv*, 233007. https://doi.org/10.1101/233007

Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P., & Husain, M. (2019). Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews*, *101*, 1–12. https://doi.org/10.1016/j.neubiorev.2019.03.017

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. https://doi.org/10.1038/nature12742

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martínez-García, M., Rolls, E. T., Deco, G., & Romo, R. (2011). Neural and computational mechanisms of postponed decisions. *Proceedings of the National Academy of Sciences*, *108*(28), 11626–11631. https://doi.org/10.1073/pnas.1108137108

Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J.-P., Maeder, P. P., Clarke, S., & Meuli, R. A. (2007). Multisensory Interactions within Human Primary Cortices Revealed by BOLD Dynamics. *Cerebral Cortex*, *17*(7), 1672–1679. https://doi.org/10.1093/cercor/bhl077

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, 1. https://doi.org/10.1038/s41593-019-0414-3

McKee, J. L., Riesenhuber, M., Miller, E. K., & Freedman, D. J. (2014). Task Dependence of Visual and Category Representations in Prefrontal and Inferior Temporal Cortices. *Journal of Neuroscience*, *34*(48), 16065–16075. https://doi.org/10.1523/JNEUROSCI.1660-14.2014

Mendoza-Halliday, D., Torres, S., & Martinez-Trujillo, J. C. (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nature Neuroscience*, *17*(9), 1255–1262. https://doi.org/10.1038/nn.3785

Meyer, T., Qi, X.-L., Stanford, T. R., & Constantinidis, C. (2011). Stimulus Selectivity in Dorsal and Ventral Prefrontal Cortex after Training in Working Memory Tasks. *Journal of Neuroscience*, *31*(17), 6266–6276. https://doi.org/10.1523/JNEUROSCI.6798-10.2011

Meyers, E. M. (2018). Dynamic population coding and its relationship to working memory. *Journal of Neurophysiology*, *120*(5), 2260–2268. https://doi.org/10.1152/jn.00225.2018

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. *Journal of Neurophysiology*, *100*(3), 1407–1419. https://doi.org/10.1152/jn.90248.2008

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. *Journal of Neuroscience*, *16*(16), 5154–5167.

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, *100*(2), 463–475. https://doi.org/10.1016/j.neuron.2018.09.023

Miller, M. H., & Orbach, J. (1972). Retention of spatial alternation following frontal lobe resections in stump-tailed macaques. *Neuropsychologia*, *10*(3), 291–298. https://doi.org/10.1016/0028-3932(72)90020-6

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Morrill, R. J., & Hasenstaub, A. R. (2018). Visual Information Present in Infragranular Layers of Mouse Auditory Cortex. *Journal of Neuroscience*, *38*(11), 2854–2862. https://doi.org/10.1523/JNEUROSCI.3102-17.2018

Murray, A. M., Nobre, A. C., & Stokes, M. G. (2011). Markers of preparatory attention predict visual short-term memory performance. *Neuropsychologia*, *49*(6), 1458–1465. https://doi.org/10.1016/j.neuropsychologia.2011.02.016

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., & Wang, X.-J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences*, *114*(2), 394–399. https://doi.org/10.1073/pnas.1619449114

Myers, N. E., Rohenkohl, G., Wyart, V., Woolrich, M. W., Nobre, A. C., & Stokes, M. G. (2015). Testing sensory evidence against mnemonic templates. *ELife*, *4*, e09000. https://doi.org/10.7554/eLife.09000

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2017.03.010

Nemrodov, D., Niemeier, M., Patel, A., & Nestor, A. (2018). The Neural Dynamics of Facial Identity Processing: Insights from EEG-Based Pattern Analysis and Image Reconstruction. *ENeuro*, *5*(1), ENEURO.0358-17.2018. https://doi.org/10.1523/ENEURO.0358-17.2018

Nikolić, D., Häusler, S., Singer, W., & Maass, W. (2009). Distributed Fading Memory for Stimulus Properties in the Primary Visual Cortex. *PLOS Biol*, *7*(12), e1000260. https://doi.org/10.1371/journal.pbio.1000260

Nikolić, D., Häusler, Stefan, Singer, Wolf, & Maass, Wolfgang. (2007). Temporal dynamics of information content carried by neurons in the primary visual cortex. *Advances in Neural Information Processing Systems*, *19*, 1041–1048.

Nobre, A., Correa, A., & Coull, J. (2007). The hazards of time. *Current Opinion in Neurobiology*, *17*(4), 465–470. https://doi.org/10.1016/j.conb.2007.07.006

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Oberauer, K., & Hein, L. (2012). Attention to Information in Working Memory. *Current Directions in Psychological Science*, *21*(3), 164–169. https://doi.org/10.1177/0963721412444727

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2011.05.004

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M., Oostenveld, R., Fries, P., … Schoffelen, J.-M. (2010). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data, FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience, Computational Intelligence and Neuroscience*, *2011, 2011*, e156869. https://doi.org/10.1155/2011/156869, 10.1155/2011/156869

Panichello, M. F., DePasquale, B., Pillow, J. W., & Buschman, T. J. (in press). Error-correcting dynamics in visual working memory. *Nature Communications*.

Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. *Nature Neuroscience*, 1. https://doi.org/10.1038/s41593-017-0003-2

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97–107. https://doi.org/10.1038/nrn1603

Pearson, J., & Keogh, R. (2019). Redefining Visual Working Memory: A Cognitive-Strategy, Brain-Region Approach. *Current Directions in Psychological Science*, 0963721419835210. https://doi.org/10.1177/0963721419835210

Pilat, D., & Fukasaku, Y. (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. *Data Science Journal*, *6*, OD4–OD11. https://doi.org/10.2481/dsj.6.OD4

Pochon, J.-B., Levy, R., Poline, J.-B., Crozier, S., Lehéricy, S., Pillon, B., … Dubois, B. (2001). The Role of Dorsolateral Prefrontal Cortex in the Preparation of Forthcoming Actions: An fMRI Study. *Cerebral Cortex*, *11*(3), 260–266. https://doi.org/10.1093/cercor/11.3.260

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review*, *83*(2), 157–171.

Postle, B. R. (2016). How Does the Brain Keep Information "in Mind"? *Current Directions in Psychological Science*, *25*(3), 151–156. https://doi.org/10.1177/0963721416643063

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 6–17. https://doi.org/10.1037/xhp0000302

Pratte, M. S., & Tong, F. (2014). Spatial specificity of working memory representations in the early visual cortex. *Journal of Vision*, *14*(3), 22–22. https://doi.org/10.1167/14.3.22

Prins, N., & Kingdom, F. A. A. (2009). Palamedes: Matlab routines for analyzing psychophysical data. *Http://Www.Palamedestoolbox.Org*.

Qi, X.-L., Meyer, T., Stanford, T. R., & Constantinidis, C. (2011). Changes in Prefrontal Neuronal Activity after Learning to Perform a Spatial Working Memory Task. *Cerebral Cortex*, *21*(12), 2722–2732. https://doi.org/10.1093/cercor/bhr058

Rademaker, R. L., Park, Y. E., Sack, A. T., & Tong, F. (2018). Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 925–940. https://doi.org/10.1037/xhp0000491

Rainer, G., Asaad, W. F., & Miller, E. K. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, *393*(6685), 577–579. https://doi.org/10.1038/31235

Rainer, G., Rao, S. C., & Miller, E. K. (1999). Prospective Coding for Objects in Primate Prefrontal Cortex. *The Journal of Neuroscience*, *19*(13), 5493–5505.

Ramkumar, P., Jas, M., Pannasch, S., Hari, R., & Parkkonen, L. (2013). Feature-Specific Information Processing Precedes Concerted Activation in Human Visual Cortex. *Journal of Neuroscience*, *33*(18), 7691–7699. https://doi.org/10.1523/JNEUROSCI.3905-12.2013

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of What and Where in the Primate Prefrontal Cortex. *Science*, *276*(5313), 821–824. https://doi.org/10.1126/science.276.5313.821

Rauss, K. S., Pourtois, G., Vuilleumier, P., & Schwartz, S. (2009). Attentional load modifies early activity in human primary visual cortex. *Human Brain Mapping*, *30*(5), 1723–1733. https://doi.org/10.1002/hbm.20636

Riggall, A. C., & Postle, B. R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *32*(38), 12990–12998. https://doi.org/10.1523/JNEUROSCI.1892-12.2012

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Ringach, D. L., Shapley, R. M., & Hawken, M. J. (2002). Orientation Selectivity in Macaque V1: Diversity and Laminar Dependence. *Journal of Neuroscience*, *22*(13), 5639–5651. https://doi.org/10.1523/JNEUROSCI.22-13-05639.2002

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, *35*(1), 83–92. https://doi.org/10.1016/j.intell.2006.05.004

Romo, R., Brody, C. D., Hernández, A., & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature*, *399*(6735), 470. https://doi.org/10.1038/20939

Rosanova, M., Casali, A., Bellina, V., Resta, F., Mariotti, M., & Massimini, M. (2009). Natural Frequencies of Human Corticothalamic Circuits. *The Journal of Neuroscience*, *29*(24), 7679–7685. https://doi.org/10.1523/JNEUROSCI.0445-09.2009

Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139. https://doi.org/10.1126/science.aah7011

Roux, F., & Uhlhaas, P. J. (2014). Working memory and neural oscillations: Alpha–gamma versus theta–gamma codes for distinct WM information? *Trends in Cognitive Sciences*, *18*(1), 16–25. https://doi.org/10.1016/j.tics.2013.10.010

Samaha, J., Sprague, T. C., & Postle, B. R. (2016). Decoding and Reconstructing the Focus of Spatial Attention from the Topography of Alpha-band Oscillations. *Journal of Cognitive Neuroscience*, *28*(8), 1090–1097. https://doi.org/10.1162/jocn_a_00955

Saproo, S., & Serences, J. T. (2010). Spatial Attention Improves the Quality of Population Codes in Human Visual Cortex. *Journal of Neurophysiology*, *104*(2), 885–895. https://doi.org/10.1152/jn.00369.2010

Sauseng, P., Klimesch, W., Gruber, W. R., Hanslmayr, S., Freunberger, R., & Doppelmayr, M. (2007). Are event-related potential components generated by

phase resetting of brain oscillations? A critical discussion. *Neuroscience*, *146*(4), 1435–1444. https://doi.org/10.1016/j.neuroscience.2007.03.014

