Neural effects of prioritisation on working memory

Sammi Ramdane Chekroud
St. Cross College, Oxford

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford

DPhil Thesis, Trinity Term, 2022

Neural effects of prioritisation on working memory

Sammi Ramdane Chekroud

St. Cross College, Oxford Trinity Term, 2022

Abstract

Working Memory describes the cognitive ability to store information over short timescales, and use it to guide behaviour. It is strikingly limited, and we are only able to maintain small amounts of information in this short-term store. However, it is highly flexible and we are able to prioritise working-memory contents according to their relevance for behaviour. Such attentional prioritisation is associated with enhancements in working-memory performance. However, whether we use memories to guide behaviour should depend upon their quality and value. Thus, it is important to understand how we introspect upon working memory, evaluating and learning about its quality, and assign value to its contents. This thesis explores uncertainty and value in working memory.

Chapter 2 examines how attentional prioritisation affects working-memory introspection. I show that attentional prioritisation improves the quality of introspection into working-memory contents. I also show evidence that feedback regarding the accuracy of recent confidence judgements is incorporated into working-memory confidence, shaping our beliefs about the quality of working-memory performance.

Chapter 3 describes two experiments using electroencephalography (EEG) to explore how feedback regarding working-memory performance is processed. I show that uncertainty – manipulated externally by varying working-memory load, and carefully measured through subjective confidence judgements – is represented in the neural response to feedback. When feedback contains information regarding the quality of subjective uncertainty, I find evidence of a confidence prediction error that could facilitate learning about working-memory quality.

Chapter 4 describes two novel experiments that explore how the value associated with objects held in working memory shapes performance. These experiments present reward information during working-memory maintenance to examine how reward influences the prioritisation of working-memory contents after encoding. I show that reward can guide the prioritisation of working-memory contents – even when reward is irrelevant to the task at hand. I demonstrate both global and item-specific benefits associated with reward-guided attentional prioritisation of working-memory contents.

Finally, in Chapter 5, I discuss the findings of this thesis in a wider context, and remaining theoretical questions related to working-memory guided decisions.

This thesis was funded by the Glyn Humphreys Scholarship, from the Department of Experimental Psychology, University of Oxford.

Approximate word count: 40,000 words.

Acknowledgements

There are a lot of people to thank for helping me get to where I am now, and to the end of this DPhil. It's hard to thank them all, but important to try – so bear with me.

Firstly, I'd like to express deep gratitude to Kia Nobre. Over 10 years ago, you took a chance on 16-year-old me, offering me a place to study Experimental Psychology at Oxford. Three years later, you took another chance on me and hired me as a Research Assistant. Three years after that, I was lucky enough to be accepted to do my PhD under your supervision. Since leaving home, my academic journey has been a series of doors opened by you that have allowed me to grow personally and intellectually, exploring new ideas and doing cool new things. For the series of opportunities, personal and professional guidance you have provided me, and the time and kindness you have dedicated to me and my growth over the last decade, 'thanks for literally everything' is largely inadequate, but hopefully captures the essence of it. You have shaped the way I see the brain and behaviour, something I am immensely grateful for.

Secondly, I owe a huge thanks to Nils Kolling. You were my shining light in the darkness that is GLM design matrices and model-based EEG analyses. You introduced me to the intrigue of introspection, guiding me through it and its importance. When life got hard you were patiently there to help me get back on track. Sometimes I left our meetings not knowing whether my questions were answered, but the questions I had afterwards were always much cooler than the ones I went in with. Thank you for all of your support.

Over the last 10 years, I've had the privilege of working with many of the Brain and Cognition Lab. Some particularly stand out for shaping my journey. I owe a debt to Freek van Ede, Nick Myers and Mark Stokes. I had the fortune of working with you as an undergrad and RA, experiencing science with you all. I invariably wouldn't be here without your help along the way. Andrew Quinn also deserves a special mention. Your GLM toolbox was the workhorse of my EEG analyses in Chapter 3, and helped me avoid MATLAB for 4 years. Beyond your ninja python skills, I will forever remember our evenings playing pool, the BBQ chicken wings and the support you gave me from the Star. Sage and Jasper were constant supports when needed and always there to help, I'm very lucky to have had you guys with me. Dejan, Nahid, Zita, Katharina, Ryszard, Nora, Giedré, Jemma, Lauren and many others have made B&C an excellent environment as a grad student.

I was lucky enough to be awarded the Glyn Humphreys Scholarship for my DPhil. It was an honour to receive this as Glyn was Head of Department when I was a fresh-faced undergrad. I'd also like to thank all the participants who took part in my studies – my research was only completed thanks to their time.

Outside science, I was lucky enough to be supported by lots of great friends. I've had great times with, and help from, others over the last four years who deserve thanks: Sam Rowan, Youssuf Saleh, Eri Ichijo, the lads from LMH M1 2017-2019, Sorcha Bolton, and more. Thanks for stopping me from being a total hermit, you guys are all great.

Alex Board is the best friend anybody could ever have. I'm glad you took that RA job all those years ago, I'm not sure what I'd do without your voice notes. Thanks for being you.

Almost 7 years ago I was lucky enough to join another family. Penny and Henning have been incredibly supportive throughout this process. Thank you for always trying to get me to switch off and relax – even when I didn't appreciate it at the time. Yours was a home away from home when I needed to get away, especially during covid – thank you.

I wouldn't be where I am today if it weren't for my family. I was given great educational opportunities by my parents, Djamel and Sarah - a privilege that enabled me to start and finish this thesis. Mum is always at the end of the phone whether it's good or bad, patiently helping me get through whatever bothers me. You're the best. I'm also lucky enough to have two of the most annoying brothers in existence. Adam inspires me to be better in most walks of life, and has been my hidden guide along academia so far. Ameen is somehow the absolute best and absolute worst at the same time, as only a twin can be. I couldn't ask for better brothers and supports in my life.

Finally, Megan. 'Thanks' doesn't really cut it. You've been with me through the highs and lows – enjoying them and supporting me through them. I'm sorry for being such a big grump these last few years, but thanks for helping me see everything more clearly along the way. You make me a better person and make life brighter. Thank you for everything, you deserve all the stickmen in the world.

Statement of Contribution

Science is collaborative by nature. Thus, so is this thesis. This work was conducted together with colleagues both in and outside of my supervisory team. Chapter 2 of this thesis was designed with Nils Kolling and Kia Nobre. I collected the data for this experiment and ran the analysis. The first experiment in Chapter 3 was designed by Frederik van Ede and Kia Nobre to investigate non-feedback aspects of working memory. I collected the data for this experiment as a Research Assistant before beginning my DPhil, and I conducted the re-analysis of this data during the Coronavirus pandemic. The second experiment in Chapter 2 was designed in collaboration with Nils Kolling and Kia Nobre. I collected the data for this experiment and led the analysis for this with Nils Kolling and Kia Nobre. It is also of note that the analysis of EEG data in this thesis was undertaken using (at the time unpublished) analytical tools developed by Andrew Quinn. I designed the online experiments in Chapter 4 with Nils Kolling and Kia Nobre. I programmed these experiments to collect the data, and led the analysis for these experiments. The writing of this thesis, while my own, was carried out under the expert supervision of Kia Nobre and Nils Kolling.

1 - General Introduction

1.1 General Overview

Surrounding us at all times is an incredibly complex environment. When you take the time to think about it, it is remarkable that the fleshy organ in our head can make sense of the sheer amount of input it receives. In the process of living, not only do we experience life in the present, but snippets (if not more) make their way into our memory. Anecdotally, one can intuit how memories of the past can shape actions in the future. For example, my mother disliked Brussel Sprouts as a child during Christmas dinner. As a result, she refused to make them for Christmas later in her life. Previous experiences of bony fish lead me to avoid certain fish when dining.

These are two simple examples of how long-term memories might guide ongoing behaviours. Long-term memory, however, is not the focus of this thesis. One cannot think about memory-guided behaviour without immediately going to the intuitive long-term store of our lived experience. However, memory has multiple timescales. Thus, different types of memories could guide adaptive behaviours. I just mentioned two ways in which long-term memories of experiences could guide action. Information stored over much shorter time periods – held in Working Memory – could also guide action. If you are queuing in a café and see a waiter drop the piece of cake you were planning on ordering, you may change your order to avoid eating gritty cake.

This, however, is a mundane example. The world is dynamic, and our cognition and abilities are dynamic. In a recent conversation, a music teacher, who asked about the focus of my research, gave me insight into her own experience of memory-guided behaviour from her time as a professional harpist. She described endless time practising to learn a piece, putting in the hours to engrave the music into her memory. She then described performing as a dynamic process – her hands plucking strings while, in the back of her mind, the sheet music moves on, held in her visual imagination. As she progressed through the piece, the sheet music moved, allowing her to keep track of where she was. Even more, imagining where she *would* be in the music enabled her to anticipate her future hand movements, preparing for the next bar or section of the piece to perform without mistakes.

Her description struck me as an evocative example of the complexity of both memory-guided action and memory-guided anticipation. It captures many aspects of what memory truly is under an embedded-components, or nested-processes, framework such as those proposed by Nelson Cowan or Klaus Oberauer (e.g. Cowan, 2001; Oberauer, 2009). In such models, the interplay between the types of memory is emphasised. Long-term memories hold all the information and experience we have. Within this long-term store, there is the *activated long-term memory* – the long-term memories that could be relevant for behaviour. Information in activated long-term memory holds a privileged state, more active than irrelevant memories, for supporting goal-directed behaviour. Condensing down, there is what Cowan calls the *focus of attention* – a limited-capacity store of information. We know this more commonly as "*working memory*."

More recently the focus of attention concerning working memory has a different meaning. It describes how selective attention operates upon specific objects or items held in working memory, marking them as more task-relevant. It allows the prioritisation of one (or a small subset) of the contents held in working memory to allow them to guide behaviour. This interaction between attention and working memory is characterised across tasks by improvements in behavioural performance (Griffin & Nobre, 2003; Landman et al., 2003; Myers et al., 2018; Niklaus et al., 2017; Souza, Rerko, Lin, et al., 2014) and modulation in cortical brain activity (Kuo et al., 2012; Myers et al., 2015; Nobre et al., 2004; van Ede et al., 2017). Attention within working memory will be discussed in more depth in a later section.

One thing seems clear - our memories of the past are not merely representations that are stored without purpose. Instead, they are maintained to support our ongoing, and future, behaviour. However, the extent to which our memories (working or long-term) guide behaviour, may depend on how well we remember them. If I'm looking for my keys while trying to rush out of the house in the morning, being confident that I did *not* leave my keys in the kitchen shapes the way I search my house. If I remember clearly that I left them on my bedside table, I will search there first. The way we introspect upon our memories can shape the way we use them in the present.

But introspection is not always perfect. We can be confident in remembering something and be wrong. I might believe I remember my cousin's birthday and miss her actual birthday because of my error. As a result, it is important not only to introspect upon

our memory but also to learn about our accuracies and errors to properly decide when we should allow certain memories to guide what we do. Learning that my memory is generally quite bad might discourage me from relying upon what I think I remember and instead ask a family member to check the birthday date. I may, as a result of knowing whether my memory is highly fallible, rely more upon external sources of information rather than the information I hold myself.

The empirical research in this thesis focuses on trying to understand the 'working' in working memory. Specifically, there are two main themes: The first theme explores introspection: I have investigated how we introspect on the contents of our working memory and how we learn about its quality through feedback (Chapter 2). I use electroencephalography (EEG) to understand how we represent working-memory uncertainty during feedback processing (Chapter 3). The second theme focuses on reward as a possible dimension of selective attention that can act upon the contents of working memory. I look at how rewards guide the prioritisation of objects during working-memory maintenance. I consider if reward-based cues act independently of, or synergistically with, identity-based attentional retrocues (Chapter 4).

Before reporting on my empirical work, I will start with an overview of a few topics to contextualise this research. Firstly, I will review how research has focused on representational theories of working memory, before exploring the flexibility of this store as demonstrated by attention and working-memory interactions. Secondly, I will discuss the influence of reward on behaviour. I will discuss how reward shapes the way we interact

with the world around us and make decisions as well as learn through feedback. I will then explore what we already know about how reward affects attentional processes and affects working memory.

1.2 Working Memory

Over the years, cognitive psychologists have focused on understanding the structure of working memory as a system (e.g. Bays et al., 2009; Brady et al., 2011; Cowan, 2001; Luck & Vogel, 1997; Ma et al., 2014). A general capacity limit is widely acknowledged (e.g., Cowan, 2001), although disputes continue about finer details. For example, the underlying cause of this limit has been debated, and psychologists have aimed to understand the structure of the working-memory representation and why we can only maintain so little content within this short-term store.

1.2.1 Representational theories of working memory

One foundational theory of working-memory structure posits that working memory consists of fixed 'slots' in which integrated objects can be stored (e.g. Luck & Vogel, 1997; Zhang & Luck, 2008). It argues that a fixed number of objects can be encoded into working memory. When under capacity, all of the to-be-encoded items can be stored, and remembered faithfully. When working-memory load surpasses capacity, only a subset of stimuli enters into working memory.

This ‘slot’ structure of memory is likely to have been shaped by the types of tasks - more specifically, the types of responses or actions - that were used to understand the format of (visual) working memories. The seminal paper by Luck and Vogel (1997) used a change-detection task that contained discrete stimuli (coloured squares or shapes) in an encoding array. Participants were required to report whether the original stimulus array was different to a test array (Luck & Vogel, 1997). Given the discrete nature of the task, a discrete structure of working memory would be intuitive.

This slot-based model, based on discrete tasks, was challenged by tasks that used continuous reports (e.g. Bays et al., 2009; Bays & Husain, 2008; Gorgoraptis et al., 2011; Wilken & Ma, 2004). In precision-recall tasks, the stimulus arrays are similar to change-detection tasks, but the response requirements differ. Rather than reporting a change in a test array, participants were required to report an exact feature of an item from working memory (e.g. the item's precise orientation or colour). Rather than simply reporting differences, participants are required to actively use the quality of information held in working memory to deliver a response. This task led to a new theory of working-memory representation: rather than a fixed number of slots, a limited pool of ‘resources’ is allocated across the array of items to be remembered. As the number of to-be-remembered items increases, resources are distributed across a larger array, resulting in lower precision of the maintained stimuli.

Alternative models of working-memory representation focus less on cognitive frameworks (slots vs resources in working memory) and more on the neurobiology of

memory representations. These theories of visual working memory propose that the brain encodes a noisy representation of the encoded item (e.g. Bays, 2014; Matthey et al., 2015; Schneegans & Bays, 2017; Wilken & Ma, 2004) via population coding. The retrieval of the encoded information occurs through readouts of probabilistically active neuronal populations. Such models recreate a variety of performance patterns related to working-memory recall. Notably, these models attempt to explain the phenomenon of *'misbinding'* – how humans often misbind features together in memory and report features of non-probed items. While concerned with the representation of working memories, these computational models also attempt to understand human behaviour by modelling the retrieval of information. Such population-coding models explain variability in working-memory performance due to noise in maintained representations and processes of retrieval. These computational models based on neurobiological principles encourage the consideration of working memory as more than a store.

1.2.2 Flexibility of working memory: dimensions of attention

Conceptualising working memory as a store of information appears, on the face of it, to imply some rigidity. It suggests that once something is stored in memory, it is fixed - whether this representation is noisy or precise, and whether resources are allocated in a discrete or distributed way. Instead, working memory is incredibly flexible. Indeed, it needs to be so to guide behaviour adaptively. The environment around us changes over time, and our goals change as a function of the task at hand. In order to support effective behaviour,

working memory *should* be flexible enough to allow us to adapt. The operation of selective attention upon working-memory contents has highlighted this flexibility.

Since the early 2000s, cognitive psychologists have demonstrated that retrospective attentional cues, termed 'retrocues', can operate upon the contents of working memory (Griffin & Nobre, 2003; Landman et al., 2003). In retrocue tasks, an attentional cue is presented during working-memory maintenance, orienting the individual towards a specific object held in memory by marking it as more useful for behaviour compared to other maintained memoranda. The behavioural consequences of these cues are then examined. Studies using retrocues have highlighted clear benefits following these internally directed attentional cues. When reporting a cued item from working memory, participants are both faster and more accurate in their responses based on the validly cued memorandum (e.g. Astle et al., 2012; Griffin & Nobre, 2003; Hajonides et al., 2020; Landman et al., 2003; Niklaus et al., 2017).

Early work on the orienting of attention to working-memory representations focused on the use of spatial retrocues (e.g. Griffin & Nobre, 2003; Landman et al., 2003). This is natural, given the seminal attentional work by Michael Posner based on spatial orienting. Posner showed that attention can be covertly oriented towards spatial locations in a perceptual task (Posner, 1980). Participants were either presented with an uninformative plus sign or an informative arrow (80% valid cue) that indicated the likely location of an upcoming stimulus they had to detect. Early work demonstrating that attention can be oriented internally, rather than externally, took a similar approach.

Landman and colleagues (Landman et al., 2003) used a line cue during maintenance, presented just adjacent to fixation, that directed attention towards a spatial location where a change would occur in the test array. Griffin and Nobre (Griffin & Nobre, 2003) used a square stimulus, in which two adjacent sides of the square would brighten to form an arrow and direct spatial attention to a location in memory.

Space is not the only defining feature of the world around us, though. Things have colour, shape, and motion direction and speed. They appear at different times, change over time, and are important at different times. Things have value. The world is complex, and so is our attention. There are many dimensions to what can guide attention.

Feature-based information has been shown to guide attention to the external world around us, with both behavioural and neural consequences. Attention that selects features of objects rather than their spatial location – e.g. their orientation, colour, or motion – modulates single-unit activity recorded in macaques. Treue and Martinez Trujillo showed attentional modulation of activity in direction-selective neurons in the middle temporal visual area (MT) in macaques (Treue & Martínez Trujillo, 1999). They showed that attending to a neuron's preferred motion direction resulted in an increase in firing rate even when the attended stimulus was *outside* of the neuron's receptive field (i.e. when the stimulus inside the recorded receptive field was behaviourally irrelevant). This suggests a spreading of feature-based attention when attending to a specific motion direction. Similar effects have also been observed for feature-based attentional modulations of single-unit activity in V4 based on stimulus colour (e.g. Bichot et al., 2005; McAdams & Maunsell, 2000).

Feature-based attentional effects also impact behaviour in perceptual tasks. Ling, Lieu and Carrasco modelled visual detection performance in a two-alternative forced-choice task (2AFC) in which participants were required to detect the motion direction of a random-dot motion stimulus relative to a reference direction (Ling et al., 2009). In their task, observers could be cued in advance of the reference motion direction. They found that the threshold of motion detection was lower when validly pre-cued towards the motion direction, compared to a neutral cue. Ho and colleagues found similar benefits in perception with feature-based attention (Ho et al., 2012). In their task, participants were presented with four random-dot motion stimuli, one in each quadrant of the screen, and were required to report which of these four stimuli contained coherent motion (only one of the four contained motion coherence). In half of the trials, participants were validly pre-cued to the direction of motion that would be present on that trial (but not where this would occur). Modelling showed that the rate of sensory evidence accumulation (the *drift rate* of the linear ballistic accumulator model) was significantly higher when valid feature-based attentional cues were presented, compared to neutral or invalid cues. In a second experiment, competition between stimuli was increased (target motion patches contained 80% coherent motion, and distracting patches contained 40% coherent motion). Participants were more accurate in reporting the location of the more coherent patch when validly cued to the motion direction. Not only were they better, but they were also faster at reporting the target. In this second experiment, there was also an associated cost in both accuracy and reaction time when participants were invalidly cued to the feature.

Altogether, the single-unit and human behavioural data provide compelling evidence for feature-based attentional cues affecting human perception.

These findings have also been shown to extend to the context of working memory (e.g. Hajonides et al., 2020; Niklaus et al., 2017; Park et al., 2017; Ye et al., 2016). Indeed, Niklaus, Nobre and van Ede showed that feature-based attention can operate upon representations held within working memory (Niklaus et al., 2017). In their first experiment, participants were required to encode three coloured, oriented arrows into working memory. During maintenance, a word appeared, cueing them to the feature dimension they were required to report from one of the items or conveying no information about the feature they would report. They found that performance was better when validly cued (e.g., when cued to report colour, and subsequently prompted to report the colour of a stimulus) compared to when the cue was invalid (e.g., when cued to report colour but subsequently prompted to report the orientation of a stimulus) or when neutral cues provided no information. The findings suggested that a feature dimension (one feature across all objects in working memory) could be prioritised. They extend the observation in a third experiment, showing that feature-based attention spreads even to non-cued items. In this third experiment, participants could be presented with two retrocues during maintenance. The first cue oriented attention to one feature dimension of a specific item. In one-third of the trials, they were then probed about this cued item. In the remaining two-thirds of the trials, a second retrocue appeared, orienting attention to a feature dimension of a different item in working memory. This second cued feature dimension could either be the same or

different as previously cued. They showed that when re-orienting attention to a new feature dimension (e.g. re-orienting from colour to orientation), participants were worse than if simply reporting the same feature of a different object. This suggested that although initially tied to a specific item, this feature-based attentional effect spread across the mnemonic array. Crucially, they found that when colour was cued to be behaviourally relevant, there was no cost associated with refocusing on a new item when the second retrocue appeared. This suggests that colour-based attention spreads across working memory, including to non-attended items. This is in line with similar findings in perception showing a global spread of feature-based attention (McAdams & Maunsell, 2000; Saenz et al., 2002; Treue & Martínez Trujillo, 1999). Crucially, these benefits to the cued feature are typically associated with costs to the uncued (deprioritised) feature dimensions.

Altogether, these findings highlight a more interesting story about working memory: rather than simply storing information as a representation to be recalled, working memory is highly flexible. The same information that can guide our attention to the external world can be used to orient toward contents within our memory. We can use these sources of information to prioritise and deprioritise memoranda. We can shift our focus between items, feature dimensions, or feature values of items, to best retrieve memories to serve our current task goals.

1.3 Reward

In the studies mentioned so far, one neglected aspect of the world (internal or external) is reward. Things around us have value. They can be valuable themselves (e.g. primary reinforcers such as food or water), or be associated with value through learning (e.g. money or other associative cues). Rewards can be financial - for example, linking monetary rewards to performance on task. In other cases, reward is motivational – feedback can be intrinsically rewarding and enables us to learn about the quality of our performance on task. In the next section, I provide evidence that reward shapes attention to the external world around us. I then explore how reward has been shown to influence working-memory processes and spell out what open questions I aim to answer with some of the empirical work in this thesis.

1.3.1 Reward and attention

When we think about the factors that guide visual attention, we often think of sensory attributes like space, colour, orientation, shape – aspects of the visual environment that are directly linked to the visual domain itself. In other words, these fundamental aspects of our environment are salient because of the fact we can experience them with our senses. However, attention can also be controlled according to higher-level attributes – such as the value associated with stimuli or actions.

Value has been shown to be represented widely across many brain areas (e.g. Baruni et al., 2015; Komura et al., 2001; Peck et al., 2009; Platt & Glimcher, 1999; Stanisor et al., 2013). In 1999, Platt and Glimcher showed that reward modulates the activity of neurons recorded in macaque lateral intraparietal (LIP) in a cued-saccade task (Platt & Glimcher, 1999). In their task, monkeys were required to make specific eye movements in response to a centrally presented cue, with a reward received when a correct saccade was performed. They showed that the activity of neurons with a receptive field in the periphery was higher when the expected reward associated with the cued saccade was higher. When making a value-based decision between two spatial locations, LIP neurons seem to track the relative value of the two options.

This type of reward-based modulation of single-unit activity has been shown in earlier visual areas too. Baruni and colleagues trained monkeys to perform a dual 2AFC orientation discrimination task across two different spatial locations (Baruni et al., 2015). Monkeys were presented with a stream of oriented Gabor patches at two different spatial locations. Each was pre-cued to be associated with different rewards should the monkey perform correctly. The monkeys had to determine whether probes were oriented more horizontally or vertically relative to the final (target) orientation of the stream. This was effectively designed to introduce an attentional bias towards one of the spatial locations based on the magnitude of possible reward for accurate performance. Monkeys were faster and better at this discrimination when the relative value of the probed orientation was higher. Critically, behavioural performance was mostly driven by the relative value of the

target, not its absolute value – suggesting that performance was related to the unbalanced, reward-guided distribution of attention to the two spatial locations. In this task, the researchers were able to manipulate and orthogonalize absolute value and relative value: a spatial location could be worth 0.5 ml of juice on a trial where the corresponding distractor was worth 0.1 ml, or in which the distractor was also worth 0.5 ml. This allowed them to examine the contributions of both relative and absolute value in modulating neural activity.

Activity in V4 neurons appeared to reflect mostly the *absolute value* of the stimulus in their receptive field. Enhancement was related mostly to the absolute value of the receptive-field stimulus, rather than its relative value compared to the distractor stimulus. This presents an intriguing dissociation between brain and behaviour – behavioural performance was most changed by relative value of the two spatial locations, yet modulation of single-unit V4 activity tracked the absolute value of the location. Similar modulations of neural activity have been seen in earlier visual stream regions, notably in V1 (Stanisor et al., 2013) and even at the level of the thalamus in rats (Komura et al., 2001). In sum, this paints a complicated picture for how reward-guided attention affects the brain and behaviour. Reward is represented in many areas, but modulations in these areas may not all be necessarily linked to changes in behaviour. Instead, higher cortical regions associated with action planning and control (e.g., LIP) may integrate information about the reward associated with spatial locations to reflect the relative value of perceptual information, in advance of initiating a response action.

Reward has also been shown to affect both behaviour and neural activity in humans across a variety of tasks (e.g. Bourgeois et al., 2016; Camara et al., 2013; Chelazzi et al., 2013; Della Libera & Chelazzi, 2009; Engelmann et al., 2009; Engelmann & Pessoa, 2007; Gruber et al., 2013; Kiss et al., 2009; Serences, 2008; Small et al., 2005). Various effects on human perception have been reported. For example, reward has been shown to affect the perceptual sensitivity of our visual system in a variant of the classic Posner-cueing task (Engelmann & Pessoa, 2007). In this task, a peripherally presented cue predicted the location of a perceptual target (a faint red dot inside a face or house stimulus) with 70% validity, while motivation was manipulated by varying the magnitude and valence of an expected reward based on accurate performance. Across two experiments, they showed that perceptual sensitivity to a target increased as the incentive value (the possible reward on a trial) increased. This suggests that increases in motivation (due to the prospect of reward) influence attention. Notably, the prospect of gaining reward, or avoiding losing reward, results in motivational changes that cause increased attention at the relevant spatial location. Indeed, given previous neurophysiological results in animal models showing increased spatial selectivity for rewarded receptive fields (increased processing of rewarded areas of space) and increases in firing rate based on absolute or relative reward, this could be supported by increased neuronal gain in neurons processing rewarded areas of the visual environment. In a follow-up study, Engelmann and colleagues used fMRI to explore the mechanisms that might underpin such reward-guided attentional prioritisation in space (Engelmann et al., 2009). Here they found widespread modulation of areas implicated in spatial attentional orienting (parts of the putative attention network, e.g.,

Corbetta & Shulman, 2002; Nobre et al., 1997) alongside parts of the anterior cingulate cortex (ACC) – an area strongly implicated in reward-guided learning and performance monitoring (Baker & Holroyd, 2011; Hauser et al., 2014; Rushworth & Behrens, 2008; Yeung et al., 2004). Others have suggested that there might be additional recruitment of areas implicated in affective and motivational processing when reward is involved in the deployment of attention (Mesulam, 1990; Small et al., 2005).

Behaviourally, reward has other consequences and has been shown to shape attention even when the reward is not linked to performance. In a *negative-priming task,* Della Libera and Chelazzi showed that the priming effect only occurred when responses were rewarded (Della Libera & Chelazzi, 2006). In a negative-priming task, two stimulus arrays are presented sequentially, both of which require a perceptual judgement – the first is termed the *prime* and the second is termed the *probe*. Negative priming refers to the behavioural impairments (longer reaction times and reduced accuracy) seen when the target of the second stimulus array matches the distractor in the initial prime array. In their study, Della Libera and Chelazzi rewarded participants based upon their response to the prime display. While they informed participants that high and low reward related to the quality of their performance, in truth the reward was unrelated to behaviour. They found that the negative-priming effect was influenced by reward feedback – following low reward (which participants were informed was related to poor performance), negative priming was not seen. Performance has also been demonstrated to be affected by reward processing in another core attentional paradigm – the *attentional blink.* When viewing a sequence of

images, attending to a first target stimulus (target one, T1) influences the ability to detect a subsequent target stimulus (target two, T2) presented soon after (Husain et al., 1997; Raymond et al., 1992). Within this paradigm, reward has profound effects. When the T2 stimulus was associated with a high probability of reward, its recognition was not impaired as would be expected within an attentional blink (Raymond & O'Brien, 2009). This suggests that reward can affect early stages of attentional processes.

Another important way in which reward can influence visual selective attention is through *value-driven attentional capture (VDAC)*. This describes the powerful phenomenon by which stimuli that were previously rewarded continue to guide attentional processes – even when the previously learned reward contingencies are no longer in place (Anderson, 2013; Anderson et al., 2011; Camara et al., 2013; Tankelevitch et al., 2020; Wallis et al., 2015). In VDAC paradigms, participants first learn to associate basic features of an object (for example their shape or colour) with a specific reward magnitude during a learning phase. In a secondary task, the consequences of the learned association are tested. VDAC has revealed performance benefits on visual search (Anderson et al., 2011) when task-irrelevant objects in the visual environment were previously associated with reward. It also affects free-choice responses – when participants can freely choose which of two targets to saccade to (neither of which is rewarded), participants more frequently choose targets at the location that was previously rewarded in an irrelevant task (Camara et al., 2013).

Reward also has demonstrable consequences on neural indices of selective attention to the visual environment. In both EEG and magnetoencephalography (MEG),

reward has been shown to modulate the N2pc (Kiss et al., 2009; Tankelevitch et al., 2020) – a commonly-investigated index of spatial selection to both externally-presented stimuli and internally-held representations (e.g. Dell'Acqua et al., 2010; Doallo et al., 2013; Eimer, 1996; Kiss et al., 2009; Kuo et al., 2009). Multivariate decoding techniques that leverage whole-brain signals recorded in MEG have also shown value-based modulations of spatial attention when locations are cued by stimuli that were previously associated with reward but are now no longer predictive of reward (Tankelevitch et al., 2020). In fMRI, reward value also modulates activity across many areas of visual cortex, and sub-regions of prefrontal and parietal cortices (Serences, 2008).

It is clear that reward is powerful in guiding selective attention to the visual world around us. Alongside spatial location and the features of objects in our environment, reward is an important dimension to consider when interacting with the external world around us.

1.3.2 Reward and working memory

While reward has clearly been shown to guide attention to the world around us, our understanding of how reward is associated with the control of working memory is much less clear. Most of what we know has focused on how reward influences the flow of information *into* working memory, often using principles from value-driven attentional capture paradigms that have typically focused on reward effects on working-memory

encoding. In such tasks, reward has been shown to have interesting effects on working-memory performance.

In VDAC-like working-memory tasks, participants learn to associate differing levels of reward to different stimuli that will be used in an upcoming working-memory task (e.g. Infanti et al., 2015; Klink et al., 2017; Wallis et al., 2015). Wallis and colleagues showed that when such stimuli were used in the encoding array of a separate working-memory task, performance on the working-memory task was higher when the value of the entire encoding array was larger – even though the reward associations were not relevant in the working-memory task (Wallis et al., 2015). Infanti and colleagues used a visual-search task (rather than a binary choice learning task) to have participants learn reward associations and then test their effect on working-memory performance (Infanti et al., 2015). Participants were required to search for a target defined by one of two pre-defined colours and report its orientation. One colour was associated with high reward and the other with low reward. In their working-memory task, eight oriented lines (either vertical or horizontal) were briefly presented evenly spaced around, and equidistant from, fixation. All of the lines were surrounded by small placeholder circles. On each trial, one of the placeholders was coloured to match one of the colours that had been associated with reward in the visual search task (all other placeholders were grey). Participants were subsequently required to report the orientation of one of the encoded items. Accuracy was higher for items associated with any level of reward compared to no reward. Further, they found that performance was diminished for items presented adjacent to the rewarded item,

suggesting that rewarded items can generate interference for objects encoded in their vicinity.

Klink and colleagues have tried to determine exactly where reward can operate upon working-memory representations, by presenting reward cues at different phases of a working-memory task (Klink et al., 2017). In their task, participants learned to associate specific colours with rewards in a learning task. These colours were then used as the colour of placeholder rings in a subsequent working-memory task where participants had to report the orientation of one item from a set of three Gabor stimuli encoded on each trial, where the reward contingencies were still relevant. When placeholders changed colour during encoding - indicating reward information at the point of input into working memory - performance was higher (response error was lower) when reporting orientations associated with higher reward values. When reward cues were presented during the maintenance period or at the onset of the response probe, no significant effect on response error occurred (though this could be related to low statistical power due to small participant samples). In a second experiment, they showed that this reward-guided bias of encoding into working memory persisted even when the previously learned reward contingency was task-irrelevant (see also Infanti et al., 2015).

More recently, Lin and Fougnie used reward-value to elucidate the cognitive mechanisms that support the retrocue effect (Lin & Fougnie, 2022). Participants were presented with three coloured circles, before reporting the colour of one of them. During the working-memory delay, they were either cued to remember two of the items that they

could be probed about (allowing them to drop one item from working memory) or cued that all three could be probed. Participants were rewarded proportionately to the error in their response. At the beginning of the experiment, participants learned to associate specific reward values to the three possible spatial locations that were used in the task, resulting in a reward-guided encoding bias when objects entered into working memory. They found that working-memory performance was higher for items presented in more valuable spatial locations (although participants were also slower to respond), suggesting that increasing reward was associated with more accurate formation of working memories (as in Klink et al., 2017). They also found that the retrocue benefit did not differ across levels of reward and that dropping rewarding items did not confer increased benefit to uncued items. Their findings are directly inconsistent with resource-allocation theories of retrocue benefit (e.g. Pertzov et al., 2013; Souza, Rerko, Lin, et al., 2014), which posit that retrocues allow the flexible shifting of mnemonic resources to cued items at the cost of uncued items. Under this framework, high-reward items (associated with higher memory quality) would be allocated more resources during encoding. Thus, cues that allow dropping the high-reward item should free up larger amounts of cognitive resources, which could then be reallocated to other items, conferring extra benefit compared to when low-reward items are dropped. This was not the case.

Altogether, these studies highlight some interesting consequences of reward on working memory. Firstly, the reward information presented at the point of encoding into working memory leads to higher quality responses (a proxy for higher quality working-

memory representations) for objects associated with high vs low reward (Infanti et al., 2015; Klink et al., 2017; Lin & Fougnie, 2022). Such reward-based improvements in performance are also seen in other memory contexts (e.g. Doallo et al., 2013; Gruber et al., 2013; Wittmann et al., 2011). At least in the working-memory context specifically, these effects are found even when reward associations are no longer applicable (Infanti et al., 2015; Klink et al., 2017). This suggests that reward can bias (and enhance) the transformation from external sensory information, to internal mnemonic information.

Secondly, when reward cues were presented during maintenance (where reward is assigned retrospectively after encoding), significant benefits were not found for working-memory performance (Klink et al., 2017). Despite the wealth of evidence showing high levels of flexibility in how attention can be selectively allocated to working-memory representations, these researchers concluded that reward-guided attention may be fundamentally limited to gating the input to working memory, rather than influencing maintenance or retrieval processes – although there may be reason to revisit this.

However, one can intuit that reward can be retrospectively assigned and that this might influence behaviour. For example, one day I may leave the lab and walk a new route home. On this particular day, I may be surprised by a particularly nice view of the sunset – something rewarding. As a result, I may choose to take that path more in the future as a result of this unexpected reward. In this anecdotal - and slightly exaggerated - example, the reward (a nice view of the sunset) is assigned retrospectively (I am unaware of the lovely view during the walk) and shapes my future decisions of how I make my way home in the

evening. Indeed, retrospectively assigned rewards have been shown to influence recognition memory in the long-term memory context (e.g. Braun et al., 2018; Patil et al., 2017) after long intervals before test. For long-term memory at least, this suggests the need for consolidation processes to integrate reward information in the formation of long-term memories.

Despite the evidence of how reward can influence working-memory encoding, there are unexplored avenues regarding the intersection of reward and other working-memory processes. Working memory is typically discretised into three components: encoding, maintenance and retrieval. Whether and how reward can influence the maintenance and retrieval of information from working memory through attentional cues is unclear.

The lack of maintenance-based effects of reward so far may be due to how reward related to the task, or was manipulated in relation to attentional processes. Thus, I take a different approach. In Chapter 4 of this thesis, I explore how reward can guide the prioritisation of information in working memory. Specifically, I adapt the retrocue approach to do this, attempting to make the distinction between *capturing* attention, and *guiding* it. In other words, instead of using reward cues to try to *capture* attention towards a spatial location tied to a learned reward history at encoding, I use retrocues that instead *guide* attention to specific internal representations by assigning value to them on that trial. Instead of using reward to bias working-memory encoding, I explore how reward can guide attentional prioritisation of objects already encoded (and assumed to be encoded

similarly). This allows me to investigate reward-guided *prioritisation* of working-memory representations during maintenance and retrieval.

1.4 Learning and uncertainty

In the realm of both perceptual and economic decisions, uncertainty is a key aspect of the environment that one learns about and can integrate into the decisions or perceptual judgements that one makes. Understanding the uncertainty in information arriving at our senses can help guide the perceptual judgements that we make. Understanding and learning about the uncertainty in reward outcomes shape behaviour when choosing between available options.

In discussing uncertainty, I make the distinction between two types of uncertainty. Firstly, there may be *computational uncertainty*. In this, I refer to uncertainty introduced in the tasks that we have participants perform – for example, varying load in a working-memory task or changing the perceptual parameters of stimuli to make the perceptual decision more difficult. Secondly, there can be *cognitive uncertainty*. With this, I refer to the uncertainty that we are aware of based on our ability to introspect on the environment and our decisions, and uncertainty that we can measure through introspective judgements. Cognitive uncertainty would relate to the confidence in the appropriateness or accuracy of the behaviours that we make.

In the context of memory-guided behaviours, one can envisage how this might be important. Take the classic example used by many: trying to find your keys. On the face of it, this is a visual search task – in working memory, we hold a template of what our keys look like, and apply this template to incoming sensory information while we visually search the house. It is also a working-memory task: we maintain a representation of what our keys look like while searching. One also keeps a running track of *where* in the house we have searched to try to find them – as we search, we increase the load on our memory system. But this particular morning, I am stressed. I thought I searched in the kitchen, but I am uncertain if I definitely did. Did I already check the pocket of the coat I was wearing yesterday? I can't quite remember. I can be unconfident in where I have already searched to find my keys (cognitive uncertainty), or even uncertain about what my keys looked like in terms of colour or shape (an internal sensory uncertainty). Depending on the *certainty* with which I remember searching a particular location, I may go back just to *become sure*. When making memory-guided decisions between competing actions or available options, it is important to understand the quality of, or uncertainty in, our memories in order to make the best possible decision.

1.4.1 Uncertainty in perceptual and working-memory guided decisions

In perceptual decisions, sensory evidence for a decision appears to be represented in relevant areas of sensory cortex. In motion-discrimination tasks, activity in the gamma frequency range (60-100 Hz) recorded from motion-sensitive areas has been shown to

correlate with the strength of visual motion (Siegel et al., 2007). When area MT is stimulated in macaques during dot-motion tasks, monkeys decide for the neurons' preferred motion direction more often (Ditterich et al., 2003), suggesting a causal role of activity in this area for perceptual-motion decisions. When decisions are simulated based on the output firing of MT neurons in macaques, they mirror (or exceed) the monkey's performance (Newsome et al., 1989). Similar effects for sensory evidence encoding have been seen for perceptual decisions in other sensory domains, such as somatosensation (e.g. Hernández et al., 2000; Romo et al., 1998, 2000).

The amount of evidence for different percepts (e.g., a particular motion direction) then appears to be integrated in higher-order areas where sensory evidence can be transformed into action. For example, activation in macaque LIP neurons appears to reflect the integration of sensory evidence toward a saccadic decision (Hanks et al., 2006; Roitman & Shadlen, 2002). Stimulating LIP neurons influences choices – speeding saccadic decisions *towards* the location of the receptive field of the stimulated neuron and slowing choices to other locations (e.g. Hanks et al., 2006; Huk & Shadlen, 2005). Indeed, in humans, activity in regions relating to the specific motor action required for response has been shown to relate to the integrated evidence for a perceptual decision in both fMRI (e.g. Tosoni et al., 2008) and MEG (e.g. Donner et al., 2009; Gould et al., 2012).

These findings are important as they lay the foundation for two types of uncertainty in the context of perception: sensory uncertainty (related to the sensory evidence for a specific motion direction, for example) and decision or computational uncertainty (related

to the difference in evidence for available options). There is evidence that these types of uncertainty influence decisions. Kiani and Shadlen trained monkeys to make a perceptual decision about the net motion of a random-dot motion stimulus (Kiani & Shadlen, 2009). The monkey was required to saccade to a location to indicate its direction choice, and was rewarded for correct performance. On half of trials, they were able to opt-out of the decision (by making a saccade to a third target) and receive a smaller but certain reward. Monkeys opted out of deciding on the motion direction when sensory evidence was weak. However, when they did *not* opt out of the decision on these trials, choice accuracy was higher. This suggests that the monkey was incorporating information about its certainty in the percept to guide its decision across trials. This decision to opt-out was associated with lower firing rates of LIP neurons.

In humans, uncertainty is also an important factor in perceptual decisions. Subjective certainty in accuracy of responses has been shown to relate to behavioural performance in a range of tasks – generally speaking, when people are better they are more confident (e.g. Boldt et al., 2019; Boldt & Yeung, 2015; Denison et al., 2018; Honig et al., 2020; Kiani et al., 2014; Samaha et al., 2019). For example, Kiani, Corthell and Shadlen investigated the link between confidence and performance in a random-dot motion discrimination task (Kiani et al., 2014). Participants were required to detect the motion direction (up or down) and then report this motion through a saccadic eye movement to a horizontal bar target. Critically, participants also reported their confidence at the same time as this response, by saccading to a point along the length of a horizontal bar target

above or below fixation (uncertain to 100% confidence, from the left to right of the stimulus respectively). When correct, participants' confidence (displacement along the x-axis of the bar target) was systematically related to the strength of sensory evidence (motion strength) across trials. Trials with longer reaction times were also associated with lower reported confidence in the response. This suggests that confidence could be related to sensory uncertainty (motion strength), but was also based on an evaluation of how long it took to make the decision (a proxy of decision uncertainty). However, for incorrect trials, confidence was higher when motion strength was higher. This suggests that confidence in responses may not be linked only to an evaluation of decision performance, but also to other aspects of the decision process.

More recently, van Bergen and colleagues (van Bergen et al., 2015) showed that humans appear to use information about sensory uncertainty in their perceptual decisions. In a simple perceptual task (in reality, a simple working-memory task), participants were presented with an oriented Gabor stimulus and, after a delay, were required to reproduce its orientation using button presses. Typically, in studies that investigate the consequences of stimulus uncertainty, perceptual parameters are externally manipulated to induce levels of uncertainty. Here, they took advantage of the natural trial-wise variability in response error to look at endogenous changes in uncertainty. Using fMRI decoding that factored in the orientation preference of recorded voxels, van Bergen and colleagues estimated the full probability distribution of orientations across areas of the visual cortex. They took the circular mean as an estimate of the presented orientation, and the standard deviation as

an estimate of uncertainty in the percept. With this method, they showed that participants made more accurate decisions (orientation reproductions) when the precision of the decoded information (its certainty) from visual cortex was higher. They also found that the cardinality bias (a tendency to report remembered orientations as being vertical or horizontal) was lower when decoder certainty was higher, suggesting a reduced reliance on this bias when certainty about the stimulus was higher.

Rather than inferring the uncertainty of memory representations through neural activity or behavioural responses, others have looked at how cognitive uncertainty relates to working-memory judgements (e.g. Honig et al., 2020; Li et al., 2021; Rademaker et al., 2012; Samaha et al., 2019; Yoo et al., 2021). These studies have generally shown that individuals' confidence judgements (their reported estimates of uncertainty in their response) track the quality of working-memory recall in both categorical (Rademaker et al., 2012) and continuous (Geurts et al., 2022; Honig et al., 2020) confidence judgements. Working-memory load manipulations have been used to vary working-memory uncertainty, resulting in changes in subjective confidence (Rademaker et al., 2012). Under conditions of higher working-memory load, individuals were less confident in their responses on average. Further, for both set sizes (load 3 and 6), low confidence was associated with larger response errors and flatter response-error distributions that indicated more random responses. Mixture-model decomposition of response-error distributions showed higher guess rates with lower confidence levels and increased visual working-memory precision with higher confidence. Confidence also appears linked to

sequential biases across trials – with evidence that items associated with higher confidence exert stronger biases on working-memory recall on subsequent trials compared to lower confidence items (Samaha et al., 2019). Confidence in working memories, therefore, may be derived from evaluation of uncertainty in different working-memory processes. It is not directly read out by evaluating working-memory representations, but likely combines information from different sources to allow us to infer our uncertainty. This includes, but is not limited to, uncertainty during maintenance (working-memory load), the quality of mnemonic representations, or how well information was encoded into memory. Understanding how we introspect upon working memory, and estimate the uncertainty therein, will undoubtedly help us better understand how working memories are used for proactive behaviours.

Earlier in this introduction, I focused on how selective attention has been shown to operate upon the contents of working memory, and the behavioural consequences of attentional prioritisation. Selective attention to specific memoranda, facilitated by retrocues, has been shown to reduce average error, reduce variability in response errors, and speed up responses. In essence, retrocues allow us to selectively reduce our working-memory uncertainty by prioritising a specific item as more relevant for behaviour - often allowing us to drop concurrently maintained information. What is unclear is whether the allocation of attention to specific memoranda necessarily results in changes in reported confidence in working-memory recall. When attention allows us to prioritise and enhance a specific memorandum, do we see commensurate changes in introspection? In Chapter 2,

I investigate whether the prioritisation of working-memory contents through attentional retrocues changes our confidence in the working-memory-guided judgements we make. In Chapter 4 I explore whether the reward value of working memories also results in changes in our estimates of working-memory uncertainty.

1.4.2 Learning through feedback

How we learn in the environment has been extensively studied in the context of rewarded decisions (for review, see e.g. Daw & Doya, 2006; Rushworth & Behrens, 2008; Sugrue et al., 2005). In the burgeoning field of neuroeconomics, participants are required to choose from a set of available courses of action in order to maximise reward. A whole class of models – *reinforcement learning* models – arose in order to discover how agents optimally use feedback regarding decisions to optimally learn about, and better understand, the reward contingencies in our environment (Behrens et al., 2007, 2008; Hunt & Hayden, 2017; Kolling et al., 2012; Niv & Schoenbaum, 2008). In such tasks, agents may be presented with two (or more) different options (stimuli) to choose from, each of which is associated with a particular probability of reward (the uncertainty of being rewarded) and a reward magnitude (the value received if the option is rewarded on a trial) (e.g. Nieuwenhuis et al., 2004; Padoa-Schioppa & Assad, 2006; Scholl et al., 2015; Yeung & Sanfey, 2004). Across trials, agents learn the value of the available options and choose the more rewarding option when present to maximise the reward they receive. Theories of economic choice propose that the *expected value* of the available options is compared to compute such decisions. By

simply multiplying the probability of reward by the magnitude of reward, an agent can estimate the reward expected for a particular decision. When given a 50% chance of gaining £10, or an 80% chance of gaining £9, a rational agent would choose the latter. Crucially, the agent must know whether the decision or behaviour was correct – through feedback – in order to determine whether behavioural change is necessary, or how much to alter behaviour.

In humans, much research into the processing of feedback has focused on EEG recordings, commonly finding a component we know as the feedback-related negativity – or FRN (Becker et al., 2014; Hajcak et al., 2006; Hauser et al., 2014; Miltner et al., 1997; Nieuwenhuis et al., 2005). It was first described in a time-estimation task where subjects had to produce a 1-second interval before receiving feedback on their performance (Miltner et al., 1997). Critically, feedback expectation was controlled for by varying the criteria for positive feedback – as subjects improved, the criterion for correctness in the response became stricter. This allowed researchers to control the probability of correct and incorrect feedback such that one type of feedback was not more probable across trials. When comparing the potentials elicited by incorrect compared to correct feedback, they observed an early relative negativity in the feedback-locked event-related potential. This feedback-related negativity for incorrect trials was observed across the midline of the scalp, from frontal (FPz) to posterior (Pz) electrodes. Crucially, similar components were observed whether the feedback was visual, auditory or somatosensory in nature, suggesting a domain-generality to this feedback component.

In keeping with the possibility of the FRN reflecting a general-purpose signal reflecting the valence of outcome information at feedback, the FRN has been seen in a multitude of tasks and is sensitive to many aspects of task structure. Gehring and Willoughby describe a medio-frontal negativity akin to the FRN, present when receiving outcome regarding gains and losses in a classic monetary gambling task where participants had to choose between two squares (Gehring & Willoughby, 2002). The component they recorded was sensitive to the valence of the outcome (a monetary gain or loss) rather than the correctness of the chosen action (whether they won the most possible or lost the least possible). When others replicated and extended these results, they found that the feedback-evoked response could be sensitive to both outcome valence (the utilitarian information present at feedback) *and* the correctness of responding – depending on what information was made salient at the time of feedback (Nieuwenhuis et al., 2004). FRN amplitude also appears to be sensitive to agency in tasks. In a monetary gambling task where participants could either make active choices between two options (choice) or were instructed to press a single button and have the choice made for them by a computer (no-choice), an FRN was seen in both conditions but was larger when decisions were actively made (Yeung et al., 2005). The FRN is also sensitive to reward probability, magnitude, and the expectedness of the information that feedback contains (Cohen et al., 2007; Holroyd et al., 2008; Pfabigan et al., 2011; Yeung & Sanfey, 2004).

Overall, the FRN appears to reflect at least the binary processing of feedback in a process of performance evaluation. Locked to the onset of feedback conveying information

about the success of performance, it may reflect many aspects of task structure, depending on what the feedback stimulus itself emphasises. In fact, studies examining the underlying neural source of the FRN have consistently identified the ACC as the likely generator of the signal – from dipole modelling in EEG (Gehring & Willoughby, 2002; Miltner et al., 1997) to EEG-informed MRI analyses of simultaneous EEG-fMRI data (Becker et al., 2014; Hauser et al., 2014). Activity in the ACC has been shown to reflect reward expectation and prediction-error signals (Amiez et al., 2005, 2006; Matsumoto et al., 2007). The area receives input from dopaminergic neurons in the striatum (Williams & Goldman-Rakic, 1998) that represent prediction errors on reward outcomes (Schultz et al., 1997; Waelti et al., 2001). As a result, ACC activity is well placed to play a key role in performance evaluation (e.g. Ito et al., 2003), with the FRN reflecting the aspect of ACC activity during this process that we can measure at the scalp.

In sum, there appears to be a neural circuitry that uses feedback to learn about the consequences of actions. Based on prediction-error signalling, we can then update our model of the world. Evidence suggests that there are domain-general signals that reflect active processing of feedback to evaluate performance – across different types of perceptual and economic tasks, and across different feedback modalities. The question then remains: how does this relate to working memory? Are we able to evaluate performance of working-memory-guided behaviours similarly to how we evaluate our ability to choose between two valuable options? How do we process and update our internal estimate of uncertainty when such uncertainty is not about external reward

contingencies, but the correspondence between internal WM content and the world? In Chapter 3, I investigate feedback-processing in the context of working-memory performance. Across two experiments, I aim to extend previous findings of a FRN when comparing feedback relating to good and poor performance in perceptual and economic tasks, to evaluations of the quality of working-memory recall. Further, using model-based analysis of the feedback-evoked response, I explore whether there are representations of error and uncertainty in the feedback response that might represent signals that can be used to learn about the quality of our memories and our memory-guided behaviours, describing a key contributing neural signal for constructing internal uncertainty estimates.

Thesis Scope

The overall goal of the thesis is to advance the understanding of how we maintain internal WM content, make memory-guided decisions and reflect about our own performance. I particularly focus on the multiple factors that drive uncertainty estimates and judgements such as reward, uncertainty feedback, load and objective error, and their neural substrates. Across three empirical chapters, I explore working memory as a proactive tool that guides behaviour.

In Chapter 2, I report two experiments that explore how individuals introspect upon the contents of working memory – how people make judgements about the uncertainty of their memory contents. I examine the relationship between subjective uncertainty and objective performance, and its influence by selective attention. Further, I look at how

feedback regarding this introspection can shape learning about the quality of memories – something I believe is critical for flexible memory-guided action.

In Chapter 3, I report two experiments using EEG recordings to examine the neural mechanisms of this feedback processing. In the first experiment, uncertainty in working memory was manipulated externally by varying working-memory load. In the second, I measured endogenous fluctuations in uncertainty by carefully measuring subjective uncertainty through a continuous confidence judgement. I show that the feedback-evoked response is sensitive to both types of uncertainty across tasks. Furthermore, I specifically isolate confidence prediction error signals separately from simple error signals. Such confidence prediction errors could guide adaptive confidence changes over time.

In Chapter 4, I investigate the role of reward in guiding attention to items held in working memory. Rather than use learned reward associations to bias encoding, I take a retrocue approach to measure how retrospective reward assignments can bias the maintenance of memoranda.

In Chapter 5, I discuss the implications of these findings and attempt to integrate them across the attention, working memory, and perceptual decision-making domains to understand how attention, feedback, uncertainty and reward are integrated in working-memory guided judgements.

2 - Confidence in working memories is shaped by attention and recent performance

2.1 Abstract

Working memory (WM) is capacity limited, and our ability to recall information from WM is variable. Directing selective attention to internal mnemonic representations reduces behavioural variability, and helps overcome capacity limits. It has been shown that people are able to make introspective judgements regarding the quality of their memories, and these judgements are linked to objective memory performance. However, it remains unknown whether benefits of internally-directed attention on memory performance occur alongside commensurate changes in memory uncertainty. Across two experiments, I used retrospective cues (retro-cues) during working-memory maintenance to direct attention to items in memory, then examined their consequence on confidence judgements made. I found that internal attention improved confidence judgements as well as performance of the probed item. Confidence judgements and memory quality were correlated. Further, I showed that participants can use feedback on the accuracy of confidence judgements to update their beliefs across time, according to performance. These results extend our appreciation of working-memory flexibility by showing we can selectively modulate our confidence about its contents.