Schneegans, S., & Bays, P. M. (2018). Drift in Neural Population Activity Causes Working Memory to Deteriorate Over Time. *Journal of Neuroscience*, *38*(21), 4859–4869. https://doi.org/10.1523/JNEUROSCI.3440-17.2018

Schneider, D., Mertes, C., & Wascher, E. (2016). The time course of visuo-spatial working memory updating revealed by a retro-cuing paradigm. *Scientific Reports*, *6*, 21442. https://doi.org/10.1038/srep21442

Scimeca, J. M., Kiyonaga, A., & D'Esposito, M. (2018). Reaffirming the Sensory Recruitment Account of Working Memory. *Trends in Cognitive Sciences*, *22*(3), 190–192. https://doi.org/10.1016/j.tics.2017.12.007

Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory. *Vision Research*, *128*, 53–67. https://doi.org/10.1016/j.visres.2016.09.010

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Serences, J. T., & Saproo, S. (2012). Computational advances towards linking BOLD and behavior. *Neuropsychologia*, *50*(4), 435–446. https://doi.org/10.1016/j.neuropsychologia.2011.07.013

Shafi, M., Zhou, Y., Quintana, J., Chow, C., Fuster, J., & Bodner, M. (2007). Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience*, *146*(3), 1082–1108. https://doi.org/10.1016/j.neuroscience.2006.12.072

Sharbrough, F., Chatrian, G. E., Lesser, R., Luders, H., Nuwer, M., & Picton, T. W. (1991). American Electroencephalographic Society Guidelines for Standard Electrode Position Nomenclature. *Journal of Clinical Neurophysiology*, *8*(2), 200.

Shen, G., Tao, X., Zhang, B., Smith, E. L., & Chino, Y. M. (2014). Oblique effect in visual area 2 of macaque monkeys. *Journal of Vision*, *14*(2), 3–3. https://doi.org/10.1167/14.2.3

Sigala, N., Kusunoki, M., Nimmo-Smith, I., Gaffan, D., & Duncan, J. (2008). Hierarchical coding for sequential task events in the monkey prefrontal cortex. *Proceedings of the National Academy of Sciences*, *105*(33), 11969–11974. https://doi.org/10.1073/pnas.0802569105

Soto, D., Hodsoll, J., Rotshtein, P., & Humphreys, G. W. (2008). Automatic guidance of attention from working memory. *Trends in Cognitive Sciences*, *12*(9), 342–348. https://doi.org/10.1016/j.tics.2008.05.007

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

Souza, A. S., Rerko, L., & Oberauer, K. (2014). Unloading and reloading working memory: Attending to one item frees capacity. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1237–1256. https://doi.org/10.1037/a0036331

Souza, A. S., Rerko, L., & Oberauer, K. (2016). Getting more from visual working memory: Retro-cues enhance retrieval and protect from visual interference. *Journal of Experimental Psychology. Human Perception and Performance*, *42*(6), 890–910. https://doi.org/10.1037/xhp0000192

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *Journal of Neuroscience*, *37*(27), 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017

Spitzer, B., & Blankenburg, F. (2012). Supramodal Parametric Working Memory Processing in Humans. *Journal of Neuroscience*, *32*(10), 3287–3295. https://doi.org/10.1523/JNEUROSCI.5280-11.2012

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707. https://doi.org/10.1016/j.neuron.2016.07.006

Sreenivasan, K. K., Curtis, C. E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82–89. https://doi.org/10.1016/j.tics.2013.12.001

Stokes, M. G. (2011). Top-down visual activity underlying VSTM and preparatory attention. *Neuropsychologia*, *49*(6), 1425–1427. https://doi.org/10.1016/j.neuropsychologia.2011.02.004

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron*, *78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Stokes, M. G., Saraiva, A., Rohenkohl, G., & Nobre, A. C. (2011). Imagery for shapes activates position-invariant representations in human visual cortex. *NeuroImage*, *56*(3), 1540–1545. https://doi.org/10.1016/j.neuroimage.2011.02.071

Stokes, M. G., & Spaak, E. (2016). The Importance of Single-Trial Analyses in Cognitive Neuroscience. *Trends in Cognitive Sciences*, *20*(7), 483–486. https://doi.org/10.1016/j.tics.2016.05.008

Stokes, M. G., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-Down Activation of Shape-Specific Population Codes in Visual Cortex during Mental Imagery.

The Journal of Neuroscience, *29*(5), 1565–1572. https://doi.org/10.1523/JNEUROSCI.4657-08.2009

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(10). https://doi.org/10.1167/13.10.9

Sugase-Miyamoto, Y., Liu, Z., Wiener, M. C., Optican, L. M., & Richmond, B. J. (2008). Short-Term Memory Trace in Rapidly Adapting Synapses of Inferior Temporal Cortex. *PLOS Comput Biol*, *4*(5), e1000073. https://doi.org/10.1371/journal.pcbi.1000073

Sussillo, D. (2014). Neural circuits as computational dynamical systems. *Current Opinion in Neurobiology*, *25*, 156–163. https://doi.org/10.1016/j.conb.2014.01.008

Sussillo, D., & Abbott, L. F. (2009). Generating Coherent Patterns of Activity from Chaotic Neural Networks. *Neuron*, *63*(4), 544–557. https://doi.org/10.1016/j.neuron.2009.07.018

Takeda, K., & Funahashi, S. (2004). Population Vector Analysis of Primate Prefrontal Activity during Spatial Working Memory. *Cerebral Cortex*, *14*(12), 1328–1339. https://doi.org/10.1093/cercor/bhh093

Taylor, P. C. J., Nobre, A. C., & Rushworth, M. F. S. (2007). FEF TMS Affects Visual Cortical Activity. *Cerebral Cortex*, *17*(2), 391–399. https://doi.org/10.1093/cercor/bhj156

Tiganj, Z., Cromer, J. A., Roy, J. E., Miller, E. K., & Howard, M. W. (2018). Compressed Timeline of Recent Experience in Monkey Lateral Prefrontal Cortex. *Journal of Cognitive Neuroscience*, *30*(7), 935–950. https://doi.org/10.1162/jocn_a_01273

Uluç, I., Schmidt, T. T., Wu, Y., & Blankenburg, F. (2018). Content-specific codes of parametric auditory working memory in humans. *NeuroImage*, *183*, 254–262. https://doi.org/10.1016/j.neuroimage.2018.08.024

van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J., & Olivers, C. N. (2018). Current and future goals are represented in opposite patterns in object-selective cortex. *ELife*, *7*. https://doi.org/10.7554/eLife.38677

Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding Sound and Imagery Content in Early Visual Cortex. *Current Biology*, *24*(11), 1256–1262. https://doi.org/10.1016/j.cub.2014.04.020

Wallis, G., Stokes, M., Cousijn, H., Woolrich, M., & Nobre, A. C. (2015). Frontoparietal and Cingulo-opercular Networks Play Dissociable Roles in Control of Working Memory. *Journal of Cognitive Neuroscience*, *27*(10), 2019–2034. https://doi.org/10.1162/jocn_a_00838

Walther, A., & van den Bosch, J. J. F. (2012). FOSE: A framework for open science evaluation. *Frontiers in Computational Neuroscience*, *6*. https://doi.org/10.3389/fncom.2012.00032

Wang, X.-J. (2001). Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences*, *24*(8), 455–463. https://doi.org/10.1016/S0166-2236(00)01868-3

Watanabe, K., & Funahashi, S. (2007). Prefrontal Delay-Period Activity Reflects the Decision Process of a Saccade Direction during a Free-Choice ODR Task. *Cerebral Cortex*, *17*(suppl 1), i88–i100. https://doi.org/10.1093/cercor/bhm102

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. https://doi.org/10.1038/nn.3667

Watanabe, Y., Takeda, K., & Funahashi, S. (2009). Population Vector Analysis of Primate Mediodorsal Thalamic Activity during Oculomotor Delayed-Response Performance. *Cerebral Cortex*, *19*(6), 1313–1321. https://doi.org/10.1093/cercor/bhn170

Wessel, J. R., & Aron, A. R. (2017). On the Globality of Motor Suppression: Unexpected Events and Their Influence on Behavior and Cognition. *Neuron*, *93*(2), 259–280. https://doi.org/10.1016/j.neuron.2016.12.013

Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, *17*(3), 431–439. https://doi.org/10.1038/nn.3645

Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience*, *9*. https://doi.org/10.3389/fnsys.2015.00123

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871. https://doi.org/10.1038/nn.4546

Wolff, M. J., Kandemir, G., Stokes, M. G., & Akyurek, E. G. (2019). Impulse responses reveal unimodal and bimodal access to visual and auditory working memory. *BioRxiv*, 623835. https://doi.org/10.1101/623835

Worden, M. S., Foxe, J. J., Wang, N., & Simpson, G. V. (2000). Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *20*(6). Retrieved from https://einstein.pure.elsevier.com/en/publications/anticipatory-biasing-of-visuospatial-attention-indexed-by-retinot-2

Wutz, A., Loonis, R., Roy, J. E., Donoghue, J. A., & Miller, E. K. (2018). Different Levels of Category Abstraction by Different Dynamics in Different Prefrontal Areas. *Neuron*, *97*(3), 716–726. https://doi.org/10.1016/j.neuron.2018.01.009

Xu, Y. (2017). Reevaluating the Sensory Account of Visual Working Memory Storage. *Trends in Cognitive Sciences*, *21*(10), 794–815. https://doi.org/10.1016/j.tics.2017.06.013

Xu, Y. (2018). Sensory Cortex Is Nonessential in Working Memory Storage. *Trends in Cognitive Sciences*, *22*(3), 192–193. https://doi.org/10.1016/j.tics.2017.12.008

Zanos, S., Rembado, I., Chen, D., & Fetz, E. E. (2018). Phase-Locked Stimulation during Cortical Beta Oscillations Produces Bidirectional Synaptic Plasticity in Awake Monkeys. *Current Biology*, *28*(16), 2515-2526.e4. https://doi.org/10.1016/j.cub.2018.07.009

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Zhang, W., & Luck, S. J. (2009). Sudden Death and Gradual Decay in Visual Working Memory. *Psychological Science*, *20*(4), 423–428. https://doi.org/10.1111/j.1467-9280.2009.02322.x