2.2 Introduction

Working memory (WM) describes a core cognitive ability to store and manipulate information on short time scales to guide behaviour. Its use is ubiquitous throughout daily life; from rummaging for a top to match the trousers you picked out to wear, to holding online the objects to collect before leaving the house in the morning. Much research has been dedicated to understanding the contents of WM representations. In contrast, our subjective insight into, or confidence in, working memory has not been a main focus of research even though it is likely to play an important role in shaping the actions we take. For example, if you are uncertain about your memory of the objects you already placed in your bag, you may check it to make sure you have all the things you need. If you are more confident in your memories, you can leave the house with gusto.

While working memory is capacity-limited (Bays & Husain, 2008; Luck & Vogel, 1997; Zhang & Luck, 2008), research has shown that selective attention can be directed internally upon its contents to improve task performances (Griffin & Nobre, 2003; Landman et al., 2003; Souza et al., 2016). Retrospective attention-directing cues ('retrocues') to locations (Landman et al., 2003; Myers et al., 2018), objects, and features (Hajonides et al., 2020; Niklaus et al., 2017) of items maintained in WM can improve performance on tasks. To measure the quality of WM contents and associated benefits of attention-directing cues, WM studies typically report variable precision of responses across trials, and this variability serves as a proxy readout of fluctuations in the precision of encoding or maintenance of the probed items. Memory uncertainty, or imprecision, is inferred based on the width of

response error distributions, without probing actual subjective uncertainty in memory across trials. Moving beyond simply accessing and reporting the objective contents of working-memory, relatively little is known about how people subjectively judge the contents of WM, how these judgements are shaped by attention, or how these judgements change the way in which WM is used.

Recently, some studies have attempted to probe the relationship between memory error and memory uncertainty more directly, requiring participants to make confidence judgements regarding the contents of their memory. Typically, these have involved making categorical confidence judgements on discrete scales when reporting orientations from memory (e.g. Rademaker et al., 2012; Samaha et al., 2019). For example, Rademaker and colleagues (2012) showed that both categorical prospective confidence judgements (made prior to reporting orientation) and retrospective confidence judgements (made after reporting orientation) varied systematically with reporting errors. Low confidence was associated with increased errors and higher guessing rates, while performance was better when participants were more confident about their response. A load manipulation (three vs six items to encode) also revealed lower prospective and retrospective confidence judgements with more items to hold within working memory (Rademaker et al., 2012). However, even when only one item is held in working memory, the same association between higher confidence and higher accuracy (lower error) has been found (Samaha et al., 2019). Interestingly, confidence judgements also seem to influence interactions between items on different trials. Investigating serial dependence biases – how things that

were previously in working memory bias newly encoded items across trials – Samaha and colleagues (2019) showed that items reported with higher confidence exert stronger biases on subsequent trials than items reported with lower confidence.

Moving beyond categorical judgements, others have investigated how confidence and error are parametrically related (Honig et al., 2020; Li et al., 2021). For example, Honig and colleagues (2020) have investigated how confidence and error are related to each other within the same circular space, showing that error and confidence are parametrically linked (Honig et al., 2020). With participants making confidence judgements and colour reports on the same rotating circular dial, they showed that confidence and error were parametrically associated: trials with larger errors were associated with lower confidence. Using permutation-based analyses, they also showed that participants were not merely using heuristics about colour difficulty to guide their confidence judgements, as the observed relationship between response error and confidence was stronger than simulated relationships based on feature similarity. As a result, confidence judgements concerning memorised items appear not to be based solely on heuristics, or stimulus features, but appear to be based, at least in part, on mnemonic representations themselves.

So far, evidence suggests that confidence judgements track the quality of working memory. But, how might orienting attention to working memory contents affect confidence judgements? Selective attention modulates objective performance based on WM – by improving access to, and sometimes the quality of items that are prioritised (see Nobre & Stokes, 2020; Souza & Oberauer, 2016). However, depending on what stage(s) of

working-memory processes confidence judgements are based on, dissociations could result between confidence reports and attention-related benefits. For example, if confidence judgements are based on encoding quality, attentional manipulations during maintenance might not lead to changes in confidence. If confidence judgements are in part formed based on accessibility of items in working memory, then attentional prioritisation should also lead to changes in subjective insight.

In order to address the question of whether deployment of selective attention during working-memory maintenance is associated with changes in subjective confidence, I performed two experiments. In the first experiment, participants reproduced the orientation of a bar previously encoded into working memory – on half of all trials, a retrocue was presented during the maintenance delay indicating which item was to be reported on that trial. Following orientation reproduction in each trial, I introduced a novel confidence judgement to allow participants to report their subjective uncertainty in their response. In a second experiment, I also included feedback on every trial, providing information about participants' response error, subjective confidence, and the error in their confidence judgement. This allowed me to explore whether and how feedback can act as a learning signal to shape confidence judgements across trials.

I used a precision-recall working-memory task in which participants were probed to report the orientation of one of two items from an encoded array. After their report, participants made a confidence judgement on the same circular space – reporting a range of angles centred upon their initial response to indicate their uncertainty. In line with

previous reports, I observed that attentional prioritisation reduced working-memory error. Supporting a link between memory states and memory confidence, I observed that retrocues also improved confidence judgements, reflecting awareness of the quality of memory. Further, I demonstrated that participants can use feedback about the error in their confidence judgements (how over- or under-confident they were) to flexibly adjust their confidence across trials.

2.3 Methods – Experiment 1

Experimental procedures for both experiments were reviewed and approved by the Central University Ethics Committee of the University of Oxford. All participants provided written, informed consent prior to their participation in the study.

Participants

Twenty-two participants were recruited for experiment 1 (7 males, 15 females), two were excluded from data analysis (one failed to confirm their response in almost 40% of trials, a second responded randomly across trials). Twenty participants were fully analysed (6 male, 14 female, age range 20-35 years, mean age 26.9 years, 1 left-handed). All participants reported having normal or corrected-to-normal vision and not being colour blind. Participants were compensated £10 per hour for their time. Participants completed 8 blocks of 32 trials in one session, resulting in a total of 256 trials.

Stimuli, procedure and task

Participants were seated in a dimly lit room at a distance of approximately 40 cm from the monitor (24-inch; resolution: 1920 x 1080 pixels; screen width: 53 cm; refresh rate: 100 Hz). The experimental script was generated using PsychoPy version 1.90.3 (Peirce et al., 2019). All code used to generate experimental stimuli and the task can be found at https://github.com/schekroud/wmConfidence_Behaviour.

Figure 2.1 shows a schematic of the task. At the start of each trial, two coloured bars appeared either side of a central fixation cross (RGB: [150, 150, 150]), for 250 ms. The centre of each bar was six degrees of visual angle (dva) from the centre of the screen, with a height and width of 5.7 and 0.8 dva respectively. Bars could either be blue (RGB: [0, 0, 255]) or yellow (RGB: [255, 255, 21]). Orientations were drawn independently from uniform distributions between 0 and 179 degrees.

After a two-second delay, the fixation cross changed colour for 250 ms. On 50% of all trials, the fixation cross changed colour to match one of the items previously encoded into memory, signalling that only this item was relevant for subsequent report (*retrocue* condition). On the other half of trials, the fixation cross changed to white, providing no information regarding item prioritisation (*neutral* condition). After the cue, there was a variable delay between 2 and 4 seconds (drawn randomly from a uniform distribution of 100-ms steps) before the probe stimulus appeared on screen. On neutral trials, the fixation cross changed colour to match the item that participants had to report. On retrocue trials, the fixation cross changed to white and participants reported the previously cued item (thus ensuring participants processed the cue). A circle appeared around the fixation cross (diameter: 5.7 dva) to indicate the start of the probe phase. Participants had unlimited time to decide on the orientation to report before pressing the 'space' bar (with their left hand) to indicate they were ready. Upon pressing the space bar, two white lines appeared on opposite sides of the probe circle. The starting orientation for the dial was randomly drawn

from a uniform distribution between 0 and 179 degrees. Participants used the mouse to rotate the dial around the circle to replicate the orientation of the probed item. A left mouse click was required to confirm their response. Although participants had an unlimited time to decide on the orientation to report, they were limited to two seconds to deliver and confirm their response. This reporting method therefore provides informative measures relating to accessing and deciding on the item to report as well as to the contents of items reported (see also Niklaus et al., 2017; van Ede et al., 2018, 2019). After clicking to confirm the response, there was a variable delay between 2 and 4 s before the confidence phase began. Participants were presented with their orientation response as a bar inside a circle, with the colour matching the probed item. Participants had unlimited time to decide how confident they were in the response they gave, before pressing the 'space' bar with their left hand to initiate their response. Upon pressing the space bar, two white lines appeared, mirrored symmetrically around the response orientation. Using the mouse, participants adjusted the angle formed by these lines to represent their confidence interval in a circular space – larger wedges corresponded to low confidence in their response, while narrower wedges indicated high levels of confidence. Participants had up to two seconds to dial up and confirm their confidence report before the trial would end. After responding, there was an inter-trial interval of between 1 and 2 s (randomly drawn from a uniform distribution of 100-ms steps) before the next trial began, during which the grey fixation cross was presented.

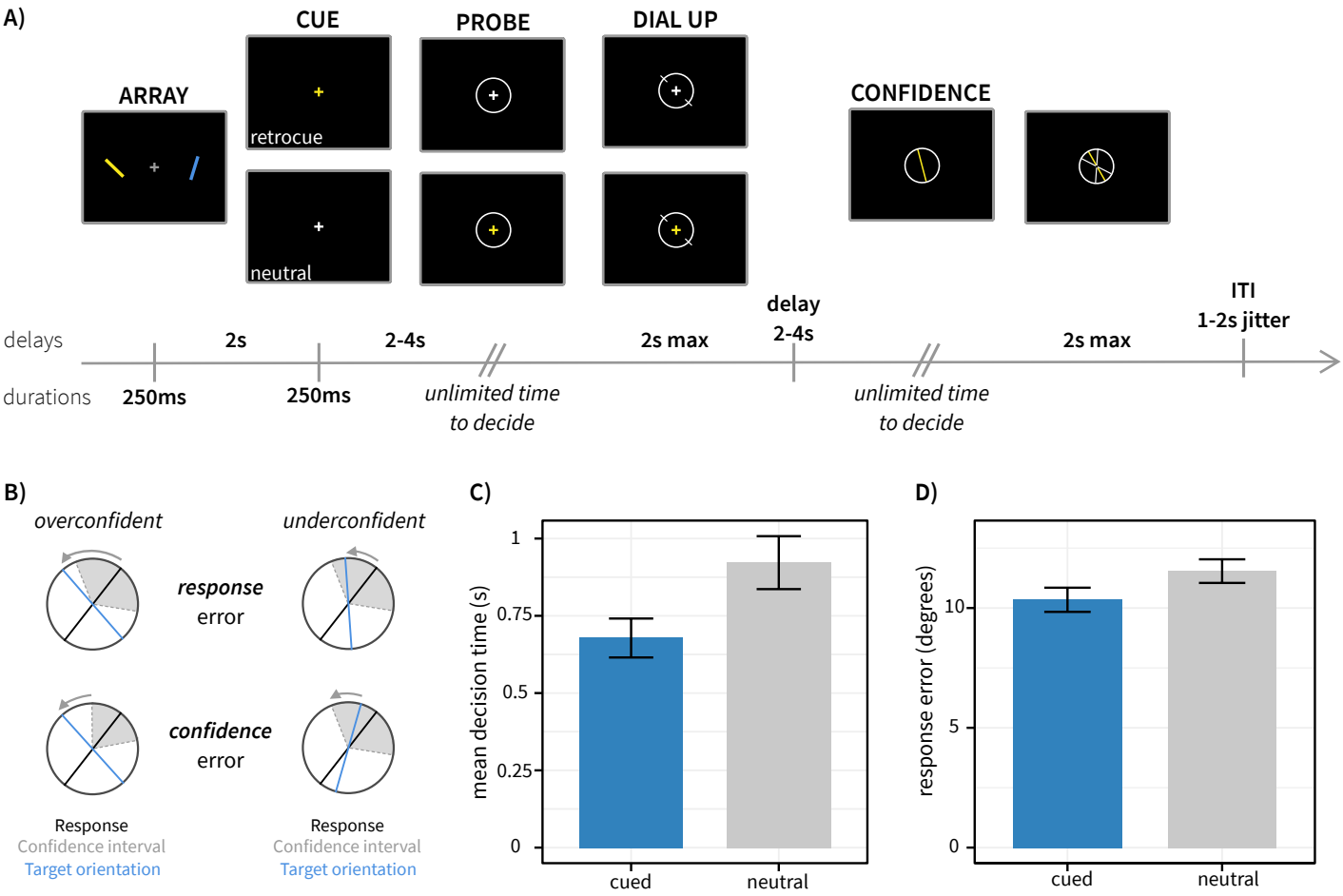

Figure 2.1 - A) Trial structure for experiment 1. Participants were presented with two coloured, oriented bars to memorise. After a delay, a cue was presented that was either informative to which item they would be probed about (retro-cue) or uninformative (neutral cue). Participants subsequently reproduced an orientation from memory. After reporting an orientation, participants reported their confidence in their response. This was done by creating a wedge that reflected a range of values they believed the target orientation could be within – larger wedges reflect lower confidence in their response. B) Schematic visualising two main types of error of interest in this task: response error (deviation between response and target), and confidence error (deviation between response and edge of the confidence wedge). C) Behavioural performance in Experiment 1. Decision time (time taken to initiate the orientation reproduction) was modulated by cueing. D) Mean absolute response error was modulated by cueing. All error bars represent the standard error of the mean.

Behavioural Analysis

I defined two main sources of error. *Response error* was calculated as the angular deviation between the target orientation and the response. *Confidence error* was calculated as the

difference between this response error and the reported confidence wedge on any given trial (i.e. the discrepancy between error and confidence). Both error sources are illustrated in Figure 2.1B. The time between probe onset and pressing 'space' to start the dial phase was taken as the decision time. Where appropriate to analyse aggregate data, repeated-measures Analysis of Variances (ANOVAs) were conducted using the *'afex'* package (Singmann et al., 2020) in the R statistical programming environment (3.4.3, R Core Team, 2020). To analyse the relationship between error and confidence in working memory, and its influence by attention, linear mixed-effects models (LMMs) were implemented using the *'lmerTest'* package (Kuznetsova et al., 2017) to provide p-value estimates for coefficients. I used LMMs as they account for all single-trial variance rather than aggregating data and allow me to look at the strength and direction of relationships while statistically controlling for other correlated effects. Confidence was log-transformed to approximate a normal distribution. To test which factors are linked to confidence judgements using LMMs, I modelled confidence as a function of response error, condition (neutral vs cued), and the interaction term for these two. I initially chose a maximal random-effects structure that included random effects for intercept, effect of error, condition, and their interaction, estimated with their six covariances and sampled per participant. However, maximal random-effects structures can lead to over-parameterisation of models and high degrees of freedom during model-fitting. In order to have models converge on more stable solutions, I used a Principal Component Analysis (PCA) of the random-effects variance-covariance estimate for the fitted mixed-effects models to identify over-parameterisation

(as in Lauer et al., 2018). Random effects that were not supported by the model, and did not contribute significantly to the goodness of fit, were removed accordingly.

2.4 Results – Experiment 1

Behavioural results are visualised in Figures 2.1 and 2.2. To assess whether retrocues affected behaviour in our task, I conducted one-way repeated-measures ANOVAs on decision time, response error, response variability and confidence. I found a main effect of cue type on decision time ($F(1, 19) = 21.41$, $p < 0.001$, pes = 0.530) and response error ($F(1, 19) = 10.66$, $p = 0.004$, pes = 0.359). Responses were faster in cued trials (M = 0.678s, SE = 0.063s) than neutral trials (M = 0.922s, SE = 0.0857s) and were more accurate in cued trials (M = 10.3 degrees, SE = 0.504 degrees) than neutral trials (M = 11.5 degrees, SE = 0.493 degrees). I also found a main effect of cue on response variability ($F(1,19) = 4.64$, $p = 0.044$, pes = 0.196), with lower variability (lower standard deviation) in the response distribution for cued trials (M = 0.167 degrees, SE = 0.00937 degrees) than neutral trials (M = 0.187 degrees, SE = 0.0103 degrees).

A key facet of this task was the inclusion of an introspective judgement on each trial of the experiment. Participants were asked to report their confidence in the working-memory response they had just made (their initial orientation reproduction). Critically, this judgement was made in the same circular space. To test whether response error and confidence were related across trials (and ensure participants were using the confidence judgement appropriately), I implemented a linear-mixed effects model to assess the

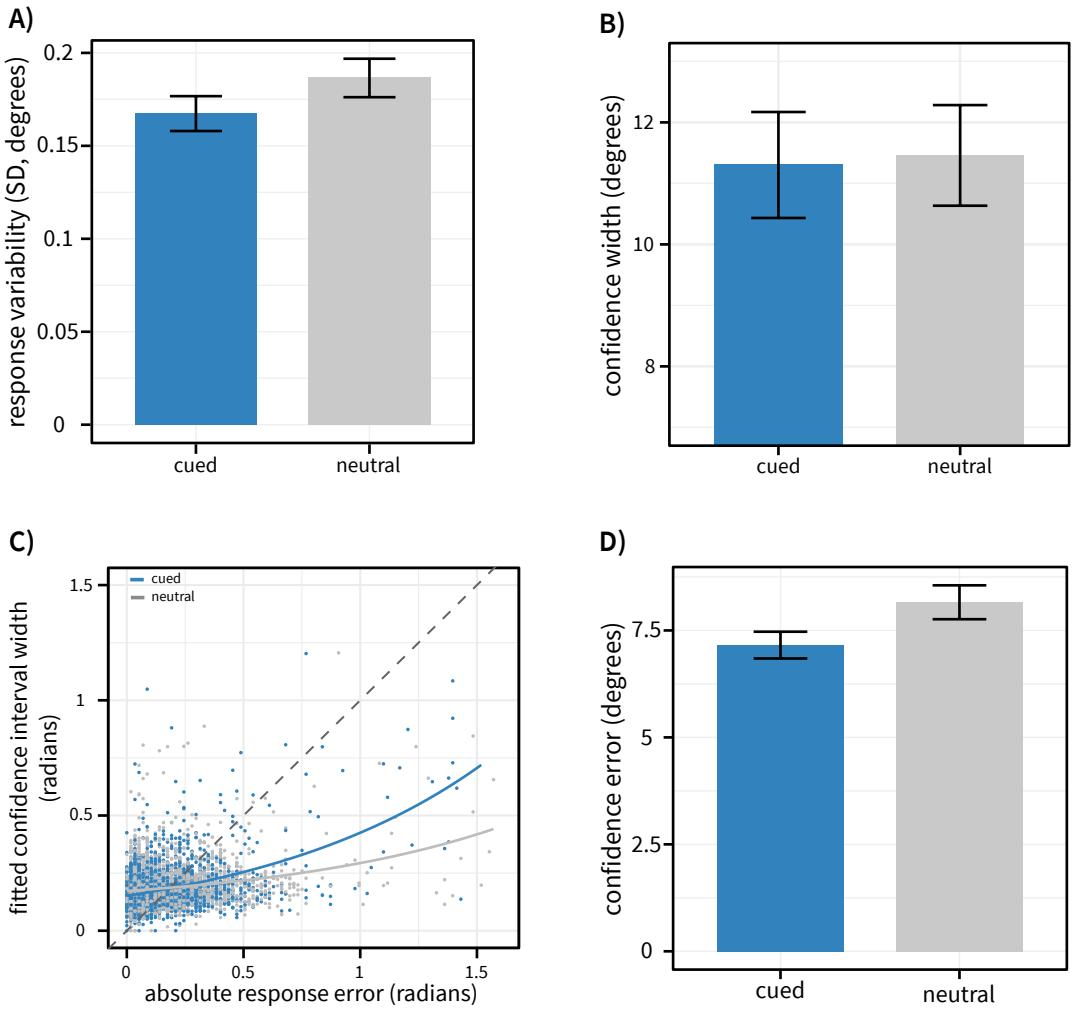

Figure 2.2 - Behavioural performance in Experiment 1. A) Average response variability (standard deviation of response errors in degrees) for cued and neutral trials. B) Average confidence wedge width (degrees) for cued and neutral trials. C) Relationship between response error (radians) and confidence wedge widths (radians) across trials and subjects, by cue type. Confidence was modelled as a function of response error, cue type (Neutral vs Cued) and their interaction, using linear-mixed effects modelling. Fitted confidence is visualised against response error, controlling for across-subject variance in effects. The dashed line represents the identity line, which would indicate perfect insight into memory quality. D) Average confidence error for cued and neutral trials. All error bars represent the standard error of the mean.

relationship between the two. This highlighted a positive association between response error and confidence across trials (see Figure 2.2C). Trials with higher response error were associated with larger confidence widths (beta = 0.651, t = 6.913, p = 1.04*10^{-6}). I also found that cue type (neutral or cued) interacted with this relationship (beta = -0.136, t = -3.127, p = 0.00178), suggesting that the relationship between error and confidence is influenced by

selective attention during the maintenance period. This was driven by a stronger association between error and confidence (more positive beta) in cued trials compared to neutral trials. To explore this further, I conducted one-way repeated-measures ANOVAs on confidence and confidence error. I saw no significant effect of cue type on confidence wedge widths (p = 0.459). However, there was a strong main effect of Cue on confidence error ($F(1,19) = 25.54$, $p < 0.001$, pes = 0.573), with confidence error lower in cued trials (M = 7.16 degrees, SE = 0.313 degrees) than neutral trials (M = 8.16 degrees, SE = 0.396 degrees).

2.5 Methods - Experiment 2

Participants

In this experiment I report the behavioural data from an EEG experiment. Twenty-six healthy human volunteers (17 female, age range 18-36, mean age 25.2 years, 2 were left-handed) participated in the study. The study involved two task sessions separated by a break, each comprising of 8 task blocks of 32 trials, resulting in a total of 512 trials. Six participants were excluded prior to data analysis: five completed only one session of task, and an additional participant was removed due to excessive noise in the raw EEG data. I do not report analyses of the EEG data in this chapter as this aimed to understand feedback processing and will be the focus of another chapter (Chapter 3). This resulted in a final sample of 20 participants (14 female, age range 18-36 years, mean age 25.3 years, 1 was left-handed). All participants reported normal or corrected-to-normal vision and not being

colour blind. All participants provided written informed consent prior to participation and were reimbursed £15 per hr for their time.

Stimuli, procedure and task

Participants were seated in an unlit, electrically-shielded booth at a distance of 100 cm from the monitor (27-inch, resolution: 1920 x 1080 pixels, screen width: 60 cm, refresh rate: 100 Hz). The experimental script was programmed using PsychoPy version 1.90.3 (Peirce et al., 2019). All code used to generate experimental stimuli and the task structure can be found at: https://github.com/schekroud/wmConfidence_Behaviour.

The structure of the task for Experiment 2 was similar to that of Experiment 1. However, in addition to the confidence judgement participants made, feedback relating to performance was presented on each trial. All stimulus durations remained the same but some delays between events changed. Figure 2.3 shows the task schematic with stimulus timings. The delay between the array offset and cue onset was reduced to 1s. The delay between cue and probe was fixed at 1.5 s and the delay between response confirmation and the confidence phase beginning was set to 1 – 1.5 s.

Uniquely to Experiment 2, a feedback routine was included on every trial after participants reported their confidence, providing comprehensive feedback on their response and confidence performance on every trial of the task. After clicking to confirm their confidence judgement, the visual display remained fixed. If a participant clicked to

confirm their confidence before the end of the two-second limit, their response remained on screen until the end of this limit, and a further one second before the feedback was presented. If a participant failed to confirm their response (i.e. it timed out after two seconds), there would still be a 1-s delay before feedback was presented. Participants were always shown their response and confidence width for at least one second. After this delay, the feedback was presented. Participants were shown the original target orientation as a thin bar (length: 5.8 dva, width: 0.1 dva, the same size and shape as the confidence interval boundary lines). This original orientation was coloured depending on the accuracy of the confidence judgement: if the target was outside the reported confidence interval, the bar was red (RGB: [204, 0, 0]); if the target was inside the interval, the bar was green (RGB: [0, 255, 0]). Therefore, the colour of the bar provided binary categorical correct/incorrect feedback about the confidence judgment, alongside the complete, continuous information available from the distance between the confidence width and the orientation of the feedback bar. This gave participants information with which to evaluate their own performance and determine whether the beliefs they held about their memory on a given trial were accurate (i.e. whether their confidence was justified). This allowed us to investigate whether participants could use this feedback information to adjust their confidence across trials. Feedback was presented for 500 ms before the visual display was replaced by the grey fixation cross for the duration of the ITI (1.5 – 2 s).