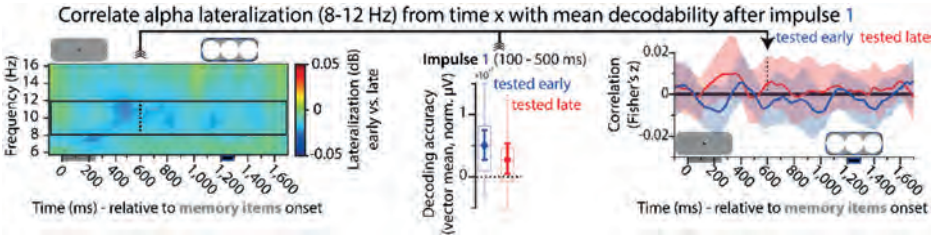
Zhou, Y. D., & Fuster, J. M. (1996). Mnemonic neuronal activity in somatosensory cortex. *Proceedings of the National Academy of Sciences*, *93*(19), 10533–10537. https://doi.org/10.1073/pnas.93.19.10533

Zokaei, N., Manohar, S., Husain, M., & Feredoes, E. (2014). Causal Evidence for a Privileged Working Memory State in Early Visual Cortex. *Journal of Neuroscience*, *34*(1), 158–162. https://doi.org/10.1523/JNEUROSCI.2899-13.2014

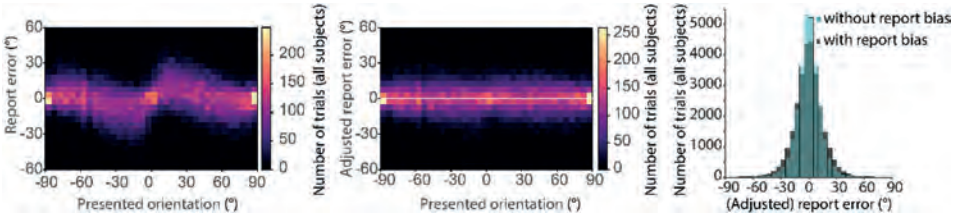
Zokaei, N., Ning, S., Manohar, S., Feredoes, E., & Husain, M. (2014). Flexibility of representational states in working memory. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00853

Zucker, R. S., & Regehr, W. G. (2002). Short-Term Synaptic Plasticity. *Annual Review of Physiology*, *64*(1), 355–405. https://doi.org/10.1146/annurev.physiol.64.092501.114547