When a participant failed to confirm the initial orientation response within the two-

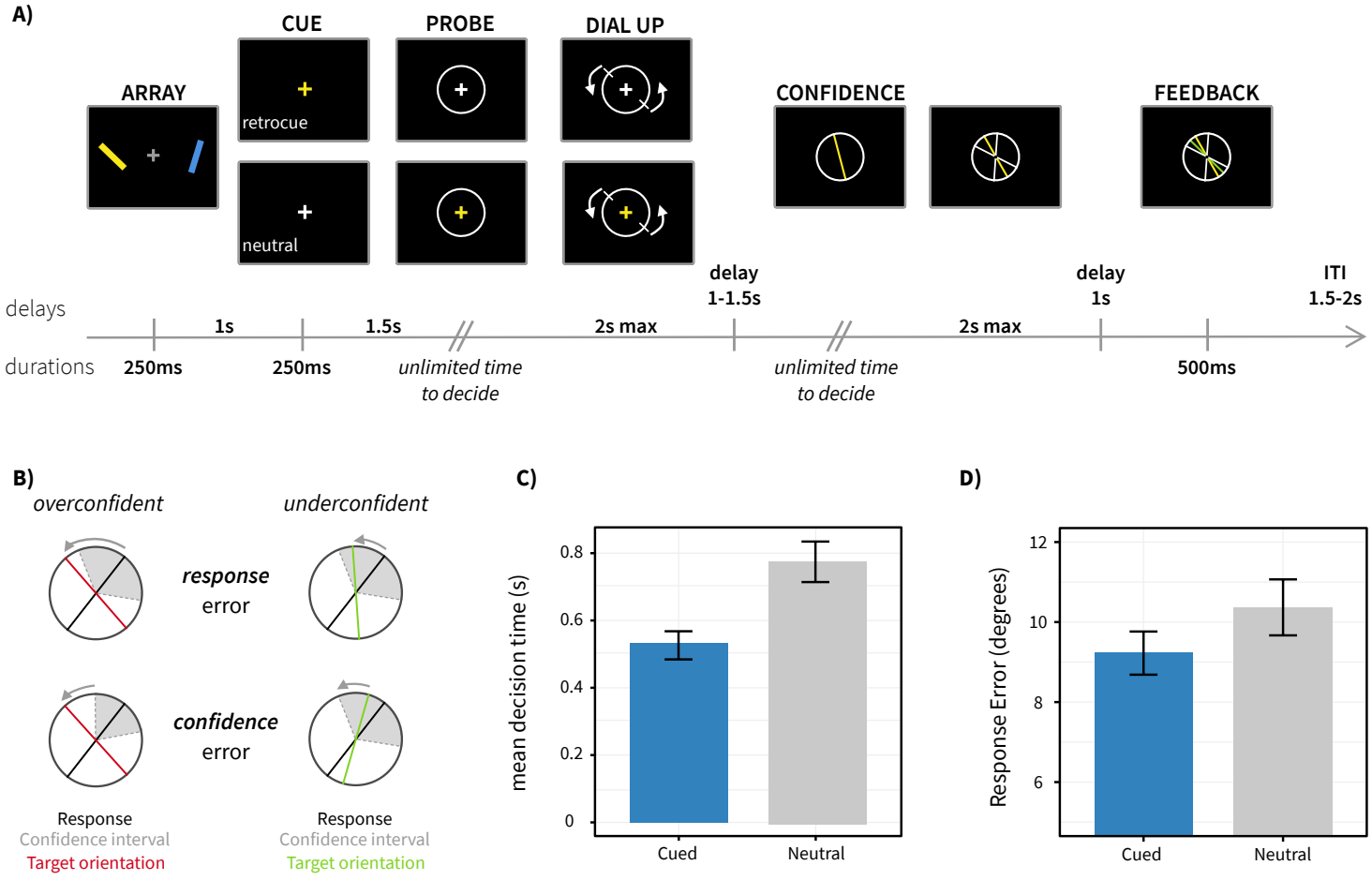

Figure 2.3 - Trial structure and behavioural performance in Experiment 2. A) Participants were presented with two coloured, oriented bars to memorise. After a delay, a cue was presented that was either informative (retro-cue) or uninformative (neutral cue) with respect to which item they would be probed about. Participants subsequently reported an orientation from memory. After reporting this orientation, they reported their confidence in their response, as in Experiment 1. After confirming their confidence response, participants were presented with feedback on their performance. Participants were presented with their responded orientation and their confidence wedge. The original target was then overlaid on top, coloured red or green depending on whether the target was outside or inside the reported confidence wedge (respectively). This gave binary information about the correctness of their confidence judgement, and continuous information regarding the accuracy of their orientation reproduction. B) Schematic visualising the two main types of error of interest and how they relate to coloration of feedback. When the original target orientation was outside the reported confidence wedge, feedback was coloured red. When the target orientation was inside the reported confidence wedge, feedback was coloured green. C) Average decision time (time taken to initiate the orientation reproduction, in seconds) for cued and neutral trials. D) Average response error (deviation between reported orientation and actual target orientation, in degrees) for cued and neutral trials. All error bars reflect the standard error of the mean.

second time limit, the last orientation on the screen was taken as the response and carried forward to the confidence phase to complete a trial. These trials were subsequently excluded from all analyses to ensure that only trials in which the participant explicitly reported an orientation were considered. At the end of each block, participants were provided with summary text informing them of their average reaction time (time taken to press the space bar to start the dial-up), their average response error (absolute angular distance between the target and response), their average tendency to be over- or under-confident, and the average degrees by which they were over- or under-confident.

2.6 Results – Experiment 2

To assess whether retrocues affected behaviour, I conducted one-way repeated-measures ANOVAs on decision time, response error, response variability, and confidence width. The behavioural results are depicted in Figures 2.3 and 2.4. I found a main effect of cue type on decision time ($F(1,19) = 66.32$, $p < 0.001$, pes = 0.78) and response error ($F(1,19) = 16.64$, $p = 0.0006$, pes = 0.47). Responses were faster in cued trials (M = 0.529 s, SE = 0.0415 s) than neutral trials (M = 0.775 s, SE = 0.0597 s) and were more accurate in cued trials (error: M = 9.22 degrees, SE = 0.539 degrees) than neutral trials (M = 10.4 degrees, SE = 0.7 degrees). I also found a main effect of cue on response variability ($F(1,19) = 5.46$, $p = 0.03$, pes = 0.22), with less variability (lower standard deviation) in the response distribution for cued trials (M = 0.149, SE = 0.0104) than neutral trials (M = 0.167, SE = 0.0117).

When analysing the confidence data, I found a main effect of cue type on the reported confidence wedge ($F(1,19) = 17.84$, $p = 0.0005$, pes = 0.48). Confidence was higher in cued trials (i.e. narrower wedges; M = 17.5 degrees, SE = 1.2 degrees) than in neutral trials (M = 18.4 degrees, SE = 1.30 degrees).

As the confidence judgement was made in the same circular space as the orientation response, it is possible to test for the relationship between these two judgements across all trials, allowing us to look at the error of participants' confidence judgements. As defined earlier, confidence error relates to the deviation between the original target orientation, and the edge of the confidence wedge that is reported on a trial. Due to how this value is computed, the sign of it holds meaning: positive values reflect over-confidence (target orientation was outside the confidence wedge) and negative values reflect under-confidence (target orientation was within the confidence wedge). The absolute confidence error therefore reflects the error in their confidence judgement, or how close they were to having perfect insight into their working memory. I found a main effect of cue when looking at absolute confidence error ($F(1,19) = 6.69$, $p = 0.02$, pes = 0.26), with absolute confidence error lower in cued trials (M = 10.8 degrees, SE = 0.815 degrees) than neutral trials (M = 11.2 degrees, SE = 0.846 degrees). Linear mixed-effects modelling further highlighted a positive association between response error and confidence wedge size. Trials with higher response error were associated with larger confidence wedges (beta = 0.574, t = 7.741, $p = 3.68 * 10^{-7}$). I also found an interaction of this relationship with cue type (beta = -0.0527, t = -2.263, $p = 0.0236$), showing that the relationship between error and confidence was influenced by

selective attention during the maintenance period. This was driven by a more positive beta (stronger association between error and confidence wedge size) in cued trials compared to neutral.

I investigated whether participants could use the continuous information in feedback to update their behaviour across trials by modelling (with a linear mixed-effects model) the relationship between the confidence error on a given trial and the change in confidence between that trial and the subsequent trial, while controlling for their confidence on the current trial (to account for regression to the mean). This showed a positive association between confidence error and changes in confidence across trials (beta = 0.0452 t = 5.651, $p = 1.64*10^{-8}$, shown in Figure 2.4E). This finding indicated that, across the experiment, participants broadened their confidence wedge (i.e., became less confident) following over-confident trials and narrowed their confidence wedge (i.e. became more confident) following under-confident trials.

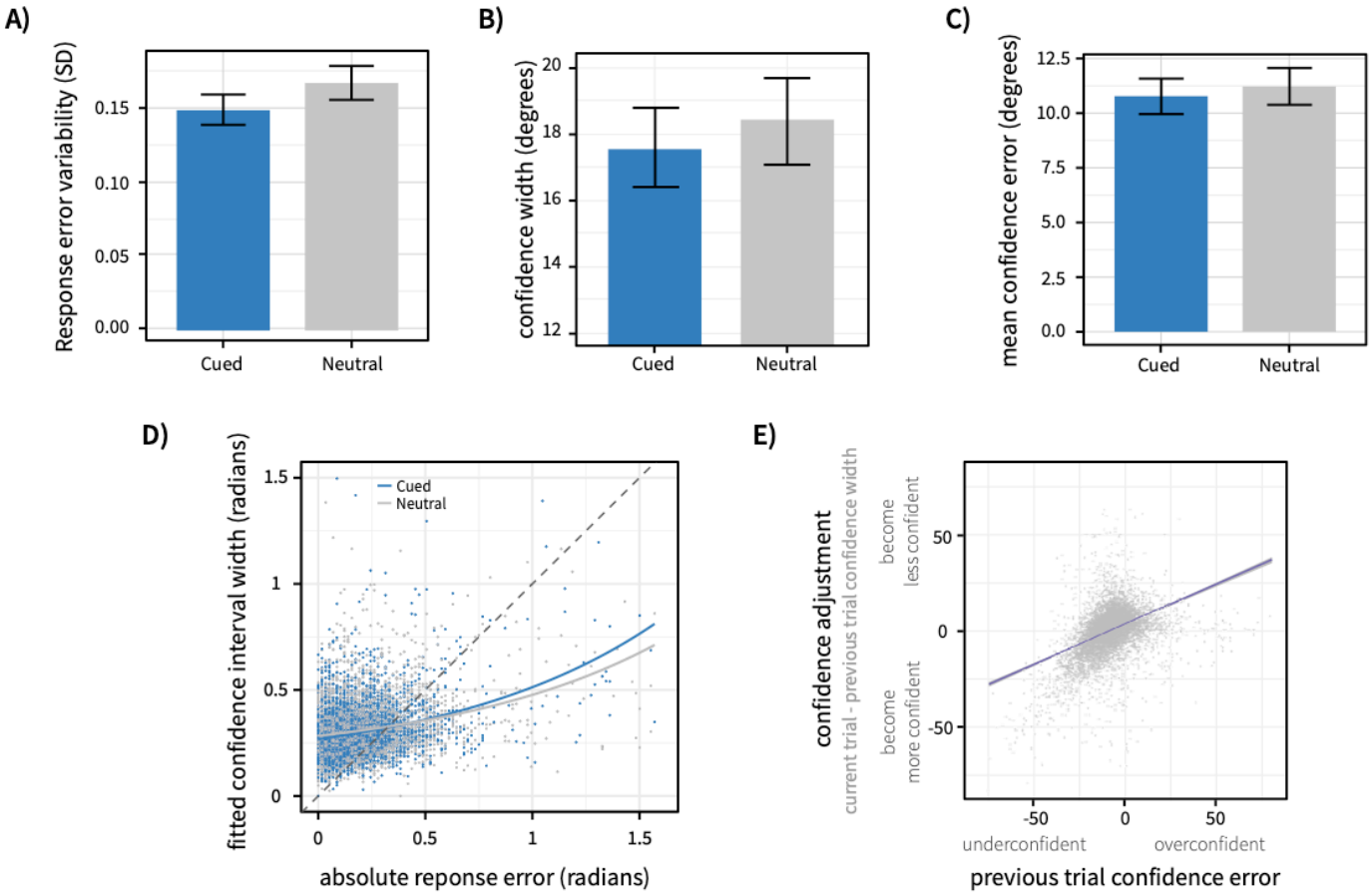

Figure 2.4 - Behavioural performance in Experiment 2. A) Average response variability (standard deviation of response errors, in degrees) for cued and neutral trials. B) Average confidence wedge width (degrees) for cued and neutral trials). C) Average absolute confidence error (how far away the target orientation was from the edge of the reported confidence wedge) for cued and neutral trials. D) Relationship between absolute response error (radians) and confidence wedge width (radians), for cued and neutral trials. Confidence wedge width was modelled as a function of response error, cue type (neutral vs cued) and their interaction, using linear mixed-effects modelling. Fitted confidence is visualised against response error, controlling for between-subject variance in effects. E) Visualisation of confidence updating across trials. Confidence error on the previous trial (degrees) is plotted against change in confidence between previous and current trial in degrees (i.e. whether confidence wedge widths increased or decreased based on confidence error on the previous trial). All error bars and shaded regions reflect the standard error of the mean.

2.7 Discussion

This study set out to address whether the deployment of attention to the contents of working memory shapes the way we introspect upon its contents. Across two experiments, I found evidence that orienting attention within working memory not only improves its quality, but improves the quality of insight into our internal representations.

I found that not only was the quality of memory associated with confidence in memory, but this relationship was, in part, influenced by attention. Participants reported orientations they had stored in memory, and subsequently made judgements about their confidence in the same continuous space, allowing us to directly link these two measures of behaviour. I found steeper slopes of association between working memory quality and confidence for trials where retrospective attention cues allowed the selection of task-relevant memoranda. The steepening of this slope suggests that the orienting of attention resulted in confidence judgements that were closer to true performance. I also found, across both experiments, that confidence error was lower in cued trials than neutral trials. Together, this suggests that representations of confidence are not static and that attention can improve the quality of introspections into working memories.

Confidence judgements made on the contents of working memory could be derived from multiple sources. Working memory is typically broken down into, and studied as, discrete phases: encoding, maintenance and retrieval. It seems intuitive that one could use heuristics about any of these stages to guide confidence judgements: if you feel you didn't see the items presented at the array very well, you may report lower confidence. If you feel

that holding things in mind during the delay period was more taxing or effortful, you may feel less confident in the orientation response that you make. If you felt it was difficult to retrieve the probed item from memory, you may use this effort as a heuristic and be uncertain in the response that you gave, reporting lower confidence on a given trial. It is unlikely that insight into the quality of memory is exclusively linked to one of these stages. For example, results from Honig and colleagues (Honig et al., 2020) suggest that insight into memory is not solely encoding driven. Specifically, they found that confidence reports in rewarded working-memory decisions cannot be entirely predicted just by stimulus features. My results support their conclusions, since benefits that retrocues bring to confidence performance necessarily occur after encoding. Metacognitive insight into the contents of working memory thus appears changeable depending on the manipulation and use of its contents.

I also found evidence that insight into working memory incorporates information from multiple sources, including recent trial history. Trials in which participants were overconfident were associated with a subsequent widening of reported confidence widths (becoming less confident), while trials where participants were underconfident were associated with narrowing of confidence widths (becoming more confident) on the subsequent trial. Thus, confidence judgements may not be exclusively bound to working-memory representations themselves, but can incorporate information from multiple sources (including recent confidence performance) to result in more accurate insights into memory quality.

Altogether, these findings suggest that not only are people able to make uncertainty judgements about the contents of their memory, but these judgements are influenced by selective attention. Further, these judgements are not fixed over time: judgements are in part shaped by recent performance – inaccuracies in reported confidence (over- or under-confidence in memory) are associated with subsequent changes in confidence that reflect the adaptive learning about our levels of insight into our mental contents. When we know we are performing better or worse, our beliefs about memory quality change adaptively. Critically, we were only able to examine such fine and dynamic changes over time thanks to the nature of the confidence judgement in this task. Confidence is commonly reported using arbitrary discrete scales. By allowing participants to report their exact level of uncertainty in their memory, I could examine the parametric association between memory quality and uncertainty, as well as how beliefs might be updated over time.

The question of what sources of information are used to construct these uncertainty estimates still remains. My findings suggest multiple information sources, but their precise nature and relative contributions will require future work. Dissociating the influence of changing encoding and maintenance phases would be informative here. For example, Honig and colleagues varied the underlying stimulus distributions to investigate the influence of stimulus features (drawing orientations from non-uniform distributions) on subjective uncertainty, providing an initial step to the question (Honig et al., 2020). Additional encoding-related manipulations would provide a fuller picture into how

encoding precision may contribute to confidence judgements. This could be done in a number of ways – e.g. by varying array presentation durations, using oriented Gabor stimuli with varying degrees of noise, or prospective attentional cues that guide attention at the point of encoding. This would further shed light on how encoding precision is factored into subjective uncertainty about the contents of working memory. Coupled with retrospective attentional cues (as in the current study) and other delay-period manipulations, it could be possible to disentangle the contributions of these working-memory processes towards confidence reports.

In these experiments, I focused specifically on the maintenance period and demonstrate that selective attention to memoranda in an internal space improves the quality of this memory and confers improvements in confidence. A natural question follows: do different kinds of attentional selection affect confidence in similar ways? We used a central cue to guide spatial selection in working memory. Other work has also shown that retrocues can also direct attention to features (or feature dimensions) of objects contained in memory (Niklaus et al., 2017) and contrasted item-based and feature-based selection directly (Hajonides et al., 2020). Integrating confidence judgements into working-memory tasks that investigate feature-based selection could help inform whether the effect of attention on confidence is driven by a spatial, item-specific reactivation, or related to the selection of the to-be-reported feature without requiring spatial information. Types of attentional selection that do not rely on sensory features (e.g. orientation, colour, location) encoded at array onset can also be explored, such as how reward-based attentional

orienting during the memory delay affects both memory quality and memory uncertainty. Reward-based cueing that is tied to an item (e.g. indicating one encoded item as more valuable than another) or not item-specific (more general regarding a potential reward that could be obtained on a given trial) could be explored along with their influence on both memory maintenance processes, and how the more general effects of reward could influence processing of feedback.

In this study I deliberately did not reward participants based on the confidence judgements that they make. There is an argument that rewarding participants based on the error in their confidence judgement (as in Honig et al., 2020) generates a trade-off that encourages accuracy in their responses – in this reward strategy, participants receive reduced numbers of points where the confidence interval is much wider than required to capture the target orientation. However, the prospect of reward on each trial can have attentional effects that may influence the endogenous characteristics of working memory – for example, generating stronger trial history effects than expected following erroneous trials due to also accruing fewer points. I aimed to look at endogenous insight into working memory in the absence of trial-wise reward-based motivation to determine the effect of attention without this potential extra contribution. However, whether and how cue effects are altered in the presence or absence of reward is an interesting question, or how reward might change the metacognitive effects that we observed, remain open questions to investigate.

In sum, across two behavioural experiments, I show that selective attention can influence metacognitive judgements made on the contents of working memory. I also provide evidence that participants track confidence across trials, with beliefs about the accuracy of confidence judgements flexibly updated over time. Confidence judgements on the contents of working memory provide a rich measure of behaviour that could better enable us to investigate fluctuations in memory performance and attention, and how people combine information about memory uncertainty when utilising memories for goal-directed behaviours.

3 - Neural signatures of error and confidence feedback processing during working-memory judgements using EEG

3.1 Abstract

Representational accounts of working memory consider it as a static information store regarding the past. More recently, research has shown that working memory is dynamic, and internal representations can be selected and prioritised to guide behaviour. While feedback has been used to guide learning of task structures in perceptual and value-based decisions, little is known about how feedback could facilitate learning about the quality of memories in order to better support memory-guided behaviour. Across two experiments I explore the electrophysiological processing of feedback regarding internally held representations. I found categorical differences in feedback processing depending on the valence of the feedback received, along with parametric representations of response error. Further, I show that the feedback response is not only sensitive to objective performance (error), but also to both external manipulations of uncertainty (experiment 1) and internal fluctuations in subjective uncertainty (experiment 2). These results extend upon previous research into feedback processing, providing evidence of evaluative processes regarding performance based on mnemonic information.

3.2 Introduction

The ability to process and learn from feedback is critical for adaptive behaviours. We learn from feedback continuously in daily life – for example, if informed that the meal you cooked for dinner last night wasn't particularly nice, you may not make it again in the near future. However, sometimes the feedback that we receive is more nuanced – beyond just categorical evaluations of good or bad performance, graded and multi-dimensional feedback can be used to allow more precise changes in behaviour during error-based learning (see Luft, 2014 for review). In this case, the meal itself may not have been the problem, but the particular balance of spices may have been sub-optimal. In other words, while categorical feedback allows an agent to tell whether or not to repeat a behaviour, graded feedback gives additional information that might be used to judge one's ability or make more fine-grained changes in approaching the task at hand. At the same time, sometimes worse performance is due to external changes in the environment (e.g., poorer quality of ingredients) or internal changes in one's own performance or ability.

Studies investigating feedback processing have often focused on the immediate external environment - on how individuals learn about reward structures, or on perceptual decisions. Rarely do these studies consider how feedback can enable us to better learn about the quality of memories. One often makes decisions based on stimuli experienced in the past - recall of which is an imperfect process. Within memory-guided decisions, there are various processes that could lead to both sensory and cognitive uncertainty, such as errors in encoding percepts into memory, degradation over time during maintenance, or

errors during recall. As a result, feedback in the real world also informs our beliefs about our own internal representations and performance, such as memory accuracy, or our ability to judge their quality. Despite the importance of knowing how much you can trust your own memory, it is unclear how previous findings about reward learning relate to feedback regarding the quality of recall from memory, or internally held representations more generally.

Thus, I explored feedback processing in the context of judgements made on internal representations – those held in working memory. Specifically, I examined the feedback processes linked to the quality of information reported from working memory and resulting confidence judgements. Across two experiments, I examined how uncertainty was reflected in the evoked neurophysiological response to feedback about internal representations when the quality of the internal representation was experimentally manipulated (experiment 1), and when it simply fluctuated internally but I carefully tracked those fluctuations trial by trial using precise confidence judgements (experiment 2).

To understand the underlying brain processes that integrate feedback and are used to change beliefs, researchers have relied on different kinds of neural recordings. In humans, electroencephalograpy (EEG) has been very useful for understanding what outcome features are rapidly integrated and when, to inform our beliefs about our own performance and the reward associations in the environment. One event-related potential (ERP) component that has helped provide insight into the ways in which feedback is processed is the feedback-related negativity (FRN).

The FRN was first described by Miltner and colleagues (Miltner et al., 1997) in a time-estimation task in which participants had to produce a 1-second interval. The potential classically manifests as a relative negativity over fronto-central sensors when contrasting the evoked response to incorrect vs. correct feedback (though this difference potential is also seen across more posterior midline sensors). The FRN is modulated by many different aspects of task structure rendering it a good candidate for a neural learning signal. While originally considered to be an error-related negativity, with emphasis on processing of feedback relating to poor performance, more recent converging evidence from EEG (Holroyd et al., 2008), MRI (Nieuwenhuis et al., 2005) and simultaneous EEG-fMRI (Becker et al., 2014) has suggested that the FRN is in fact generated by active processing changes of feedback relating to good performance. Further, feedback-evoked responses have been shown to be modulated by various task parameters such as reward probability (Cohen et al., 2007), magnitude (Yeung & Sanfey, 2004), the expectedness of the information the feedback contains (Holroyd et al., 2008; Pfabigan et al., 2011), or what type of information is being emphasised or presented at feedback (Nieuwenhuis et al., 2004). While there has been some change in what is considered to drive the FRN, there is consensus that there are categorical differences in the processing of positive and negative feedback. Altogether this highlights the sensitivity of the FRN to various aspects of task structure and performance, making it a useful tool to better understand feedback processing for memory-guided decisions.

Across two experiments, participants performed a free-recall working memory task where they were required to report the precise orientation of a bar previously encoded into working memory, while I simultaneously recorded the EEG. In the first experiment, I experimentally manipulated uncertainty externally by varying working memory load. Across blocks, participants were informed which of four coloured bars could be asked about within a block. At the end of each trial, participants were given categorical feedback regarding their performance. This manipulation of working-memory load has been shown to affect both objective working-memory performance (Bays et al., 2009; Luck & Vogel, 1997) and subjective uncertainty in working-memory guided responses (Rademaker et al., 2012). In a second experiment, participants were required to make a retrospective confidence judgement to report their subjective uncertainty in the response they made. At the end of each trial, participants were given continuous feedback regarding their exact performance on a trial, along with more categorical feedback regarding the correctness of their response. This allowed me to track participants' judgements about their own internal fluctuations in performance based on the feedback they received. Together, these experiments allowed me to examine how people monitor and evaluate their own performance during memory-guided decisions and how uncertainty is processed during such evaluations. I was also able to examine how both categorical and continuous information are represented at feedback and measure trial-wise behavioural correlates.

I aimed to replicate previous findings of feedback-related negativities in perceptual tasks, extending them to the processing of feedback regarding judgements based on

internal representations. I was able to explore whether the FRN may be driven by the processing of positive feedback by comparing categorical feedback differences in both tasks. Further, if uncertainty is represented in the feedback response, I would expect variability in the feedback ERP as a function of working-memory load (experiment 1), and trial-wise relationships between measured uncertainty (experiment 2) and the ERP. Across both experiments, I found evidence of the FRN reflecting the categorical difference in feedback between trials with good and poor performance. However, beyond a binary evaluation of performance at feedback, I also revealed evidence that the feedback-evoked response is sensitive to exact error across trials, and sensitive to uncertainty both when it is experimentally manipulated and when internal fluctuations in uncertainty are tracked across trials. Crucially, error-related processing at feedback was observed only in trials marked with good performance, suggesting that categorical differences in feedback processing of good and poor performance may be driven by more active processing of trials with positive feedback. Altogether, this provides evidence for the processing of feedback regarding working-memory guided decisions that may facilitate learning about both the objective quality and subjective uncertainty in memories to guide adaptive behaviours.

3.3 Methods – Experiment 1

Participants

Experimental procedures were reviewed and approved by the Central University Ethics Committee of the University of Oxford. Twenty-five healthy human volunteers participated in the study. Five participants were excluded prior to data analysis due to technical issues during recording. This resulted in a final sample of 20 participants (7 female, age range 19-35, mean age 24.9 years, 4 reported being left-handed). All participants reported having normal or corrected-to-normal vision, and not being colour-blind. All participants provided written informed consent prior to participation and were reimbursed £10/h for their time.

Stimuli, procedure and task

Participants were seated in an unlit, electrically shielded booth at a distance of 95 cm from the monitor (27-inch, 1920x1080 pixels resolution, 60 cm screen width, 100 Hz refresh rate). The experimental script was programmed using Presentation (Version 18.3, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com)

Participants performed 10 blocks in total, each comprising 60 trials, for a total of 600 trials. Each block consisted of three short sub-blocks (each of 20 trials) where load was manipulated. At the start of each sub-block, participants were presented with text on the screen that informed them of which colours were relevant for the sub-block they were in.

Participants could be informed that either one, two or all four colours were relevant. As such, working memory load was manipulated between sub blocks.

At the start of the trial, four coloured oriented bars appeared equidistant from a central fixation cross (RGB:[200, 200, 200]) for 500 ms. The centre of each bar was 5.7 degrees of visual angle (dva) in the horizontal and vertical axes, with a height and width of 5.7 and 0.8 dva respectively. Bars could be one of four different colours: blue ([0, 128, 255]), orange ([255, 127, 39]), magenta ([238, 0, 238]) or green ([0, 210, 63]) with orientations drawn independently from uniform distributions between 1 and 180 degrees. After a 1.5 s delay, a visual response dial appeared. The visual response dial was a circle (5.7 dva diameter) centred on the fixation cross. Two small circular handles were placed opposite each other on the radius of the circle, representing an orientation that participants adjusted in order to reproduce and orientation from memory. Participants had unlimited time to decide on the orientation they wished to reproduce, before moving the mouse to start the dial phase. Once they moved the mouse to start the response, participants had up to 2.5 s to rotate the dial and click to confirm their response. If they failed to confirm their response in this 2.5 s time window, the response would time out. During this response phase, a small rectangle appeared below fixation. This square filled up throughout the response phase to indicate the passage of time. After clicking to confirm their orientation reproduction, participants were presented with feedback. The fixation cross changed colour depending on the error in their response. For trials where the response error was over 20 degrees, the square was coloured red (RGB: [255, 0, 0]). For trials with response errors under 20 degrees,

the square was coloured green. The colour of this green feedback scaled with response error, such that trials with lower response errors were associated with a lighter green feedback stimulus. Feedback was presented for 200 ms, after which there was an inter-trial interval drawn randomly from between 500 and 800 ms. The task used in experiment 1, and associated timings, are represented in Figure 3.1A.