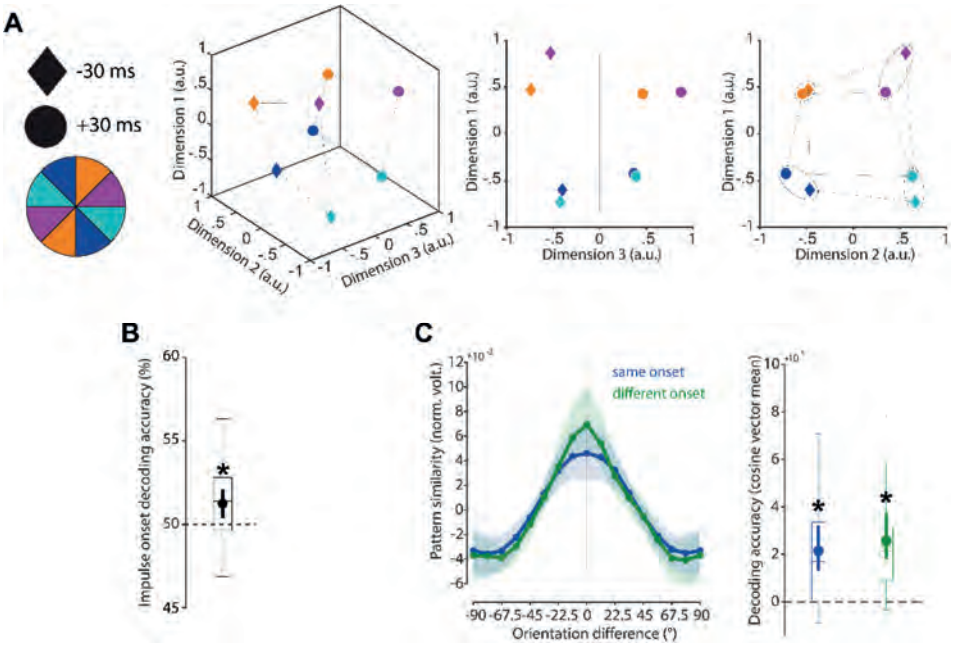
Appendix



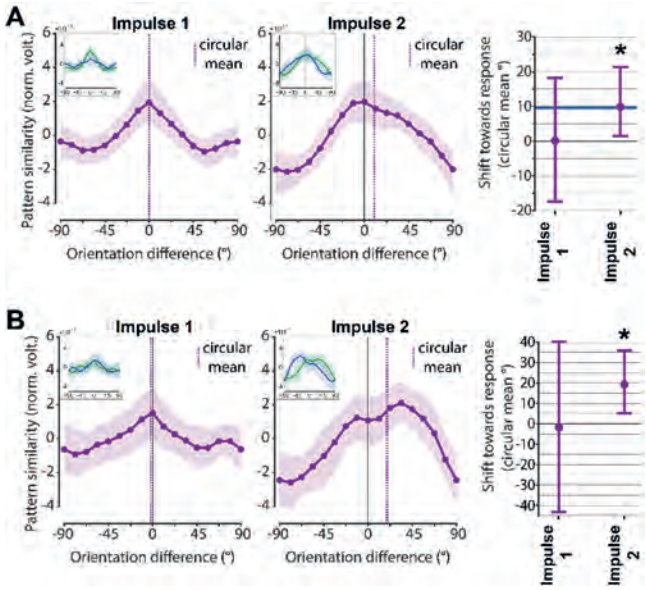
Supplemental figure 4.1. Testing the relationship between alpha-lateralization and item decoding after the first impulse in Experiment 2. Both attended and unattended memory items were decodable after the first impulse in Experiment 2; however, it remains possible that participants sometimes attended to the less-relevant item, contributing to decoding on some trials. To consider this possibility, we test whether the impulse-specific WM item decoding after impulse 1 presentation covaries with trial-wise fluctuations in spatial attention. Spatial attention was indexed by alpha-power lateralization relative to the location of the early-tested item of each time-point (left, also see Figure 4C and corresponding results), and trial-wise item decodability was estimated 100-500ms after impulse 1 onset (middle panel). The correlation time-course (right), where each time-point represents the mean correlation of the averaged item decoding (100 – 500 ms after impulse 1) with the alpha-lateralization of that time-point, shows no evidence for a relationship between item decoding and alpha-lateralization for any time-point (permutation test, $n = 19$, early-tested item: all $p > 0.058$; late-tested item: all $p > 0.148$, uncorrected). Therefore, we find no evidence that the impulse-response varies with the focus of attention, even on a trial-wise basis. Error shadings are 95% C.I. of the mean. Circles and error bars superimposed on the boxplots represent mean and 95% C.I. of the mean.



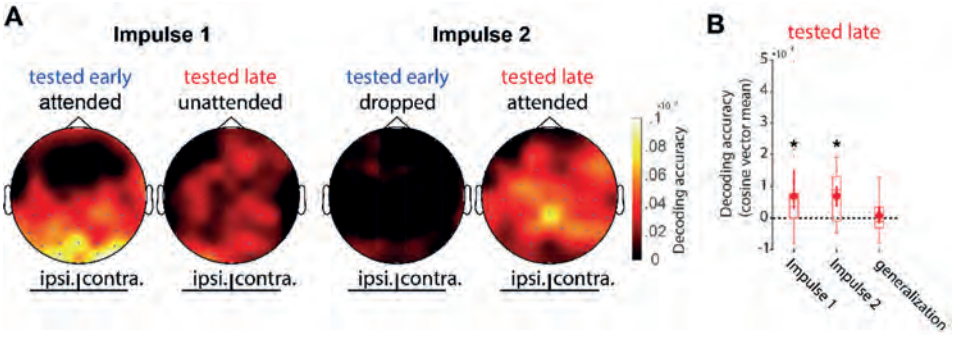
Supplemental figure 6.1. Report-bias of orientations. Participants showed a bias, exaggerating the tilt of oblique orientations, manifesting itself as a repulsion form the cardinal axes (0 and 90 degrees; *left*), similar to previous reports (Pratte et al., 2017). To ensure an unbiased estimate of a possible shift in our analysis, and to isolate random from systematic errors, the report bias was removed by subtracting the median error within 11.25 degree orientation bins (*middle*). By removing orientation-specific error, the resulting error distribution is narrower (*right*). Clockwise and counter-clockwise reports were defined as positive and negative reports relative to this "adjusted", unbiased, report error.



Supplemental figure 6.2. Cross-generalization of coding scheme between impulse onsets in reanalyses of Wolff et al. (2015). **(A)** Visualization of orientation and impulse-onset code in state-space. The third dimension discriminates between impulse-onsets. The first and second dimensions code the orientation space in both impulses. **(B)** Trial-wise accuracy (%) of impulse-onset decoding. **(C)** Orientation decoding within each impulse-onset (blue) and orientation code cross-generalizing between impulse-onsets (green). Error shadings and error bars are 95 % C.I. of the mean. Centre lines of boxplots indicate the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range. Extreme values are shown separately (dots). Asterisks indicate significant decoding accuracies or cross-generalization ($p < 0.05$).



Supplemental figure 6.3. Within impulse training and testing to estimate drift. **(A)** Response-dependent averaging of trial-wise tuning curves (Fig. 6.6A). Shift towards response: Impulse 1: p = 0.492; Impulse 2: p = 0.022, one-sided. **(B)** Response-dependent training and testing (Fig. 6.7A). Shift towards response: Impulse 1: p = 0.545; Impulse 2: p = 0.009, one-sided. Same convention as Fig. 6.6B-C and Fig. 6.7B-C.



Supplemental figure 7.1. Reanalyses of Chapter 3. **(A)** Searchlight analysis of decoding topography using each channel, and its 2 closest neighbours, separately (as in Chapter 6). **(B)** Decoding of tested-late item in experiment 2 in Chapter 3 when it is unattended at impulse 1, attended at impulse 2, and the cross-generalization between impulse 1 and impulse 2 (training on 1, testing on 2 , and vice versa). Time-window was taken from 100 to 400 ms relative to impulse onset, using the dynamic decoding approach used in Chapters 5 and 6.

Nederlandse samenvatting

translated from English by

Jasper Hajonides van der Meulen

Het onthouden en transformeren van informatie zonder perceptuele input, om voor te bereiden op toekomstige acties, is een fundamentele eigenschap van het (menselijke) brein. Sinds 1971, toen Fuster & Alexander (1971) de eerste werkgeheugen (WG) cellen beschreven, is er een voordurende discussie over de neurale basis van werkgeheugen en hoe informatie precies opgeslagen wordt. In de gebruikelijke onderzoeksopzetten wordt gekeken hoe de WG condities of WG inhoud gerepresenteerd worden in de geheugenperiode. Het is de consensus geweest dat onafgebroken neurale activiteit de informatie 'online' houdt. Dit werd geconcludeerd aan de hand van een hoeveelheid studies die dit suggereert maar ook gezien het feit dat meetbare neurale activiteit noodzakelijk is voor hersenscan onderzoek; het vinden van geen neurale activiteit tijdens WG is praktisch gezien een nul resultaat en is lastig te interpreteren.

Recentelijk is echter het idee dat WG niet afhankelijk is van een onaangebroken golf van neurale activiteit steeds meer populair geworden. Studies hebben aangetoond dat als aandacht niet naar de WG objecten gericht wordt er weinig tot geen neurale activiteit te meten is (Larocque et al., 2014; Lewis-Peacock et al., 2011; Watanabe & Funahashi, 2014) wat suggereert dat WG alleen het actief herhalen van informatie reflecteert. Verder is er gesuggereerd dat schijnbaar onafgebroken WG activiteit tijdens een geheugen periode een artefact is van het middelen over verschillende repetities en dat WG in werkelijkheid wordt gemoduleerd door korte vlagen van activiteit afgewisseld door activiteits-loze perioden (Lundqvist et al., 2016), die overbrugt worden door tijdelijke veranderingen in de connectiviteit tussen neuronen (Mongillo et al., 2008; Stokes, 2015).

In deze doctorale scriptie wordt het gebruikelijke beeld van onafgebroken neurale activiteit tijdens de geheugenperiode getest. De hypothese is dat de neurale reactie een interactie is op externe stimulatie en de huidige staat van het netwerk. Dit is te vergelijken met een sonarsysteem waar de echo de stimulatie van verborgen structuren teweeg brengt. In deze scriptie testen we deze hypothese en gebruiken we de methode om WG te onderzoeken. In verschillende studies wordt aangetoond dat de neurale 'impuls' reactie op een externe 'ping' gebruikt kan worden om verborgen WG herinneringen.

Hoofdstuk 2 demonstreert dat magnetoencephalografie (MEG) rijke spatiele informatie bevat, wat de oriëntatie kolommen in de vroege visuele cortex betreft (Cichy et al., 2015). Dit impliceert dat MEG een belangrijke speler kan zijn op het gebied van multivariate methoden die gebruik maken van het gehele spatiele patroon van activatie om 'uit te lezen' welke informatie op dat moment wordt waargenomen, verwerkt, of onthouden wordt in het brein. Dit laat zien dat zelfs informatie uit dezelfde kleine brein regio tot

een heel ander neuraal patroon kan leiden wanneer we activiteit meten over de gehele schedel. Hetzelfde geldt voor EEG, een bijna net zo robuuste multivariate techniek als MEG (Cichy & Pantazis, 2017).

In hoofdstuk 3 wordt niet alleen aangetoond dat EEG daadwerkelijk sensitief is wanneer het komt tot het uitlezen van oriëntaties van zogenoemde Gabor patches maar het principe van een neurale 'ping' wordt voor het eerst beschreven. Door middel van tijdpunt-bij-tijdpunt analyses van het EEG signaal kan onderscheid gemaakt worden tussen verschillende oriëntaties van gepresenteerde Gabor gratings. Verder onthult de presentatie van een neurale 'impuls' stimulus tijdens de geheugenperiode een neuraal signaal dat onderscheid maakt tussen de verschillende oriëntaties. Dit is bewijs voor de hypothese dat de neurale impuls de huidige staat van het WG netwerk representeert, in dit geval het nabeeld van de voorheen gepresenteerde oriëntatie.

Hoofdstuk 4 is een uitbreiding op hoofdstuk 3 en laat in een retro-cue design zien dat de impuls daadwerkelijk geheugeninhoud laat zien. Dit suggereert dat irrelevante informatie verwijderd kan worden van werkgeheugen en geen herleidbaar spoor achter laten, wat de impuls betreft. In een tweede experiment laat hoofdstuk 4 zien dat zowel items in geheugen waar aandacht wel of niet op gericht wordt uitgelezen kunnen worden na een impuls. Dit levert bewijs voor de hypothese dat WG staat alle relevante informatie bevat, afhankelijk van aandacht. In beide experimenten van dit hoofdstuk voorspelde de trial-bij-trial uitleesbaarheid van WG inhoud na de impuls de precisie van de geheugen taak.

Hoofdstuk 5 laat zien dat de impuls techniek ook toepasbaar is in het auditieve domein. Als een neutrale auditieve stimulus gepresenteerd wordt tijdens de geheugen periode resulteert dat in een neurale respons die de toon in geheugen reflecteert. Dit spiegelt de resultaten in het visuele domein en indirect impliceert dit dat de sensorische cortex betrokken zijn in het onthouden van informatie in WG (Christophel et al., 2017; Kumar et al., 2016; Serences, 2016). Verder wordt aangetoond dat een cross-modale (bijv. auditief) impuls tijdens visuele WG perioden niet inhoud specifiek is. Dit suggereert dat het visuele WG netwerk los staat van de brein regios die de auditieve informatie verwerken en dat bottom-up auditieve reacties het niet verstoren. Vice-versa is dit echter niet het geval; visuele gebieden zijn wel betrokken bij het herinneren van auditieve informatie. Dit kan direct zijn, door visualisatie, of indirect, door inhoud-specifieke connectiviteit in het auditieve WG netwerk.