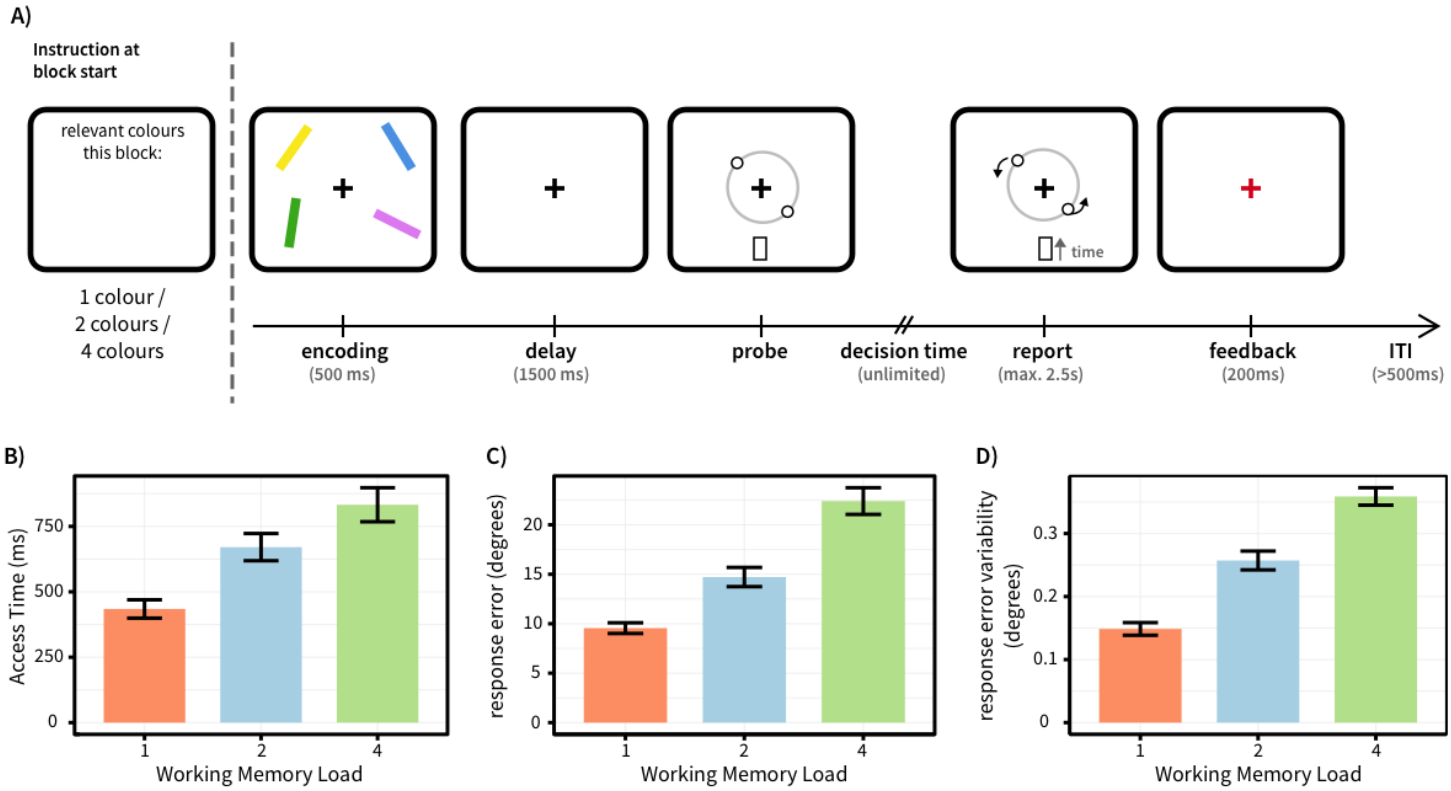

Figure 3.1 – Task design and behavioural results for Experiment 1. A) Schematic showing outline and timings for the task used in Experiment 1. Participants performed a precision working-memory recall task where they had to reported the exact orientation of an object encoded into working memory. At the beginning of each block, participants were informed which items were relevant, and thus could be asked about during the following block, manipulating working memory load. After each trial, participants received categorical feedback (red, bad vs green, good) regarding their performance on that trial. B). Access time (how long it took to retrieve an item from memory and start the report phase) as a function of working memory load. C) Response error as a function of working memory load. D) Response error variability (standard deviation of the response error distribution) as a function of working memory load. All error bars reflect the standard error of the mean.

Behavioural analysis

In this experiment I define one source of error. *Response error* is taken as the angular deviation between target orientation from memory and the response). During the probe phase, participants had unlimited time to retrieve the orientation from memory before moving the mouse to start the dial phase to report the orientation. I define *access time* as the time taken to retrieve the target orientation from working memory and initiate the dial phase. Trials were excluded from all analyses if their access times were over 5 seconds, or outside 2.5 standard deviations of the within-subject, within-condition mean. Aggregated data for conditions of the task were analysed using repeated-measures ANOVAs using the 'afex' package (Singmann et al., 2020) in the R statistical programming environment (R Core Team, 2020). All repeated-measures ANOVA include Load (Load 1, Load 2, Load 4) as within-subject factors to explore the behavioural consequences of increased uncertainty on working-memory performance.

EEG acquisition and processing

EEG data were collected using a SynAmps RT amplifier and Neuroscan acquisition Software (Compumedics Neuroscan, North Carolina, USA) using a 61 Ag/AgCl sintered electrode array (EasyCap, Herrsching, Germany) according to the 10-10 montage layout. Data were referenced to the left mastoid during recording and re-referenced to an average-mastoid reference offline. The ground was placed on the upper left arm, and two bipolar electro-oculograms (EOG) were recorded. One pair was placed above and below the left eye to

measure the vertical EOG and identify blinks, while the other was placed lateral of each eye to record the horizontal EOG and identify saccades. During acquisition, data were band-pass filtered between 0.1 and 200 Hz, digitised at 1000 Hz and stored offline for subsequent analysis.

During preprocessing, EEG data were subsequently imported, pre-processed and analysed in python using the MNE-python toolbox version 0.21 (Gramfort et al., 2013) and custom code. Raw data for each session were first band-pass filtered between 1 and 40 Hz before denoising was performed using Independent Component Analysis (ICA) using the 'InfoMax' routine to all EEG sensors. Components related to eye movements were detected by correlation with the horizontal and vertical EOG, and subsequently removed from the data.

ERP analysis.

Event-related potentials (ERPs) were analysed using Generalised Linear Models (GLMs) fitting an Ordinary Least Squares (OLS) regression (from the '*glmtools*' toolbox, https://pypi.org/project/glmtools/) to model the data. We used such a multiple regression approach for all ERP analyses rather than only relying on condition averages so that we could use parametric regressors such as behavioural parameters (e.g. response error) and test their relationship with evoked voltage changes in the EEG. This approach further allowed us to control other partially correlated factors by including them in the regression model. Bandpass filtered, epoched data were down-sampled to 500 Hz to reduce

subsequent computational demand, and baselined in the window 250 ms prior to event onset. GLMs were fit for each participant, separately for each channel at each time point, resulting in a beta weight for each time point at all channels. Beta coefficients for Contrasts of Parameter Estimates (COPEs) were used for subsequent analysis. Investigating the evoked response to feedback presentation, epochs were taken between -0.5 and 0.7 seconds around feedback onset.

Statistical analysis

I tested for significance using cluster-based permutation statistics (5,000 permutations, alpha = .05), which avoids the multiple-comparisons problem. On the ERP data, this generates a null distribution by randomly permuting the trial-average data at the participant level along the time dimension (and frequency, for spectral data). This was implemented using MNE-python's 1 sample permutation cluster test to test the observed group-level ERP against zero. In evaluating the feedback-evoked response, clustering was performed on the first-level beta weights for the contrast of interest at electrode FCz, allowing me to investigate behavioural correlations with the data while controlling for correlated variables.

3.4 Results – Experiment 1

Behaviour

To assess whether uncertainty (manipulated by varying working-memory load between task blocks) influenced behaviour in experiment 1, I conducted one-way repeated-measures ANOVAs on behaviour. I first examined whether changing working-memory load influenced the time taken to retrieve information from working memory. A one-way repeated measures ANOVA revealed a significant main effect of Load ($F(1.14, 21.71) = 62.1$, $p < 0.001$, $\eta_p^2 = 0.77$). This was driven by significant differences in access times for all levels of load (see Figure 3.1B), with shorter access time in Load 1 vs Load 2 ($t(19) = -7.49$, $p < 0.0001$), and shorter access time in Load 2 vs Load 4 ($t(19) = -7.21$, $p < 0.0001$).

I also found a significant effect of Load on response error ($F(1.25, 23.74) = 148.27$, $p < 0.0001$, $\eta_p^2 = 0.89$), with response error increasing as load increased (see Figure 3.1C). This is highlighted by significant differences in response error between Load 1 vs Load 2 ($t(19) = -9.41$, $p < 0.0001$), and between Load 2 vs Load 4 ($t(19) = -12.3$, $p < 0.0001$). Further, I found that changing working-memory load was also associated with differences in response error variability. A one-way repeated-measures ANOVA revealed a significant main effect of Load ($F(1.69, 32.04) = 195.41$, $p < 0.0001$, $\eta_p^2 = 0.91$). This was driven by increasing response error variability with increasing working-memory load (see Figure 3.1D). Response error variability was significantly lower for Load 1 vs Load 2 ($t(19) = -10.1$, $p < 0.0001$), with lower response error variability in Load 2 vs Load 4 ($t(19) = -12.2$, $p < 0.0001$).

EEG.

Analyses of feedback-evoked responses in perceptual tasks have previously demonstrated relative negativities over frontal and fronto-central midline sensors for negative feedback compared to positive feedback (Becker et al., 2014; Holroyd et al., 2008; Nieuwenhuis et al., 2004). I tested whether a similar pattern occurred when receiving feedback on judgements made based on internal representations. Further, I aimed to determine whether manipulating uncertainty by changing the demands upon working memory altered the way in which objective feedback on performance might be processed – namely, did the neural representation of feedback change as a function of how many items were held in working memory. I focus my analyses of the feedback period on electrode FCz based on these previous studies showing maximal feedback error-related negativities at this site.

I first fit a GLM to individual participants' single trial data including separate categorical regressors for correct and incorrect trials and each type of working memory load (see Equation 3.1). Condition differences were calculated at the contrast level to compare feedback processing of incorrect and correct trials with each level of working memory load.

$$EEG \sim \beta_1 * Correct + \beta_2 * Incorrect + \beta_3 * Load\ 1 + \beta_4 * Load\ 2 + \beta_5 * Load\ 4$$

Equation 3.1 – first model fit to the trial-wise EEG data per person to investigate the feedback-evoked response at all time-points and sensors

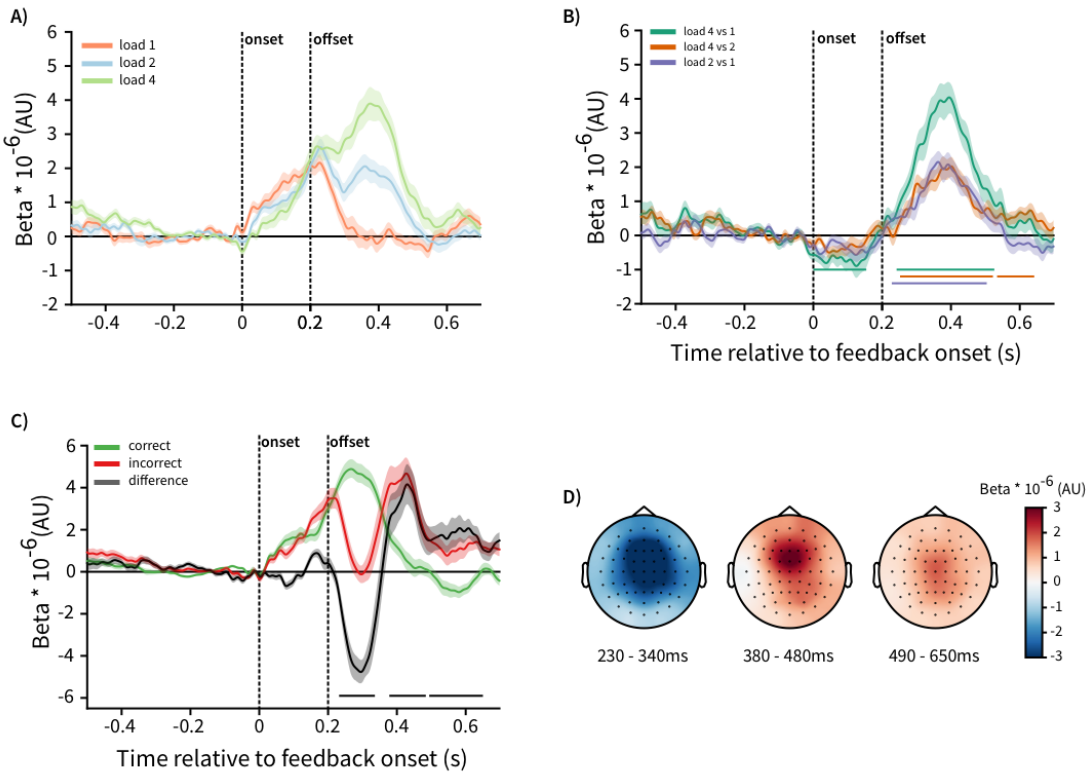

Figure 3.2 – Visualisations of the modelled feedback evoked responses. A) Grand-average modelled response to feedback separately for the load 1, load 2 and load 4 conditions. B) Differences computed between load conditions at the contrast level of the GLM, showing pairwise contrasts between the feedback-evoked response for all load conditions over time time. Coloured bars above the x-axis highlight significant time periods for the contrasts that survived permutation testing. C) Grand-average of the modelled response to correct (green line) and incorrect (red line) feedback. The difference between the two is represented in the black line, showing the presence of a feedback-related negativity. Black bars above the x-axis highlight significant time periods for the difference in feedback responses that survived permutation testing. D) Scalp topographies for significant clusters highlighted in Panel C are visualised to describe the scalp topography of these effects. All lines in A-C represent the mean beta weight for the

Figure 3.2A shows the grand-average beta weights showing the difference in the evoked-response to feedback for trials with differing working memory loads. The results show clear differences emerging between all three levels of load (one, two and four items) with graded separation in scalp voltage. Contrasts between the evoked response to feedback in the different load conditions are visualised in Figure 3.2B, highlighting clear differences in the feedback ERP for all levels of working memory load. Figure 3.2C shows that the familiar pattern of differences in processing of correct and incorrect performance

extended to feedback regarding internal working memory representations. Feedback for incorrect responses resulted in an initial relative negativity (230 – 340 ms after feedback onset), followed by later relative positivities (380 – 480 and 490 – 650 ms after feedback offset) over fronto-central sensors. Topographies for clusters that survived permutation testing are shown in Figure 3.2D.

To investigate feedback processing differences by load, I implemented a second GLM that separated the feedback-evoked response by correctness and by load (see Equation 3.2). This allowed me to model the feedback-evoked response to each type of trial separately and calculate differences at the contrast level.

$$\mathrm{EEG} \sim \beta_1 * \mathrm{Load}\ 1_{\mathrm{correct}} + \beta_2 * \mathrm{Load}\ 1_{\mathrm{incorrect}} + \beta_3 * \mathrm{Load}\ 2_{\mathrm{correct}} + \beta_4 * \mathrm{Load}\ 2_{\mathrm{incorrect}} + \beta_5 * \mathrm{Load}\ 4_{\mathrm{correct}} + \beta_6 * \mathrm{Load}\ 4_{\mathrm{incorrect}}$$

Equation 3.2 - second model fit to the trial-wise EEG data per person to investigate the feedback-evoked response at all time-points and sensors

Figure 3.3C shows clearly that the neural response seen when contrasting incorrect and correct responses occurs across all levels of working memory load. Further separating this contrast to evaluate correct and incorrect trials separately revealed no significant differences in the evoked-response to incorrect feedback as a function of load (see Figure 3.3A). However, clear differences emerged between load conditions when comparing the feedback response to correct trials. Significant clusters survived permutation testing for all three contrasts of load types for correct trials (see Figure 3.3D for visualisation).

In order to determine whether angular error across trials was associated with variability in the feedback-evoked response, and whether this representation of error varied across working memory load, a third GLM was fit to the data. I modelled the EEG data for each type of load and correctness separately and included separate regressors of angular error for each of these conditions (see Equation 3.3). This allowed me to model the variance associated with angular error beyond the main feedback ERP itself.

$$EEG \sim \beta_1 * Load\ 1_{correct} + \beta_2 * Load\ 1_{incorrect} + \beta_3 * Load\ 2_{correct} + \beta_4 * Load\ 2_{incorrect} + \beta_5 * Load\ 4_{correct} + \beta_6 * Load\ 4_{incorrect} + \beta_7 * Load\ 1\ Error_{correct} + \beta_8 * Load\ 1\ Error_{incorrect} + \beta_9 * Load\ 2\ Error_{correct} + \beta_{10} * Load\ 2\ Error_{incorrect} + \beta_{11} * Load\ 4\ Error_{correct} + \beta_{12} * Load\ 4\ Error_{incorrect}$$

Equation 3.3 – third model fit to the trial-wise EEG data per person to investigate the relationship between angular error and the feedback-evoked response at all time-points and sensors

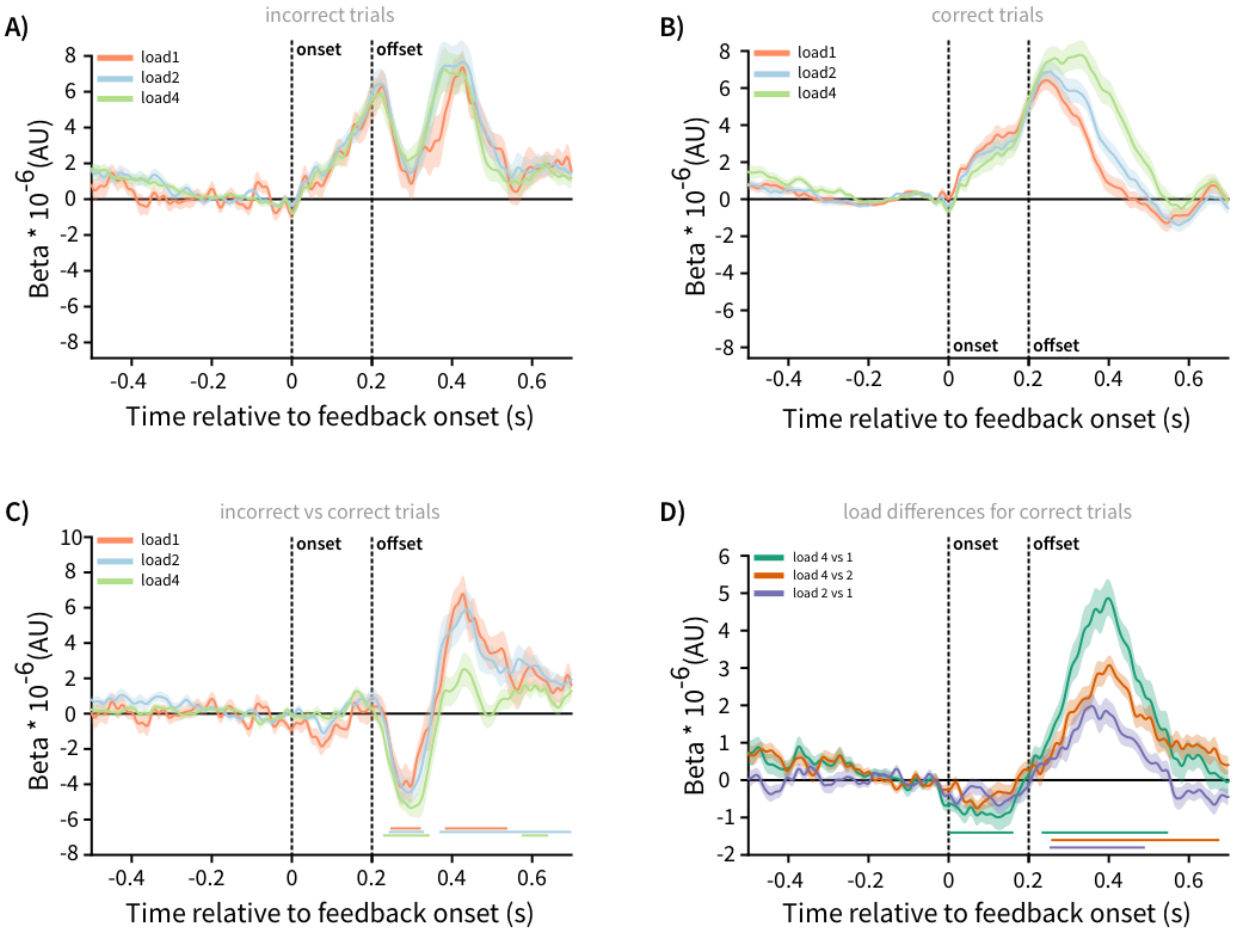

Figure 3.3 – Visualisations of the grand-average modelled responses to feeedback. A) Visualisation of the modelled response to feedback for incorrect trials, separately by load condition. B) Modelled response to feedback for correct trials, separately by load condition. C) Visulisation of the contrast between incorrect and correct feedback trials, separately for each load condition. Significant differences from zero are highlighted for each load condition by the coloured bars above the x-axis, showing time periods that survived permutation testing. D). Visualisation of the pairwise contrasts of the feedback response between load conditions for correct trials only. This highlights working memory load-related variability in the processing differences between incorrect and correct feedback. Coloured bars above the x-axis highlight time periods for pairwise contrasts that survive permutation testing. All lines represent the mean beta coefficient for the relevant regressor at electrode FCz, with shaded regions reflecting the standard error of the mean.

I show modulation of the feedback-evoked response by angular error for correct trials across all levels of working memory load – a significant early negative beta weight (reflecting that higher angular error was associated with lower scalp voltage), and a later positive beta weight (reflecting higher angular error was associated with more positive scalp voltages in this time period, see Figure 3.4A, topographies for these clusters are

visualised in Figure 3.4B). At the contrast level, I compared the relationship between the feedback ERP and error across different working memory loads. Significant differences in error modulation arose when comparing load four and load one (two clusters survived permutation testing: 240 – 340 ms and 430 – 500 ms after feedback onset) and when comparing load four and load two (one cluster survived permutation testing: 410 – 500 ms,

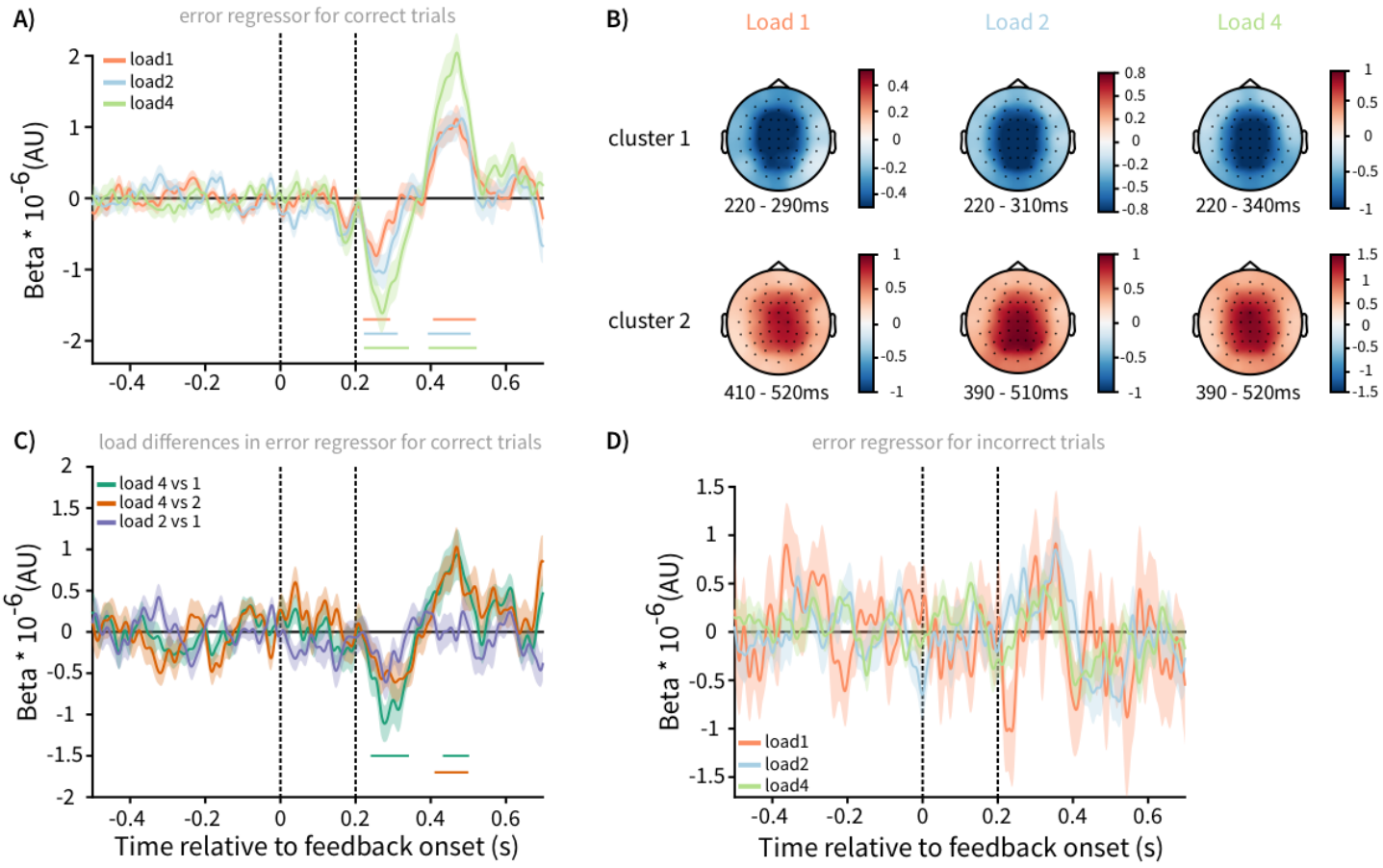

Figure 3.4 – Visualisations of the trial-wise relationship between the modelled feedback-evoked response and working memory error. A) grand-average of the error regressor for correct trials, showing the relationship between trialwise working memory error and scalp voltage over time, separately for each load condition. B) Visualisation of the scalp topography for clusters in A that survived permutation testing. C) pairwise contrasts of the error regressor between load conditions, showing differences in error processing between load conditions for correct trials only. Coloured bars above the x-axis highlight time periods of the contrasts that survived permutation testing. D). pairwise contrasts of the error regressor between load conditions for incorrect trials only. No significant time periods survived permutation testing. All lines represent mean regressor coefficients for the effect or contrast of interest at electrode FCz. Shaded regions represent the standard error of the mean.

see Figure 3.4C). No significant differences were found when comparing the error modulation for load two and load one. Further, I found no significant trial-wise association between the feedback-evoked response and angular error for incorrect trials (see Figure 3.4D).

3.5 Methods - Experiment 2

Participants

Experimental procedures were reviewed and approved by the Central University Ethics Committee of the University of Oxford. Twenty-six healthy human volunteers (17 female, age range 18-36, mean age 25.2 years, 2 reported being left-handed) participated in the study. Six participants were excluded prior to data analysis: five completed only one session of the task (the full task involved two sessions) and an additional participant was removed due to excessive noise in the raw EEG data. This resulted in a final sample of 20 participants. All participants reported having normal or corrected-to-normal vision, and not being colour-blind. All participants provided written informed consent prior to participation and were reimbursed £15/h for their time.

Stimuli, procedure and task

Participants were seated in an unlit, electrically shielded booth at a distance of 100 cm from the monitor (27-inch, resolution: 1920x1080 pixels, screen width: 60 cm, refresh rate: 100 Hz). The experimental script was programmed using Psychopy version 1.90.3 (Peirce et al., 2019).