Hoofdstuk 6 gebruikt de impuls aanpak om de relatie tussen neurale WG representatie en gedragsmatige variabele te onderzoeken. Dit levert neurofysiologisch bewijs voor de hypothese dat neurale representaties of continue WG concepten (bijv. oriëntaties of locaties) willekeurig variëren over tijd, wat resulteert in inaccurate gedragsmatige data. De impuls respons kon deze verschuiving van de neurale representatie van een oriëntatie in WG tegen het einde van de trial en de bijkomstige gedragsmatige reacties voorspellen. Verder werd gevonden dat het brein informatie contant representeert, zelfs wanneer de neurale reacties op de impulsen erg variëren over

de geheugenperiode. Dit suggereert dat de *coding schemes* stabiel zijn, ongeacht neurale dynamica die varieert over tijd (Murray et al., 2017).

Kortom, deze thesis laat de bruikbaarheid zien van een relatief simple 'ping' aanpank om voorheen verborgen neurale WG staten te onderzoeken.

Acknowledgments

First off, I would like to thank my day-to-day supervisor Elkan Akyürek, for enabling me to do my PhD and for the guidance and wisdom you have offered throughout it. You have been incredibly supportive and patient over these years. I appreciate the extreme flexibility and openness you have shown not only towards new research ideas but also towards my personal situation, as well as your practical and professional approach towards science in general.

Likewise, I would like to thank my promoter Monicque Lorist for the continued support and dedication to the successful completion of my thesis.

Of course, I want to thank my collaborator Mark Stokes, without whom this thesis would have looked completely different! Thank you very much for making me a member of your lab and for the extensive collaboration, great new research ideas, mentorship and general support throughout my PhD.

Thank you to all my collaborators in Oxford, Groningen, and elsewhere, all of whom played a major role in the completion of this thesis. Thanks to Eelke Spaak and Nick Myers, for always being ready to help and always being able to give smart advice, no matter the question or problem. Thank you, Janina Jochim, for your incredible help and effort. Thank you, Güven Kandemir, for your diligence and going above and beyond! Thank you, Jacqueline Ding, for having been a bright and motivated undergrad that I got to co-supervise. And thank you Tim Buschman for a fruitful collaboration across the pond. Thank you also to my undergraduate supervisor, Candice Morey, for sparking my interest in working memory research.

I would also like to thank the whole Experimental Psychology Department in Groningen, which has a great atmosphere, and has generally been a great place to work. In particular Rob, Aytaç, and Edyta (thank you for your support, patience, and for keeping me sane), my office mates Michael and Florian (thank you for always being helpful) as well as all other fellow PhD students: Nadine, Berry, Wisnu, Atser, Jefta, Sanne, Tineke, Erik and Marlon as well as Steffen, Wanja, and Nico.

I also want to thank my friends and colleagues in Oxford, who made me feel at home whenever I visited. Thank you, Maryann and Ben for showing me the ropes in my early days in Oxford. And thank you to Dante, Frida, Sam, Ilenia, Freek, Paul, Darinka, Jasper, Andrea, Lev, Emilia, Edwin, Michal, Alex, Sofia, and all former lab members for the great camaraderie and for providing such a stimulating environment.

Thank you to my old friends, Jul, Nick, and Anna-Sophia for staying in touch for all these years.

Thank you to my family, especially my parents and my siblings, Bettina and Alexander, for the support and interest in my research, and my grandmother for cheering me on. And of course, thank you to Helen, who has been nothing but supportive and understanding for the entirety of my PhD.

Publication List

Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2020). Drifting codes within a stable coding scheme for working memory. PLOS Biology, 18(3), e3000625.

Wolff, M. J., Kandemir, G., Stokes, M. G., & Akyürek, E. G. (2020). Unimodal and Bimodal Access to Sensory Working Memories by Auditory and Visual Impulses. Journal of Neuroscience, 40(3), 671–681.

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. Nature Neuroscience, 20(6), 864–871.

Akyürek, E. G., & **Wolff, M. J.** (2016). Extended temporal integration in rapid serial visual presentation: Attentional control at Lag 1 and beyond. Acta Psychologica, 168, 50–64.

Stokes, M. G., **Wolff, M. J.**, & Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution. Trends in Cognitive Sciences, 19(11), 636–638.

Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. Frontiers in Systems Neuroscience, 9.

Wolff, M. J., Scholz, S., Akyürek, E. G., & van Rijn, H. (2015). Two visual targets for the price of one? Pupil dilation shows reduced mental effort through temporal integration. Psychonomic Bulletin & Review, 22(1), 251–257.

Mall, J. T., Morey, C. C., **Wolff, M. J.**, & Lehnert, F. (2014). Visual selective attention is equally functional for individuals with low and high working memory capacity: Evidence from accuracy and eye movements. Attention, Perception, & Psychophysics, 76(7), 1998–2014.