At the start of the trial, two coloured oriented bars appeared on either side of a central fixation cross (RGB: [150, 150, 150]) for 250 ms. The centre of each bar was six degrees of visual angle (dva) from the centre of the screen, with a height and width of 5.7 and 0.8 dva respectively. Bars could either be blue (RGB: [0, 0, 255]) or yellow (RGB: [255, 255, 21]) with orientations drawn independently from uniform distributions between 0 and 179 degrees). After a 1-s delay, the fixation cross changed colour for 250 ms before returning to grey. On 50% of all trials, the fixation cross changed to match one of the items that were previously encoded into memory, signalling that only this item was relevant for subsequent report (*retrocue* condition). On the other half of trials, the fixation cross changed to white, providing no information regarding item prioritisation (*neutral* condition). After the cue, a fixed 1.5-s delay followed before the probe stimulus appeared on screen. On neutral trials the fixation cross changed colour to match the item that participants had to report. On retrocue trials, the fixation cross changed to white and participants would have to report the previously cued item (ensuring that participants processed the cue). A circle appeared around the fixation cross (diameter: 5.7 dva) to indicate the start of the probe phase. Participants had unlimited time to decide on the orientation before pressing the 'space' bar (with their left hand) to indicate they were ready to report the orientation. Upon pressing the space bar, two white lines appeared on opposite sides of the probe circle. The starting orientation for the dial was randomly drawn from a uniform distribution between 0 and 179 degrees. Participants used the mouse to rotate the dial around the circle to replicate the orientation of the probed item. A left mouse click was required to confirm their response.

As participants had an unlimited time to decide on the orientation, they were limited to two seconds to report and confirm their response.

After clicking to confirm the response, there was a variable delay of between 1 and 1.5s before the confidence phase began. Participants were presented with their orientation response as a bar inside a circle, with the colour matching the item they were asked about. Participants had unlimited time to decide how confident they were in the response they gave, before pressing the 'space' bar with their left hand to allow them to report their confidence. Upon pressing the space bar, two white lines appeared, mirrored symmetrically around the response orientation. Using the mouse, participants adjusted the width between these lines to represent their confidence interval in a circular space – larger widths corresponded to low confidence in their response, while narrower widths indicated high levels of confidence. Participants had up to two seconds to dial up and confirm their confidence report. After clicking to confirm their confidence judgement, the visual display remained fixed, keeping visual stimulation constant prior to the presentation of feedback to allow both pupillary and neural responses to return back to baseline levels. If a participant clicked to confirm their confidence before the end of the two-second limit, their response remained on screen until the end of this limit, and a further one second before the feedback was presented. If a participant failed to confirm their response (i.e. it timed out after two seconds), there would still be a 1 s delay before feedback was presented. Participants were always shown their response and confidence width for at least one second. After this delay, the feedback was presented. Participants were shown the original

target orientation as a thin bar (length: 5.8 dva, width: 0.1 dva, the same size and shape as the confidence interval boundary lines). This original orientation was coloured depending on the accuracy of the confidence judgement: if the target was outside of the reported confidence interval, the bar was red (RGB: [204, 0, 0]), while if the target was inside the interval it was coloured green (RGB: [0, 255, 0]), giving binary correct/incorrect feedback alongside the complete information of how responses linked to confidence. Feedback was presented for 500 ms before the visual display was replaced by the grey fixation cross for the duration of the ITI (1.5-2 s) to allow pupillary and neural responses to return to baseline.

Where a participant had failed to click to confirm their initial orientation response within the two-second time limit, the last orientation on the screen was taken as their response and carried forward to the confidence phase to allow them to complete a trial. These trials were subsequently excluded from all analyses to ensure that only trials where the participant explicitly reported an orientation were considered. At the end of each block, participants were informed of their average reaction time (time taken to press the space bar to start the dial-up), their average deviation (or error, the absolute angular distance between the target and their response), and whether they were on average over- or under-confident, and by how many degrees. Participants completed two sessions of 8 blocks (of 32 trials), separated by a break between sessions, resulting in a total of 512 trials. The task used in Experiment 2, and its associated timings, is shown in Figure 3.5A.

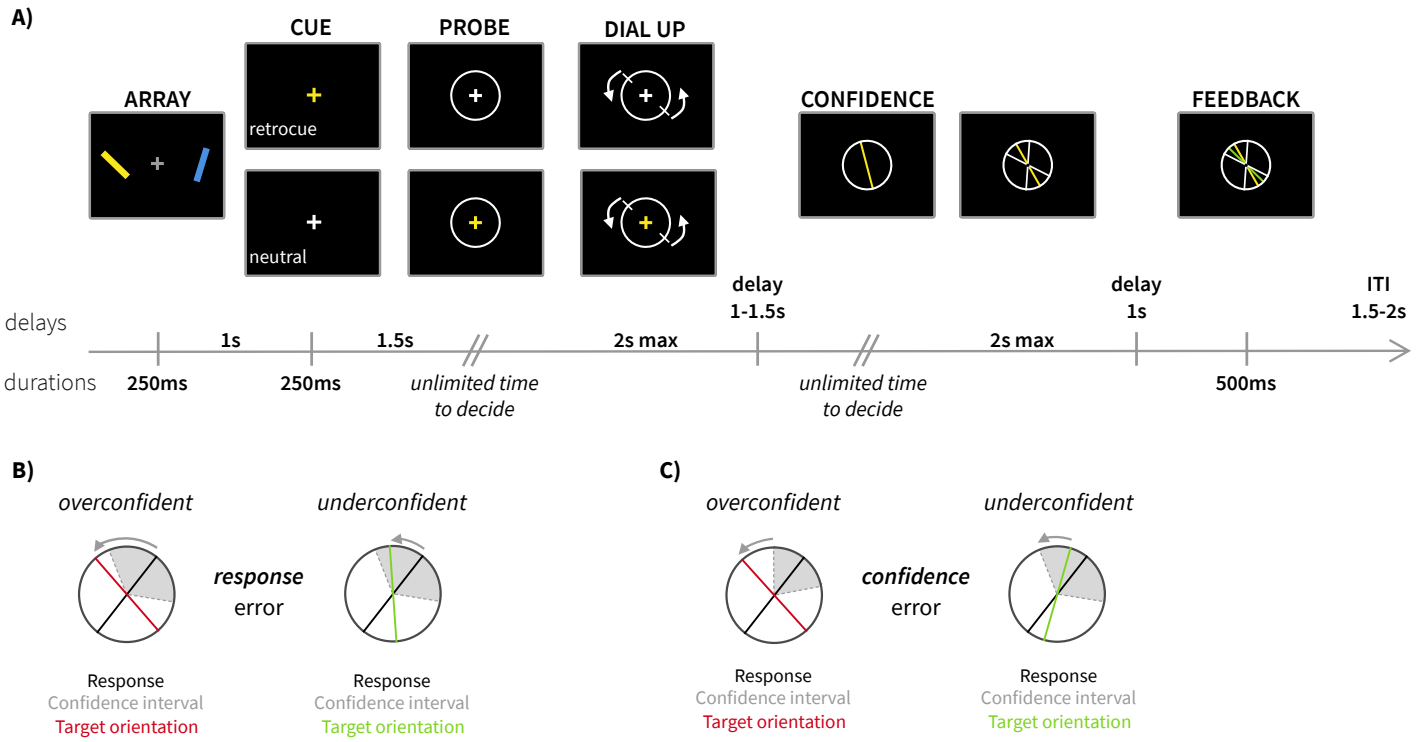

Figure 3.5 - A) Schematic showing the outline and timings for the task used in experiment 2. Participants engaged in a precision-working memory recall task where they had to report the orientation of a bar encoded into memory. Participants were, on 50% of trials, cued in advance to inform them which orientation they were required to report. Participants reported their precise confidence in their report from memory, before receiving feedback on their performance. Panels B and C visualise the two sources of error we investigated. B) Visualisation of response error – the deviation between the reported orientation and the true target orientation. C) Visualisation of confidence error – the deviation between the true target orientation and the edge of the reported confidence wedge. If the target orientation lay outside of this wedge (overconfident), participants received red (incorrect) feedback. If the target orientation lay inside this wedge (correct, or underconfident), participants received green feedback.

EEG acquisition and processing

EEG data were collected using a SynAmps RT amplifier and Curry 8 Software (Compumedics Neuroscan, North Carolina, USA) using a 61 Ag/AgCl sintered electrode array (EasyCap, Herrsching, Germany) according to the 10-10 montage layout. Data were referenced to the left mastoid during recording and re-referenced to an average-mastoid reference offline. The ground was placed on the upper left arm, and two bipolar electro-oculograms (EOG)

were recorded. One pair was placed above and below the left eye to measure the vertical EOG and identify blinks, while the other was placed lateral of each eye to record the horizontal EOG and identify saccades. Data were sampled at 1000 Hz during acquisition and stored for subsequent analysis.

During preprocessing, data were converted from the native CDT format to SET format using the EEGLAB toolbox and custom code written in Matlab (version 2015b, Mathworks). EEG data were subsequently imported, pre-processed and analysed in python using the MNE-python toolbox version 0.21 (Gramfort et al., 2013) and custom code. Raw data for each session were first band-pass filtered between 1 and 40 Hz. Denoising was then performed using Independent Component Analysis (ICA) using the 'InfoMax' routine to all EEG sensors. Components related to eye movements were detected by correlation with the horizontal and vertical EOG and subsequently removed from the data. I discarded trials if blinks were detected around the array onset and cue onset period (as blinks in these periods would prevent processing of task-relevant information, and thus influence behaviour). An automated routine was implemented to detect blinks around these two events of interest. 500-ms epochs were taken from the EOG traces around these events, bandpass filtered between 1 and 40 Hz, and baselined to the 250 ms prior to the event. For each epoch, the difference between the peak and minimal voltage in the vertical EOG was calculated and z-scored across epochs. Epochs containing blinks were identified as those with z-scored difference values larger than two and the raw difference values greater than 250 microvolts. This was done separately for both cue and array events. Trial identifiers

containing blinks were saved to mark trials to be excluded from further analyses. Data were subsequently epoched around events of interest, and trials were removed based on within-trial variance using an automated procedure that looks at the variance in the broadband signal at a significance threshold of 0.05 (generalised extreme studentized deviate test, GESD). This procedure was run separately for each event of interest. Trials were subsequently visually inspected to identify any remaining trials with excessive noise, for example in the baseline period, to be removed from analysis. I only retained trials in which participants explicitly confirmed their orientation reproduction response (clicking to confirm at the dial up phase) and decision time was within 2.5 standard deviations of the mean.

ERP analysis

Event-related potentials (ERPs) were analysed using Generalised Linear Models (GLMs) fitting an Ordinary Least Squares (OLS) regression (from the '*glmtools*' toolbox, https://pypi.org/project/glmtools/) to model the data. I used such a multiple regression approach for all ERP analyses rather than only relying on condition averages so that I could use parametric regressors such as behavioural parameters (e.g. confidence judgements) and test their relationship with evoked voltage changes in the EEG. This approach further allowed me to control other partially correlated factors by including them in the regression model. Bandpass filtered, epoched data were down-sampled to 500 Hz to reduce subsequent computational demand, and baselined in the window 250 ms prior to event

onset. GLMs were fit for each participant, separately for each channel at each time point, resulting in a beta weight for each time point at all channels. Beta coefficients for Contrasts of Parameter Estimates (COPEs) were used for subsequent analysis. Investigating the evoked response to feedback presentation, epochs were taken between -0.5 and 1 second around feedback onset.

Spectral analysis.

I estimated spectral power at frequencies using a Morlet wavelet decomposition. For the cue-locked period, I took epochs from -0.5 to +2 s to capture the entire maintenance period after cue presentation, and downsampled the data to 100 Hz. The single-trial data were convolved with Morlet wavelets between 1 and 40 Hz. For each frequency, I used a fixed 300-ms time window such that the number of cycles changed with the frequency being estimated. Where necessary to do so, time-frequency resolved data were baselined between -2 and -1.5 s prior to the event of interest (functionally equivalent to a 500-ms period prior to the presentation of the array, at the start of the trial). Subsequent analyses used GLMs (fit separately at each sensor and time point) to investigate stimulus-induced responses and their relationship with behaviour. For statistical inferences, the output beta-weights for contrasts of interest were used as the input for cluster-based permutation testing. Visualisations of the time-frequency analysis reflect the t-statistics for beta weights across participants.

Statistical analysis

I do not report the results of behavioural analyses of Experiment 2 here, as we report those in a different chapter (Chapter 2, Experiment 2). Here, I focus on the analysis of neurophysiological signals relating to feedback processing.

I tested for significance using cluster-based permutation statistics (5,000 permutations, alpha = .05), which avoids the multiple-comparisons problem. On the ERP data, this generates a null distribution by randomly permuting the trial-average data at the participant level along the time dimension (and frequency, for spectral data). This was implemented using MNE-python's 1 sample permutation cluster test to test the observed group-level ERP against zero. In evaluating the feedback-evoked response, clustering was performed on the first-level beta weights for the contrast of interest at electrode FCz, allowing me to investigate behavioural correlations with the data while controlling for correlated variables.

3.6 Results – Experiment 2

EEG results

I focused the analyses on the main manipulations of interest, namely the retrospective orienting of attention during the cue period and the processing of performance and confidence errors during the feedback period.

Orienting attention in WM.

Replicating various findings of spatial orienting during working memory maintenance, a GLM-based time-frequency analysis investigating induced brain activity (1-40 Hz) over lateral posterior electrodes (PO3/4, PO7/8, O1/2) showed significant lateralisation of power in the alpha-band range, with lower alpha power contralateral to the retrocue-induced locus of spatial attention ($p < 0.05$, see Figure 3.6). Notably, I found that induced modulation of alpha power was transient, rather than sustained during the entire maintenance period (similar to e.g. Mok et al., 2016; Myers et al., 2015).

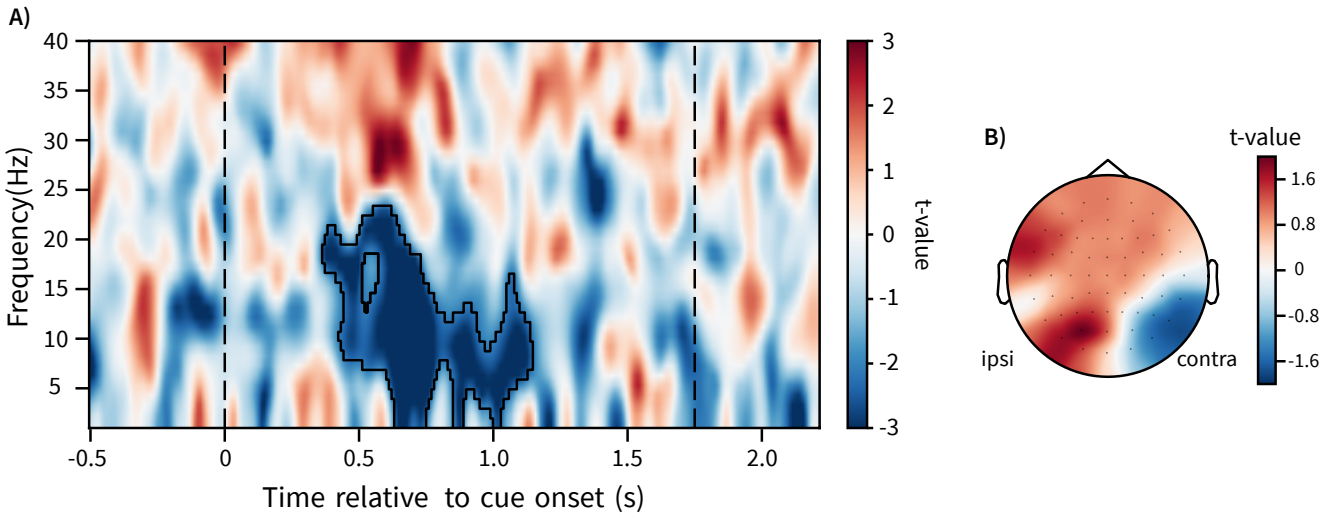

Figure 3.6 – Visualisation of cue-induced modulations in neural activity. A) Time-frequency representation showing the lateralisation (contra- vs ipsi-lateral) of modelled activity in visual sensors when comparing cued left vs cued right trials. Values represent the t-stat across participants of the first level t-stats of the contrast. Black outlines highlight regions that survived permutation cluster testing. B) visualisation of the scalp topography of alpha-frequency power for the cued left vs cued right trials in the time range of the cluster surviving permutation testing in A. This shows lower relative alpha power over right occipital sensors for cued left trials compared to cued right trials.

Feedback

Analysis of the feedback-evoked response aimed to replicate feedback-related negativity, and relationships with working-memory error observed in Experiment 1. Furthermore, I explored associations between feedback-evoked responses and behaviour when graded feedback is presented regarding working memory performance.

I fit GLMs to individual participants' single trial data including separate categorical regressors for the category of response (correct or incorrect) and additional parametric regressors for the degree of performance error and degree of confidence in these two categories. A regressor for the original side of the probed item was also included to account for any possibly lateralised effects (the full regression model for this analysis is depicted in Equation 3.4). Condition differences were calculated at the contrast level.

$$EEG \sim \beta_1 * Correct + \beta_2 * Incorrect + \beta_3 * Error_{correct} + \beta_4 * Error_{incorrect}$$
$$+ \beta_5 * Confidence_{correct} + \beta_6 * Confidence_{incorrect} + \beta_7 * Probed\ Side$$

Equation 3.4 – first model fit to the trial-wise EEG data per person to investigate the feedback-evoked response at all time-points and sensors

Figure 3.7 shows the grand-average beta weights showing the effects of the two categorical feedback types across participants. The results replicate the familiar pattern of differences in processing of correct vs. incorrect performance, extending this to continuous feedback rather than feedback valence alone. Feedback for incorrect responses resulted in

an initial relative negativity (120-180 ms after feedback onset), followed by a later relative positivity over fronto-central sensors (310–600 ms after feedback onset)

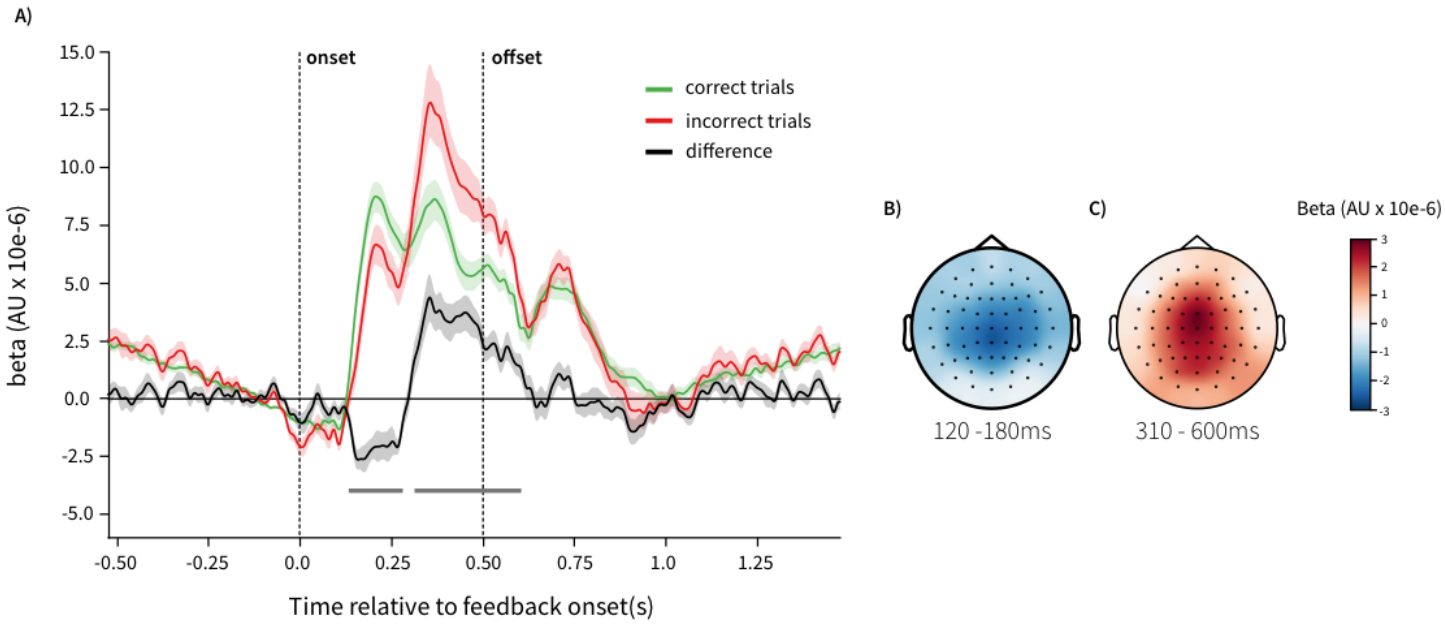

Figure 3.7 – Visualisation of the modelled response to feedback at electrode FCz. A) Grand-average modelled response to feedback for incorrect trials (red), correct trials (green) and their difference (black). Black bars above the x-axis outline time periods of the difference contrast that survived permutation cluster testing against zero. ERP traces represent the across-participant mean beta coefficient for the condition or contrast of interest, shaded regions represent the standard error of the mean. B and C) Scalp topographies for the clusters in panel A visualising the topography of the difference across sensors.

<u>Variability in the feedback ERP is associated with error and confidence in working memory</u>

Associations between behaviour and the modelled feedback response are shown in Figure 3.8. Results are plotted at electrode FCz, where feedback-related ERPs are typically reported to be maximal.

To investigate how these feedback-induced differences in scalp voltage were associated with changes in behaviour, I examined the output beta weights for the parametric regressors of error and confidence at this electrode.

A significant association occurred between the feedback-evoked response and angular response error for correct trials, around the same time frame as the FRN. This was found 150-220 ms after feedback onset, (the scalp topography for this significant cluster is visualised in Figure 3.8B). This negative association reflects that in trials where the target orientation was within reported confidence interval (receiving correct, green feedback), larger negativities were seen for trials with larger response error. No significant clusters emerged in the association between the feedback ERP and error for incorrect trials, or for the difference contrast between trial types.

A significant association also occurred between subjective confidence and the feedback ERP across trials, for both correct and incorrect feedback types (see Figure 3.8C). Two significant clusters emerged for the modelled response of correct-feedback trials (230-330 ms and 360-460 ms post feedback onset), with positive betas reflecting that greater confidence (narrower confidence interval) was associated with larger amplitudes in these time ranges. We also saw a significant cluster of positive betas emerge after cluster permutation testing for the relationship between confidence and the ERP in incorrect trials (310-410 ms post feedback onset), with larger amplitudes in this time range for trials with higher confidence (narrower confidence ranges). No significant clusters emerged from the

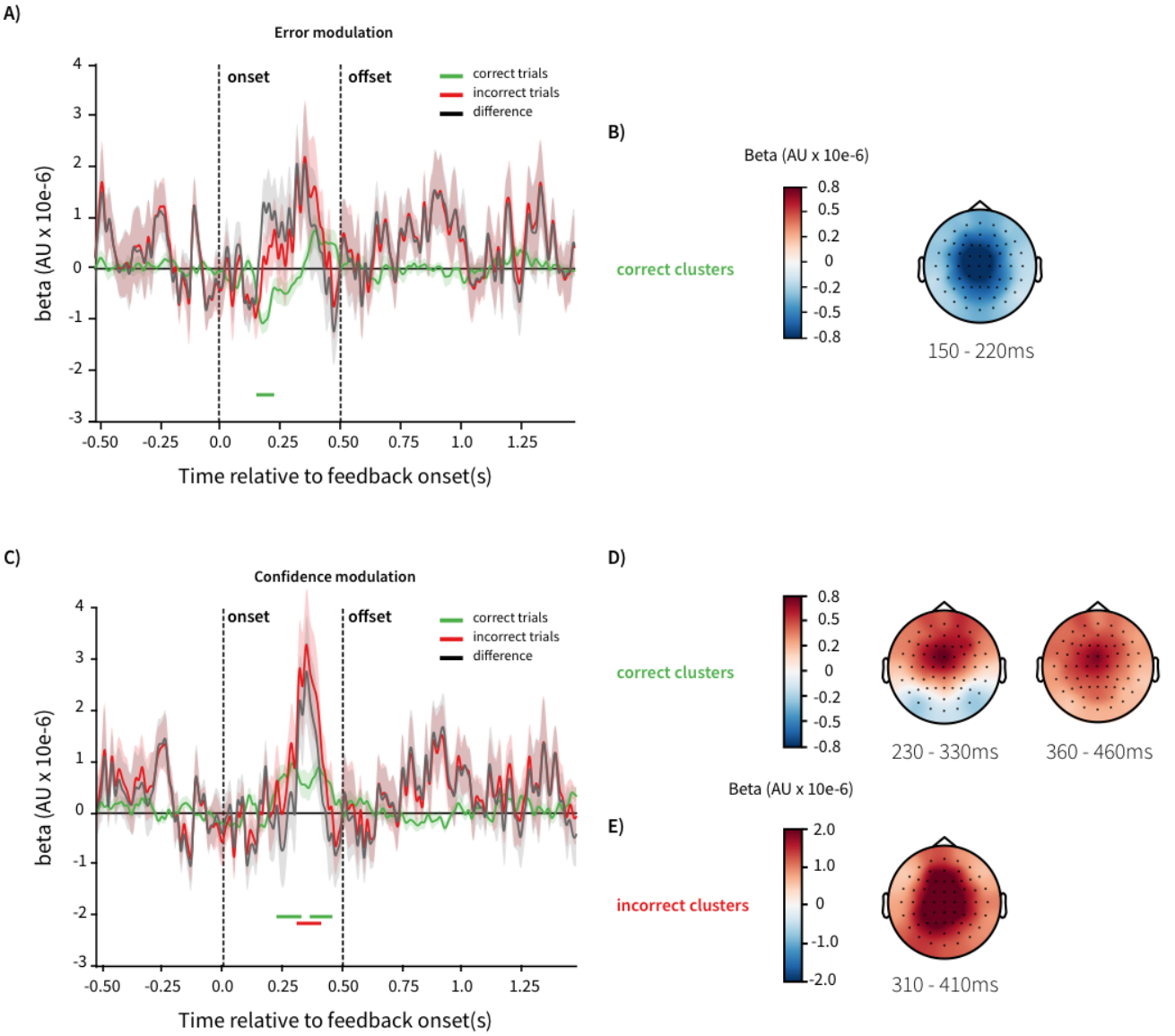

Figure 3.8 – Visualisation of the trial-wise relationship between the feedback evoked response and behavioural performance. A) Modelled relationship between the feedback-evoked response and working memory error, separately for incorrect (red) and correct trials (green). The difference between the two (calculated at the contrast-level) is represented in black. Coloured bar above the x-axis represents time periods that survived permutation testing for the relevant effect of interest. B) Visualisation of scalp topography for the error regressor cluster in correct trials that survived permutation testing. C) Modelled relationship between the feedback-evoked response and subjective confidence in working memory report, separately for incorrect (red) and correct (green) trials. The difference between trial types (calculated at the contrast-level) is depicted in black. Coloured bars above the x-axis highlight significant time periods that survived permutation testing against zero. D) Visualisation of scalp topographies for the two significant clusters in the confidence regressor for correct trials that survived permutation testing. E) Visualisation of scalp topography for the significant cluster in the confidence regressor for incorrect trials that survived permutation testing. All solid lines in A and C represent the across-participant mean beta coefficient for the effect of interest at electrode FCz, shaded regions represent the standard error of the mean.

contrast of these two feedback types. Together, these findings suggest that error is

represented in the neural response to feedback for correct trials – despite 'correct' feedback as the target orientation was contained within the confidence judgement. Likewise, ERP variance tracked confidence for trials with both feedback types, suggesting that regardless of the binary content (correct vs incorrect) of the feedback, when full information was presented regarding behaviour, the brain still represents this information when available.

Trial-wise variability in the feedback ERP is associated with confidence error.

I analysed the feedback period with a second model to test for the existence of a neural marker of confidence error at feedback that could be used to guide trial-wise adjustments of confidence in subsequent behaviour. In this model I included separate categorical regressors for feedback type (underconfident or overconfident), and parametric regressors of signed confidence error for the two feedback valences. I also included a regressor for the original side of the probed item to control for any lateralised effects (the full regression formula for this analysis is depicted in Equation 3.5).

$$EEG \sim \beta_1 * Correct + \beta_2 * Incorrect + \beta_3 * Confidence\ Error_{correct} + \beta_4 * Confidence\ Error_{incorrect} + \beta_5 * Probed\ Side$$

Equation 3.5 - second model fit to the trial-wise data per subject to investigate representations of metacognitive error in the feedback-evoked response

Two potential models for the representation of metacognitive error (confidence error) are considered, with potential relationships outlined in Figure 3.9A – 3.9D. In one possibility, *absolute* confidence error could be represented, reflecting the unsigned confidence error – the absolute distance between the original orientation and the edge of the reported confidence width. If confidence error were to be represented in this unsigned way, it would be agnostic to whether performance was over- or under-confident (Figure 3.9A). For this to be the case, one would expect the inverse sign for the beta weights in underconfident and overconfident trials (see Figure 3.9C), reflecting an equivalent representation of overconfident and underconfident errors of the same magnitude (e.g., -20 and 20 degrees). In the alternative, *signed* confidence error could be represented (see Figure 3.9B), which would indicate that the direction of confidence mattered (e.g. 20 degrees overconfident and underconfident would be represented differently). In this scenario, an inverse sign between conditions would not be expected in order to represent the direction of confidence error and not just its magnitude (see Figure 3.9D).

Two significant clusters emerged for the association between confidence error and the ERP at electrode FCz across trials. I found a significant cluster when evaluating underconfident trials only (360-460 ms after feedback onset). I visualise the topography of this cluster in Figure 3.9F. Crucially, I found no evidence of a sign-flip between the output beta weights of confidence error in underconfident and overconfident trial types. This suggests that the representation of confidence error at feedback is signed, representing

both the degree of confidence error (how far away the target orientation was from the edge of the confidence interval) and direction (whether the trial was over- or under-confident).

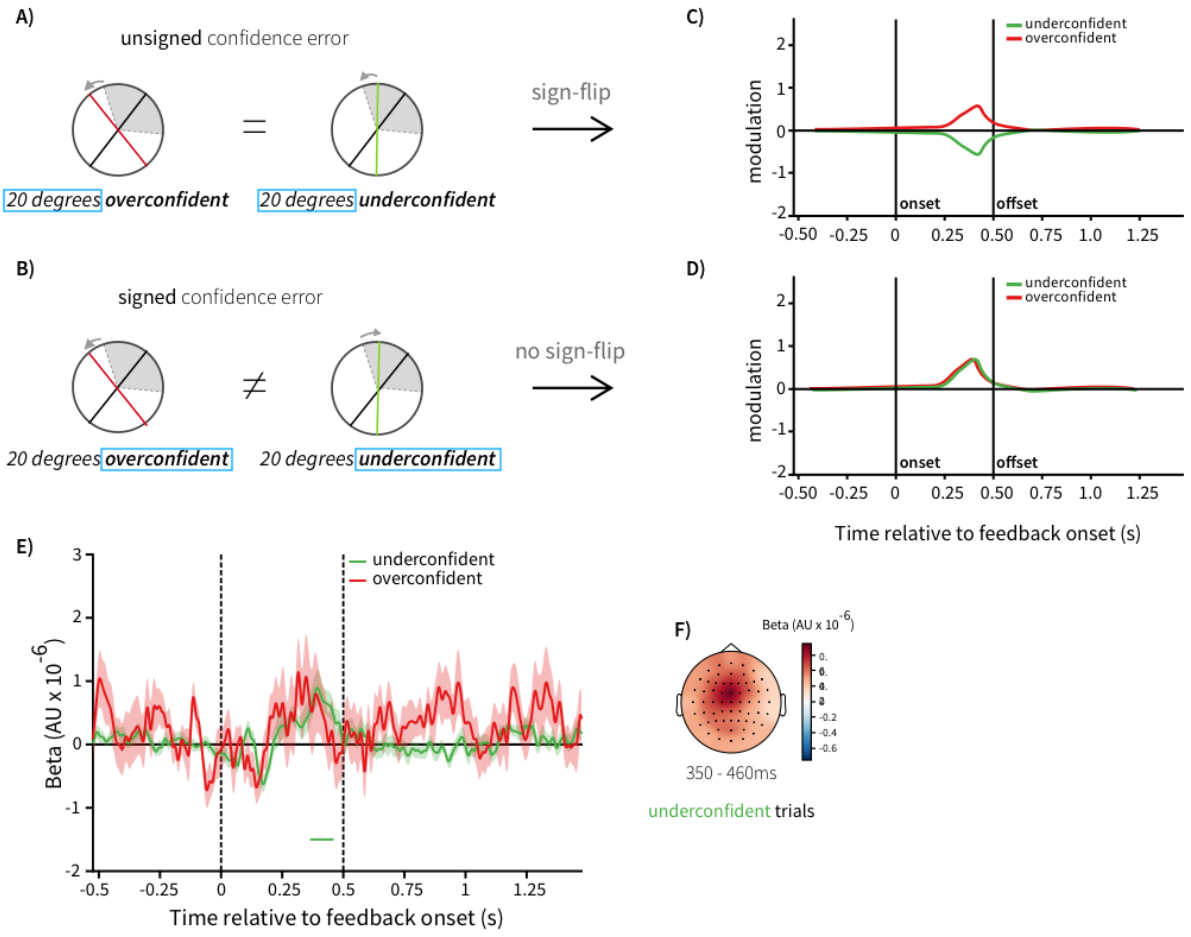

Figure 3.9 – A – D) Outline of theoretical models of confidence error representation at feedback. A) unsigned confidence error, reflecting only the magnitude of confidence error, where e.g. 20 degrees over- and under-confidence would be represented similarly. B) hypothetical visualisation of what an inverse sign between trial-types would look like to suggest an unsigned confidence error representation. C) Description of signed confidence error, reflecting that both the magnitude and direction of confidence error could be represented. D) hypothetical visualisation of what a signed confidence error representation may look like at feedback, where no sign-flip is seen between trial types. E) Modelled relationship between the feedback-evoked response and signed confidence error across trials, for overconfident (red) and underconfident (green) trials. Coloured bars above the x-axis highlight significant time periods that survived permutation testing against zero. Solid lines reflect the across-participant mean beta coefficient for the regressor of interest at electrode FCz, shaded regions represent the standard error of the mean. F) Visualisation of scalp topography for the significant cluster of the confidence error regressor that survived permutation testing in underconfident trials.

3.7 Discussion

This study set out to examine how judgements about ability and uncertainty in internal representations are processed using feedback. Across two experiments, I found evidence of signals at feedback relating to error and uncertainty regarding internal representations. I show transient modulations of the feedback-evoked response associated with both the quality (working memory error) of, and uncertainty in working memories. Importantly, I show that the categorical valence of feedback (reflecting correct vs incorrect performance) fundamentally changed the information represented at feedback.

These results build and extend upon previous research investigating feedback processing in a number of ways. Primarily, I extend upon previous findings of the categorical difference between good and poor performance by demonstrating their presence when receiving feedback on working-memory guided actions. I find evidence of a feedback-related negativity when receiving feedback on working-memory performance, suggesting processes of performance evaluation related to internal representations.

Further, I focus upon the electrophysiological signatures of feedback processing and highlight their sensitivity to features of working memory. A major empirical contribution of these studies is showing sensitivity of the feedback response to not only objective performance (error), but both external manipulations of uncertainty (experiment 1) and internal fluctuations of uncertainty (experiment 2). This suggests that feedback processing may not be solely linked to the evaluation of objective performance, but may

also be further shaped by the different types of information presented therein, and uncertainty in the quality of memories.

Meta-cognitive research frequently uses confidence judgements to infer participants' ability to consciously judge perceptual experiences or reward-guided actions. Rarely do such judgements concern internal mnemonic representations. Further, such judgements are often categorical in nature, rather than precise or on the same scale as the primary response. In experiment 2, I had participants make retrospective confidence judgements on the same scale as their primary response. Extending on findings from experiment 1 showing representations of uncertainty linked to task demands, I find a relationship between the feedback ERP and subjective uncertainty measured by retrospective confidence judgements. Importantly this representation of subjective uncertainty occurred later in time than modulations related to objective performance. This suggests that feedback processing may not be one unitary process, and instead may involve multiple steps whereby different types of information may be processed in turn. Specifically, in this case, I suggest an initial processing of objective performance and the categorical information conveyed by feedback, followed by evaluation of the metacognitive judgements made about internal representations.

Beyond representations of error and uncertainty, we find neural signals relating to confidence prediction errors in experiment 2 (see figure 3.9E). Using an exploratory model-based analysis to control for the effect of the categorical valence of feedback, I demonstrate sensitivity of the feedback response to confidence error – the difference between expected

performance (from the reported confidence judgement) and objective performance. This highlights the possibility of a neural signal at feedback that could facilitate learning about subjective uncertainty. Specifically, I show a signed confidence prediction error signal that, in this task, incorporates both the direction (over- and under-confidence) and magnitude of the difference between confidence and error. Crucially I can only explore this as participants make their confidence judgements in the same circular space as their primary response. This is important in showing a signal that could be used to guide adaptive confidence judgements, and also shows the utility of using graded feedback stimuli that give precise information regarding performance.

Previous studies looking at performance evaluation using feedback have demonstrated categorical differences in feedback processing to correct and incorrect performance. More recently, it has been suggested that these differences are driven by active processing of feedback regarding good or correct performance. Indeed, the present findings support this, as I see modulations of the feedback ERP by error only when feedback indicated good performance (Figures 3.4C and 3.8A) but not poor performance (Figure 3.4D and 3.8A). This highlights that the importance of feedback valence in shaping what may be represented, and further supports findings that the FRN may be driven by more refined, active processing of feedback indicating good performance. However, one important consideration to the feedback presented in the first experiment is that the colour of the feedback stimulus for good performance varied as a function of error – the green coloration of the feedback stimulus was darker for larger errors that lay within the bounds of 'good'

performance. This could naturally lead to covariation between the feedback response and working memory error. However, I find evidence of differences in the relationship between the feedback ERP and error as a function of uncertainty, suggesting that the variability associated with objective performance was not solely due to this change in colour.

The threshold that defines what is 'good' or 'poor' performance is also an important consideration. In binary decisions this is simple – there often is a correct or incorrect choice. As more complex behaviours are explored, and responses are continuous, giving binary feedback becomes a decision that experimenters take. Previous studies have shown that feedback processing is sensitive to various task characteristics (Cohen et al., 2007; Holroyd et al., 2008; Pfabigan et al., 2011; Yeung & Sanfey, 2004) and depends on the information emphasised at feedback (Nieuwenhuis et al., 2004). In experiment 2, feedback valence is determined by the relationship between error and confidence, so it would make sense for confidence to be represented at feedback, given its emphasis on deciding what is 'correct'. Further, this also means that trials with poorer performance need not be categorised as such based on the objective inaccuracy of working memory, but instead based on error in introspective mechanisms. If this were the case, one may expect not to see parametric modulation of the feedback ERP by error in incorrect as error may be acceptable while confidence is inappropriate. One way of determining whether the external definition of 'correctness' on a trial is driving feedback effects would be not to convey this information at all. It is possible that feedback processing is naturally influenced by task goals, and what is required to successfully perform a given task – where confidence is important for correct

performance of a task, it may be precisely represented at feedback in order to succeed. Further, individuals may be able to set their own definitions of correctness to optimise performance in tasks with less structured feedback. This could also highlight differences across individuals in the levels of performance error that individuals deem to be tolerable. Future work needs to test whether results such as ours may be driven by the emphasis on appropriate confidence, not low overall error, in determining successful performance.

Overall, these findings raise an important theoretical question regarding how to present feedback about performance when performance is continuous. While categorical feedback in perceptual or choice tasks (like flanker tasks, or binary decision tasks) is intuitive due to the binary nature of performance, in tasks requiring continuous reports this fails to convey feedback regarding the rich nature of responses. The types of information presented at feedback, and their relative emphasis, may shape the way in which individuals adjust both their responses and their beliefs about the quality of their internal representations.

While out of the scope of this study, which focusses on electrophysiological signals of feedback processing in working memory, an important question remains: Can we make decisions based on our uncertainty estimates? It has been shown that both sensory uncertainty and cognitive uncertainty are integrated into perceptual decisions and are associated with variability in behaviour (Denison et al., 2018; van Bergen et al., 2015). In this study I show representations of error and uncertainty in internal representations at feedback, and a signal reflecting a confidence prediction error. However, it remains to be

seen if this uncertainty judgement is functional – do individuals use confidence in their memories to shape decisions that they make based on them, and how might this confidence (a cognitive uncertainty) relate to internal fluctuations in sensory information? That is, are confidence judgements about internal representations made based on the quality of incoming perceptual stimuli that is subsequently encoded, or based on introspections into post-perceptual processes (e.g. the maintenance of memories or their retrieval). Future studies could manipulate these processes experimentally (e.g. varying the quality of encoded items, or altering maintenance delays) to determine their relative contributions towards the insight people have into their memories and how feedback regarding these memories is processed.

4 - Reward-guided prioritisation within working memory

4.1 Abstract

Reward is fundamental to adaptive behaviour, yet relatively little is known about how reward guides attention to the contents of working memory. It has been shown that learned reward associations can influence the encoding of sensory information into working memory, biasing encoding of objects with larger associated values. However, it remains unknown whether and how we can retrospectively assign value to working-memory contents after they have already been encoded into working memory. Across two experiments, I use retrospective attentional cues (retrocues) to assign value to objects held in working memory, and test the behavioural consequences of these assigned values. I found that reward information, conveyed through retrocues, improved performance. Even when reward was irrelevant for task performance, I found benefits for reporting the more valuable item from a working-memory array, suggesting item-specific enhancements Further, reward-related benefits are seen when reward is associated with trials rather than specific memoranda, suggesting general motivational effects of reward on working-memory performance. These results extend our understanding of how attention interacts with working-memory maintenance, highlighting the important role of reward in the prioritisation of working-memory contents.

4.2 Introduction

A key function of working memory is the ability to retain information over short periods of time to guide behaviour. While working memory is fundamentally capacity-limited (e.g. Cowan, 2001), it is also highly flexible (Nobre & Stokes, 2020). Evidence has suggested that by guiding attention to the internal representations held in working memory, capacity limits can be partially overcome (Sligte et al., 2008). Orienting attention to specific items in working memory, marking them as more relevant for the task at hand, has highlighted interesting behavioural consequences – it improves reaction times (e.g. Griffin & Nobre, 2003; Hajonides et al., 2020; Souza, Rerko, & Oberauer, 2014) and reduces errors (e.g. Griffin & Nobre, 2003; Hajonides et al., 2020; Niklaus et al., 2017; Rerko et al., 2014) for the functionally prioritised items, and can even reduce the interference between items concurrently maintained within this short-term store (e.g. Myers et al., 2018).

Studies investigating the deployment of attention to internal representations have typically used explicit cues to mark specific memoranda as task-relevant. In these "retrocue" tasks, cues are presented during maintenance periods to guide attention to internally held representations after objects are removed from view (Griffin & Nobre, 2003; Landman et al., 2003; Souza et al., 2016). Most retrocue tasks use cues that indicate which memorised object is relevant for subsequent performance by cueing the item itself (e.g. Kuo et al., 2014) or its location (Griffin & Nobre, 2003; Landman et al., 2003; Pertzov et al., 2013). Retrocues can also focus attention on the timing of items (Myers et al., 2018; van Ede et al., 2017; Zokaei et al., 2019) and on feature dimensions and feature values across items

(Hajonides et al., 2020; Niklaus et al., 2017). Overall, this highlights the flexibility of attentional orienting in visual working memory, being guided by many aspects of the world around us. However, most of these studies use cues based on relevance or probabilities of items needed on a trial. These often neglect an important dimension from consideration – that objects in the world around us can have reward value.

Reward is fundamental to guiding adaptive behaviour – choices, in both humans and monkeys, are shaped by the expected value of the available options (e.g. Behrens et al., 2007, 2008; Padoa-Schioppa & Assad, 2006; Platt & Glimcher, 1999), and reward can guide attentional processes (Anderson et al., 2011; Chelazzi et al., 2013; Tankelevitch et al., 2020). Learned associations between stimuli and reward values influence behaviour, even when those reward associations no longer apply (Anderson, 2013; Chelazzi et al., 2013; Tankelevitch et al., 2020). Further, reward shapes neural activity, resulting in increased sensitivity to targets in potentially rewarded regions of space (Kennerley & Wallis, 2009; Serences, 2008; Tankelevitch et al., 2020). Thus, it is clear that reward can shape the deployment of attention to, and our interaction with, the external world around us.

Given how reward can capture selective attention, and the strong interactions between attention and working memory, it is likely that reward can affect working-memory performance. It could do so in two core ways: general effects – influencing performance related to all objects currently held in working memory due to the prospect of reward – and item-specific effects depending on the task.

The literature on whether and how reward value influences working memory performance is mixed. Various studies have shown the impact of learned reward associations on working memory (Infanti et al., 2015; Klink et al., 2017; Wallis et al., 2015). In these tasks, participants typically first learn the reward value associated with a stimulus feature in a reward-learning task. They then subsequently take part in a visual working memory task, where the behavioural consequences of these learned associations are explored. These studies have shown that when reward cues are presented during encoding into working memory, performance is enhanced for rewarded items (Wallis et al., 2015), with increased performance on trials where the encoded array had a higher net value. These reward effects have been shown even when learned value associations are no longer predictive of reward (Infanti et al., 2015; Klink et al., 2017). Such reward-guided attention has been shown to affect working-memory encoding.

However, differing effects have been reported for value incentive effects on working memory in continuous-report tasks. Some have found benefits of reward on working-memory performance (e.g. Klink et al., 2017; Cho et al., 2022; Grogan et al., 2022; Brissenden et al., 2022; Manga et al., 2020), while others have argued against working-memory benefits by motivation (van den Berg et al., 2023; Mystakidou & van den Berg, 2020). Across two experiments, van den Berg et al. (2023) argue that reward has no effect on working-memory performance. In their task, participants were required to report the exact orientation from one of a set of previously presented oriented Gabor stimuli. Set size was manipulated to examine whether reward incentive motivation may depend on working-memory load, and

participants were randomly assigned to groups with differing total amounts of performance-related bonus. They found no effect of reward value (based on the between-subject performance bonus groupings) on the circular variance of working-memory performance – a measure of response error variability. No difference in performance was seen between people who could receive different amounts of reward at the end of the task. In a second experiment where participants were repeatedly reminded about the performance-related bonus throughout the task (to ensure it was on their mind) and received no trial-wise feedback on performance (which could result in performances closer to ceilings, reducing the effect of motivation), they still found no difference in performance for different reward levels. Indeed, Bayesian analyses argued for evidence of the null hypothesis that reward incentives have no effect on working-memory performance. Finally in a third experiment manipulating value on a trial-by-trial basis using a score multiplier that changed the value of performance on each trial, they still found no effect of reward value on working-memory performance. Finding no effect of reward at set sizes larger than the typical working-memory capacity also suggests that reward cannot be used to overcome capacity limits for the sake of increased financial gain.

Others, however, have found reward incentive benefits on working-memory performance (Cho et al., 2022;, Klink et al., 2017; Grogan et al., 2022; Brissenden et al., 2022; Manga et al., 2020). Cho et al., 2022 found that both the ability to gain reward, and to lose less reward (in gain vs loss conditions) in a spatial working-memory task where participants had to report the exact location of a previously encoded item were *both* associated with

performance benefits compared to neutral (no-reward) conditions. Brissenden et al., 2022 used a cue that conveyed both reward magnitude (the value of a trial) and identity information (through a percent predictive cue) to examine how reward value interacts with the encoding of information. Participants were better when reporting a cued item (even when validity of the cue was reduced from 80% to 66% and 50% across follow-up experiments), suggesting attentional orienting at encoding due to the presence of the pre-cue. This cueing effect interacted with reward: high-reward items were associated with lower recall error than low-reward items. When they changed the timing of this cue such that it appeared *after* encoding (as a post-cue), they found no effect of reward.

Klink et al. 2017 used a precision-recall working-memory task where participants reported the orientation of rewarded items (Klink et al., 2017). In this task, placeholder rings for the working-memory stimuli could change colour during encoding, maintenance or retrieval phases. Each placeholder ring was a different colour, assigning value to each item that was unpredictive of whether the memoranda would be probed. Here they found significant effects of reward on working-memory error when presented at encoding, but not when value information was presented during maintenance or retrieval (although this could be related to low statistical power due to the small participant samples used). Together, these suggest that reward-incentive effects are restricted to the input-gating stage, influencing the encoding of sensory information into working-memory, and cannot alter the prioritisation of an item that has already been encoded into working memory.

However, others *have* found effects when reward-related information is presented during working-memory maintenance. Grogan et al., 2022 find improvements in working-memory performance due to motivation. Across four experiments, they explore the stages at which reward motivation may influence working-memory performance, and its links to attention. Across experiments they report lower response error when participants were playing for large (50p) vs small (1p) rewards, but this did not interact with the onset time of the cue (pre-encoding or post-encoding), suggesting that reward benefits are not solely explained by improved encoding into working memory. When incidental retrocues were used during maintenance to test whether reward had differential effects on objects inside or outside the Focus of Attention, they found that reward benefit didn't interact with congruency of the attentional cue. This suggests that the motivational benefits of reward on working-memory performance are not dependent upon the items' attentional priority state. Overall, they argue that motivational benefits of reward on working-memory performance can be explained by changing decision thresholds – more conservative thresholds causing individuals to trade of speed (get slower) to increase the ability to select and retrieve the item from memory.

Further, modelling of precision responses after encoding items with different values has suggested that improvements to working-memory recall are related to changes in the probability of target selection. Reductions in guessing for high-value items compared to low-value items (Wallis et al., 2015; Grogan et al., 2022; Manga et al., 2020), or reductions in the propensity to swap and report non-target items (Grogan et al., 2022; Manga et al., 2020).

These patterns of results suggest that reward can either operate on an input-gating mechanism, altering the degree to which sensory information is transferred into working memory, or change the way in which we decide which item to report from memory, and subsequently retrieve it. One often suggested possibility is that reward-related information cannot alter the representation of items that have already been encoded into working memory.

In short, previous studies have demonstrated that reward can gate input into working memory. We aimed to test whether reward can equally alter functional working-memory states during maintenance. If so, it was also of interest to understand whether reward can have general effects on working-memory performance, item-specific effects, or both.

Across two experiments, we tested whether reward can guide attention to objects already encoded into working memory. We adapted the retrocue task design to explore spatial attention effects on working memory, presenting cues during the maintenance period that informed participants about the value of objects already encoded into working memory. In both experiments, participants viewed two coloured, oriented bars that they had to encode into memory. After a delay, they were required to report the orientation of one of these items from memory. During this delay, participants were presented with cues that indicated the value of either an item, or a trial, to explore how reward can guide prioritisation to working memories.

Experiment 1 manipulated the value of individual items maintained in working memory. Cues indicated the value of each item, which together always summed to 100. The values could be balanced (50/50 or unbalanced in favour of one item 100/0, 80/20, 70/30, or 60/40). Participants would earn the maximum number of points if they reported the item orientation perfectly and would earn a proportion of the maximum relative to the proportion of accuracy. In half of the blocks, participants were free to choose which item to report from memory – as such, they could voluntarily prioritise information held in working memory based on its value. In the other half of all blocks, participants were probed to report a specific item – thus value-based prioritisation would not necessarily facilitate task performance.

Experiment 2 used reward and identity-based retrocues manipulated in a factorial design. Reward retrocues indicated the presence or absence reward available in a trial, whereas identity cues indicated which item was relevant or provided no information. Retrocues contained shape and colour. The shape signalled whether participants would earn points on that trial, scaling with performance. The colour of the cue could prioritise one of the items maintained in working memory. The factorial crossing of cue dimensions allowed us to explore general effects of reward, as well as item-specific effects when combined with identity cues. It was also possible to test for strong interactions, such as whether reward-based effects required identity information in order to confer benefits.

To pre-empt our findings, we found that knowledge of reward value during working-memory maintenance significantly impacted performance. Extending prior findings

focused on the impact of learned reward associations on attentional capture and input-gating during encoding, we highlight the continued role of reward information in determining general and item-specific information. Reward thus plays an important role in the highly dynamic prioritisation of working-memory contents.

Experiment 1 – Behavioural consequences of reward and choice on error and confidence in working memory

4.3 Methods

This online study was reviewed and approved by the Central University Ethics Committee of the University of Oxford. All participants provided informed consent prior to their participation.

Participants

Participants were recruited from Prolific Academic (https://www.prolific.co), a platform for online participant recruitment for web-based research. Participants were pre-screened to include those who reported being aged 18-40, fluent in English, with normal or corrected-to-normal vision, and no ongoing mental health conditions. Participants were also screened for their previous participation history, requiring a minimum approval rate of 90% from their previous studies, having participated in at least 10 studies. Participants were compensated £7.50/hr for their time. An additional performance-based reward could be earned across this task, based on their performance across trials. Fifty participants were recruited (age range 19-40, mean age 29.3 years, 30 female). Participants received an average performance bonus of £5.97 (SD = £0.79).

Stimuli, procedure, and task

Participants performed a web browser-based visual working-memory task requiring them to report the exact orientation of an oriented bar from working memory. There were two main manipulations in the current task: Firstly, the potential value of the encoded items was manipulated across trials, leading to an asymmetry in the number of points that could be earned when reporting either item on any given trial. Secondly, across blocks, we manipulated whether participants were able to choose which item to report from memory (*free choice* blocks) or were instructed to report a specific item from memory (*forced choice* blocks, see Figure 4.1A for schematic visualisation). The experimental script was generated using the PsychoPy Builder version 2021.2.4 (Peirce et al., 2019) and hosted online using Pavlovia (https://www.pavlovia.org). Participants completed 8 blocks of 32 trials each, for a total of 288 trials. Four blocks were forced-choice blocks, and four blocks were free-choice, with a total of 144 trials per choice condition.

Participants performed the task using a full-screen web browser, and all visual units reported are "height" units. This generates all stimulus locations and sizes relative to the height of a window, allowing it to scale to fit different browser and screen dimensions. This was used to accommodate differences in hardware used across participants to ensure similarity in what is presented (see https://www.psychopy.org/general/units.html#height-units).

At the start of each trial, two randomly oriented coloured bars appeared 0.3 height units laterally either side of a central fixation cross (RGB:[255, 255, 255], 0.05 height and

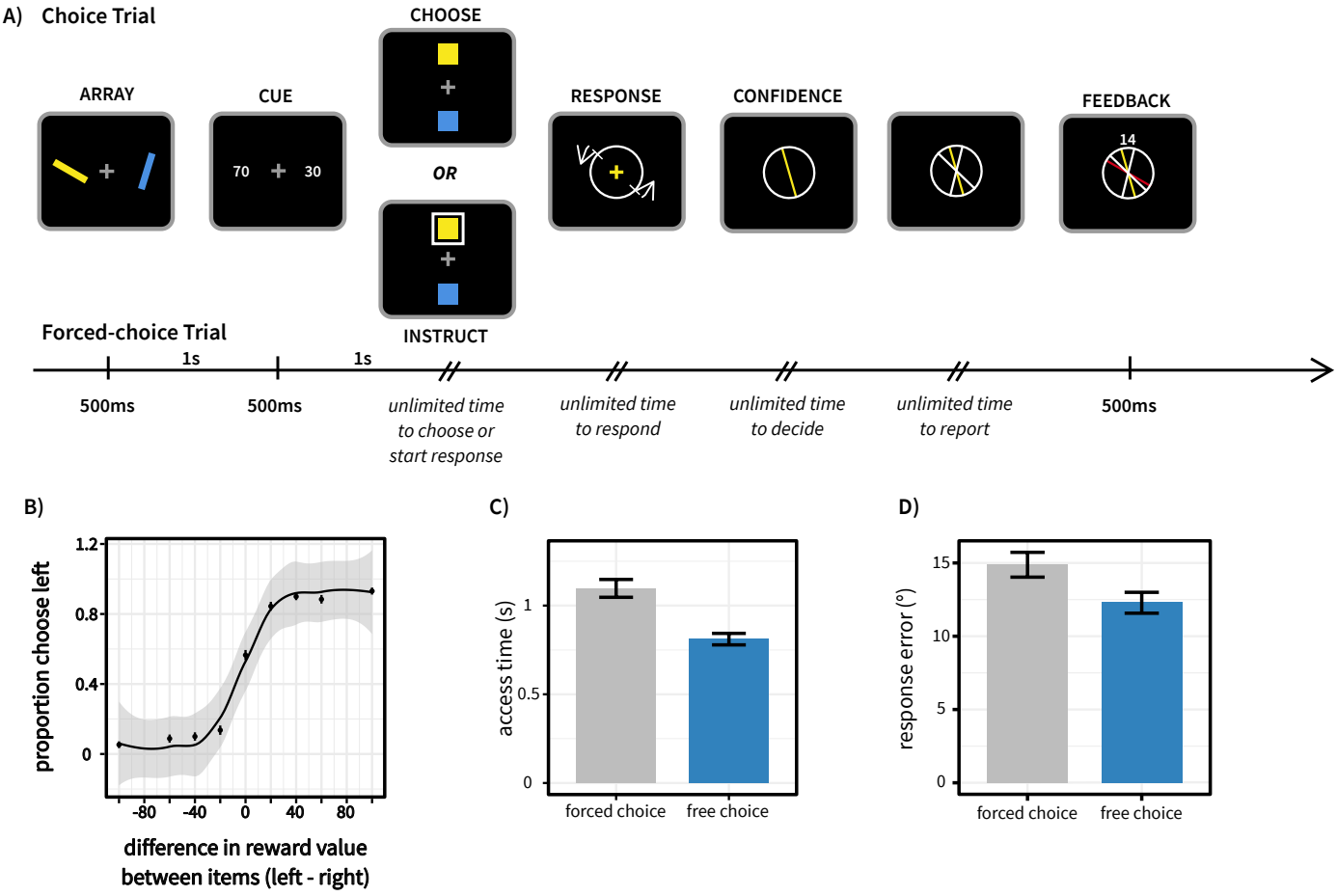

Figure 4.1 - A) Schematic showing the outline and timings for the task used in Experiment 1. Participants engaged in a precision-working memory recall task where they had to report the orientation of a bar encoded from memory. In different blocks, participants were either cued to report a specific bar (forced-choice), or were able to freely choose which orientation to report (free-choice). Participants reported their confidence in the response they gave, and received feedback on their performance. B) Choice curve showing proportion of trials where participants chose to report the left item, as a function of reward value difference between items. More positive value differences indicate larger relative value of the left item. Shaded region represents 95% confidence interval around the mean. C) Access time (how long it took to retrieve an item from memory) as a function of whether participants were instructed to report a specific item (forced-choice) or were able to choose which item to report (free-choice). D) Response error as a function of choice and whether the reported item was the most or least valuable item on that trial. All error bars represent the standard error of the mean

width) for 500 ms. Each bar had a height and width of 0.3 and 0.05 height units respectively. Bars could either be blue (RGB: [0, 0, 255]) or yellow (RGB: [255, 255, 0]). The colour of the bars was counterbalanced across trials. Orientations were drawn independently from uniform distributions between 0 and 179 degrees.

After a 500-ms delay, participants were shown the maximum value of the encoded items for 500 ms. Reward values were presented 0.1 height units laterally either side of fixation (with a letter height of 0.05 height units) for 500 ms. The values of the two bars were created with the constraint that they summed up to a maximum of 100 points – for example, if the left bar was worth 70 points, the right bar would be worth 30. The reward values of a bar could be: 100, 80, 70, 60, 50, 40, 30, 20 or 0 points. Maximum reward value was balanced across trials such that each reward value was equally likely to be associated with either item during each block of the task.

After a one-second delay, two coloured squares (height and width of 0.05 height units) were presented 0.1 height units above and below the fixation cross, matching the colour of the originally presented bars. The colours of the top and bottom bars were randomly decided across trials. In *free choice* blocks, participants had as long as they wished to decide which orientation they wished to report, before clicking on the square that matched the colour of the item they wished to report. In *forced choice* blocks, a white square outline (height and width 0.07 height units, line width 3) was presented centred on the item they were required to report. Participants had as long as they wished to decide on the orientation of the probed item before clicking on the square to proceed to the response phase. In both block types, participants had as long as they wished either to decide on the orientation of the item they chose to report (*free choice blocks*) or were instructed to report (*forced choice blocks*).

Upon starting the response phase, a grey probe circle (RGB: [128, 128, 128]) was presented at the centre of the screen (diameter 0.3 height units) with two white lines on opposite sides of the circle. Participants used the mouse to rotate the dial around the circle to replicate the precise orientation of the probed item. A left mouse click was required to confirm the response. Participants had unlimited time to reproduce the orientation and confirm their response.

After clicking to confirm their response, the confidence phase started. The orientation participants reported was shown as a thin, coloured bar (height and width 0.3 and 0.00625 height units respectively) inside the probe circle. Participants had unlimited time to decide how confident they felt, before pressing the space bar to continue to the confidence report. Upon pressing the space bar, two white lines appeared, mirrored symmetrically around the reported orientation. Participants used the mouse to adjust the angle formed by these to express their confidence interval in a circular space. Larger wedges corresponded to lower confidence in their response, while narrower wedges indicated higher levels of confidence. Participants had an unlimited time to generate and confirm their confidence report. After reporting their confidence, participants received feedback on their performance. The original target orientation was presented on top of their responses, coloured green (RGB: [0, 255, 0]) if the target was inside the reported confidence interval, and red (RGB: [204, 0, 0]) if it was outside the reported confidence interval. Participants were also given feedback on how many points they had earned. Above the response dial, a number was presented in text, showing the number of points earned

on the current trial. Points were earned according to a non-linear cost function to encourage accurate performance. Points were awarded for errors up to a threshold of 45 degrees. Beyond 45 degrees of error, no points were awarded on a trial. Reward was unrelated to the reported confidence wedge, and based solely on response error.

Behavioural analysis

I defined two main sources of error. *Response error* was calculated as the angular deviation between the target orientation and the response. *Confidence error* was calculated as the difference between this response, and the reported confidence wedge on any given trial (i.e. the difference between error and confidence). On all trials in the task, participants were presented with two small coloured squares which they were required to click (to select an item to report) shortly after seeing the reward values of the items in a trial. The time between these squares appearing on screen and participants clicking on the appropriate square to initiate the response was taken as the *access time* (i.e. the time taken to retrieve the orientation of the to-be-reported item memoranda). Trials were excluded from all analyses if their access times were over 5 seconds, or outside 2.5 standard deviations of the within-subject, within-condition mean. An average of 6.29 trials ± 0.390 (mean ± standard error) trials were discarded per participant. Where appropriate to analyse aggregate data, repeated-measures Analysis of Variance (ANOVAs) were conducted using the *'afex'* package (Singmann et al., 2020) in the R statistical programming environment (R Core Team, 2020).

4.4 Results

To explore how reward value affected the choices people made when deciding what to report from working memory, we first examined free-choice trials. We grouped trials based on the difference in value between the left and right item, then looked at the proportion of trials where participants chose to report the left item (see Figure 4.1B). This highlighted that when participants were able to choose what to report from working memory, they nearly always decided to report the item with the highest reward potential. Even when the difference in value was its lowest (20 points between items), participants chose to report the more valuable item on average for 84% of trials. This resulted in many fewer trials relating to reporting of the least rewarded item in memory.

Due to this bias in choosing the more rewarding item, we focused on analysis of forced-choice trials - where participants were instructed which item to report – as trial numbers were controlled for when the most valuable or least valuable item on a trial were reported.

To assess whether value affected the time taken to retrieve the orientation to report, we investigated access time in forced-choice trials. Planned-comparisons revealed a significant difference in access time when reporting the more valuable item on a trial compared to the least valuable item ($t(49) = 3.08$, $p = 0.003$), with faster access times when reporting the more valuable item ($M = 1.08$ s, $SE = 0.0508$ s) compared to the least valuable item ($M = 1.11$ s, $SE = 0.0504$ s). No significant difference emerged when comparing access

time on trials where the probed item was more valuable (over 50 points) compared to trials where both items had the same value of 50 points ($p = 0.145$). There was also no significant difference when comparing access time when the probed item was the least valuable on a trial (less than 50 points) compared to when both items had the same value ($p = 0.827$).

We then investigated whether item value affected response error across trials. Planned comparisons revealed a significant difference when reporting the orientation of the more valuable item in a memorised array compared to the least valuable ($t(49) = 2.58$, $p = 0.0128$), with lower response error when reporting the more valuable item in an array ($M = 14.4$ degrees, $SE = 0.811$ degrees) compared to the least valuable item ($M = 15.4$ degrees, $SE = 0.960$ degrees). We saw no significant difference when comparing trials where both items had the same value vs reporting the more valuable item ($p = 0.304$) or comparing equal reward trials to those where the probed item was the least valuable in the array ($p = 0.698$).

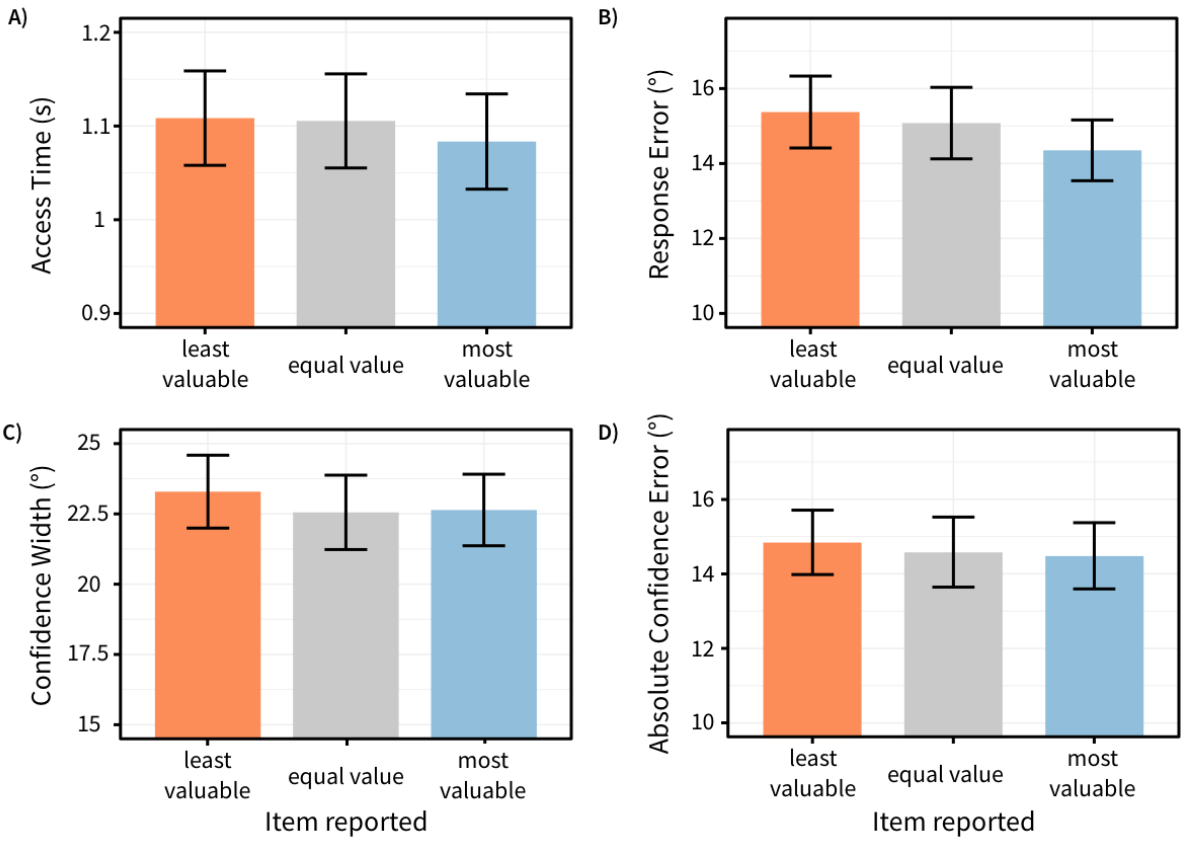

Figure 4.2 - Behavioural performance in forced-choice blocks in Experiment 1. In all panels, we visualise behavioural performance, averaged across participants, for trials where the least valuable item in the array was reported (red), where both items were worth the same value of 50 points (grey), and where the most valuable item in an array was reported (blue). A) Access Time (in seconds). B) Angular response error (in degrees). C) Reported confidence width (degrees). D) Absolute confidence error (the magnitude of over- or under-confidence across trials, in degrees). All error bars represent the standard error of the mean.

Participants were instructed to report their confidence in their orientation report on each trial of the task – as a result, we were able to test whether reward was associated with changes in subjective confidence. Here we found a significant difference in confidence by reward, with higher reported confidence (lower confidence widths) when reporting the most valuable item in an array vs the least valuable item ($t(49) = 2.90$, $p = 0.0055$. There was also a significant difference in confidence when reporting a lower value item compared to when both items were the same ($t(49) = -2.58$, $p = 0.0128$). No significant difference emerged

when comparing trials where both items were the same value vs trials where the most valuable item in an array was reported ($p = 0.772$).

As participants reported their confidence in responses on each trial, we were able to examine the quality of the reported confidence judgements by measuring their confidence error – how over- or under-confident they were on a given trial. This allowed us to look at the quality of insight into working memories, and whether this changes as a function of reward in forced-choice trials. Here we found no effect of reward on the quality of confidence judgements, with no significant difference in confidence error when reporting the most valuable item from an array, the least valuable item, or when both memoranda were worth the same 50 points (all $p > 0.28$).

4.5 Interim Discussion

Experiment 1 shows that reward cues presented during maintenance can influence working-memory performance, even when such reward is not immediately relevant to task performance. When participants were able to choose what to report from memory – making reward directly relevant to success in the task – they overwhelmingly choose to report the more valuable item. Even when reward difference is at its lowest (only 20 points separating the value of the two memoranda), participants chose to report the more valuable item over 80% of the time.

While this may seem intuitive – with the goal of the individual being to accumulate as much reward as possible across trials – this may not always be the optimal strategy for success. Value cues were presented at maintenance to prevent reward from acting upon the input-stage, meaning the error associated with objects could not be consistently in favour of the more valuable item on a trial. Given that working-memory error varies across trials, this leads to the situation where depending on the precision associated with an encoded item, it can be optimal to choose the less valuable item when one is more certain about its precision.

Moving beyond choice, where reward is relevant for task performance, we additionally show behavioural differences when reporting items from working memory according to their value differences. Performance is, on average, better when reporting the most valuable item in an array, with shorter access times (Figure 4.2A), and lower response error (Figure 4.2B). This suggests that where items have different value, there is some degree of prioritisation of working memories based on their associated values.

Performance when both items had equal value was intermediate and did not differ significantly from performance to high- or low-valued items in trials with value disparity. As a result, it is unclear what may be causing the behavioural effects that we observed – be it a facilitation of the more valuable item, or a detriment to the less valuable item (e.g. due to de-prioritisation). However, it is possible that these "equal reward" trials are not a true baseline. In such trials, all items are worth 50 points. On trials where the least valuable item was probed, these items could range from being worth 0 to 40 points. Alternatively, trials

where the most valuable item was probed were associated with item values of 60 points or more. As such, while there could be no prioritisation related to differences in value between items, it is difficult to estimate the performance contribution related purely to the intrinsic motivation provided by the value itself.

In this experiment, the net value of the memory array was consistent across trials, with item specific variation in value. This allowed us to look at whether reward could guide attentional prioritisation to specific items during working-memory maintenance. We performed Experiment 2 In order to explore the general effects of reward on working memories, and how they interact with item-specific information conveyed by identity cues.

Experiment 2 – Comparing prioritisation by reward and identity cues in working memory

4.6 Methods

Experimental procedures were reviewed and approved by the Central University Ethics Committee of the University of Oxford. All participants provided informed consent prior to their participation in the study.

Participants

Fifty-six participants were recruited for Experiment 2. Six were excluded from data analysis (for chance performance in at least one condition of the task). Data from 50 participants were fully analysed (21 male, 29 female, age range 18-40 years). Six participants did not report their precise age. Of those who did, the mean age was 25.11 years. All participants reported having normal or corrected-to-normal vision, and not being colour blind. Participants were compensated £7.50/hr for their time. An additional performance-based reward could be earned based on their performance across trials, as participants accumulated points on a subset of trials. Participants received an average performance bonus of £5.61 (SD = £0.77)

Stimuli, procedure and task.

Participants performed a web browser-based visual working-memory task requiring them to report the exact orientation of an oriented bar from working memory. On each trial, participants were presented with a retrocue during the working memory maintenance period. This cue could convey two key types of information: the identity of the item that was going to be probed at the end of each trial (Identity) and the availability of points to be earned based on the quality of item reproduction on a given trial (Reward). This resulted in a 2x2 design for cueing that allowed us to separate the contributions of identity and reward information in prioritising working-memory contents. The experimental script was generated using the PsychoPy Builder version 2020.1 (Peirce et al., 2019) and hosted online using Pavlovia (https://www.pavlovia.org). Participants completed 8 blocks of 32 trials, for a total of 256 trials, with 64 trials per cue condition.

Participants performed the task using a full-screen web browser, and all visual units reported are "height" units. This generates all stimulus locations and sizes relative to the height of a window, allowing it to scale to fit different browser and screen dimensions. This accommodates differences in hardware used across participants to ensure similarity in what is presented.

At the start of each trial (see Figure 4.3A for task schematic), two randomly oriented, coloured bars appeared laterally either side of a fixation cross (RGB:[150, 150, 150], 0.05 height and width) for 500 ms. The centre of each bar was 0.3 height units lateral from the centre of the screen, with a height and width of 0.3 and 0.05 height units respectively. Bars

could either be blue (RGB: [0, 0, 255]) or yellow (RGB: [255, 255, 21]). Orientations were drawn independently from uniform distributions between 0 and 179 degrees.

After a 1.5 second delay, a cue appeared at the centre of the screen. The cue was a small square (edge length of 0.05 height units) that could be rotated 45 degrees to appear as a diamond. The shape of the cue (square, diamond) indicated whether a reward could be gained on a trial. The mapping was counterbalanced across participants. On 50% of trials, the cue changed colour to match one of the previously encoded bars, indicating which item should be reported. In the remaining 50% of trials, the cue appeared in white (RGB: [255, 255, 255]), and provided no information about which item would be probed. This resulted in a 2x2 design whereby cues could indicate both reward availability and item identity (RI), reward availability but no identity (R), no reward availability and identity (I), and no reward availability and no identity (N). This allowed us to explore whether reward availability interacted with the identity information provided by a retrocue to confer performance benefits, and whether reward information on its own could result in recall benefits similar to those previously shown for other types of retrocues (i.e., identity, location, feature, timing).

After a 1.5-second delay, participants were probed about one of the orientations held in working memory. If the cue on a given trial did not indicate the identity of the item to be reported (e.g. a neutral or reward-only cue), the fixation cross changed colour to signal which item to report. If the cue contained identity information (i.e. changed colour to match an encoded item), the probe stimulus changed colour to be white. This made it

imperative for participants to pay attention to, and process, the cue stimulus in order to be sure what item they had to report. Participants had unlimited time to retrieve the orientation from memory, before pressing the space bar to continue to the dial up. Allowing participants an unlimited time during the probe phase provided an accurate measurement of how long it takes to retrieve the item from memory.

After pressing the space bar to begin the report phase, a fixation cross was presented at the centre of the screen (height and width 0.05 height units), with a white circular dial surrounding it (diameter 0.3 height units). Participants had to click the fixation cross at the centre of the screen to make the dial appear. This routine was added to try to standardise the mouse position during online experiments such that participants could not move the mouse in advance of the dial up in order to speed up their reporting. After clicking the central fixation cross, two small white lines appeared on opposite sides of the circle. These lines followed the position of the mouse in a full circle allowing participants to reproduce the exact orientation from memory. Participants had unlimited time to report the orientation of the item from memory, before clicking to confirm their response. After confirming their response, participants received feedback on their performance. The

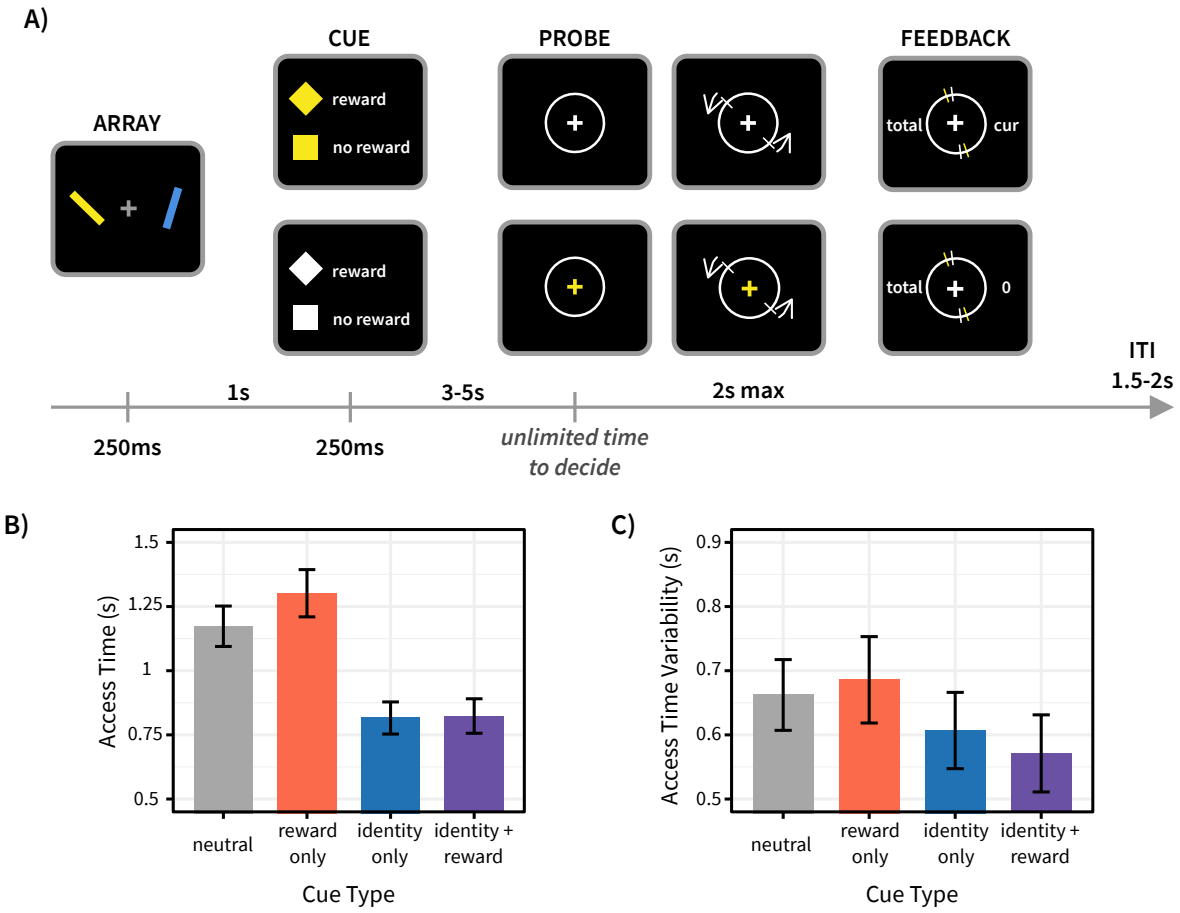

Figure 4.3 - A) Task schematic showing structure and timings of Experiment 2. B) Behavioural performance in Experiment 2. Mean decision time as a function of retrocue type. C) Decision time variability (standard deviation of decision time) as a function of retrocue type. All bar heights represent condition means. All error bars reflect the standard error of the mean.

original target orientation was presented on the probe circle as two small lines, colours matching the probed item, presented opposite each other on the circle. This gave participants continuous information regarding their performance, enabling them to determine exactly how incorrect they were on any trial. Slightly left of fixation (*x,y* coordinates (-0.1, -0.01) height units), participants were presented with text showing the total number of points they had earned on the current block so far. To the right of the fixation cross (*x,y* coordinates (0.1, -0.01) height units), participants were presented with

text showing the number of points earned on the current trial. Points were earned according to a non-linear function similar to Experiment 1. Feedback was presented for 500 ms, followed by an intertrial interval randomly drawn from between 1.5 and 2 seconds, with steps of 100 ms.

Behavioural analysis

For this experiment, I defined one source of error. *Response error* was, as in Experiment 1, calculated as the angular deviation between the target orientation and the response. During the probe phase, participants were instructed to press the 'space' bar when they had decided on the orientation they wished to report from memory, before then being able to click the central fixation cross to have the dial stimulus appear. I define *access time* as the time taken to decide on the orientation of the to-be-reported item. Trials were excluded from all analyses if their access times were over 10 seconds, or outside 2.5 standard deviations of the within-subject, within-condition mean. An average of 7.74 trials ± 0.362 (mean ± standard error) trials were discarded per participant. Trials were collapsed over item side and colour for statistical analysis. Aggregated data for conditions of the task were analysed using repeated-measures ANOVAs using the 'afex' package (Singmann et al., 2020) in the R statistical programming environment (R Core Team, 2020). All repeated-measures ANOVA include Reward (rewarded, non-rewarded) and Identity (informative, non-informative) as factors to explore the behavioural consequences of reward and spatial information both independently, and together, within attentional cues.

4.7 Results

To explore the effect of reward and item retrocues on the time taken to retrieve an orientation from memory, we conducted a 2 x 2 repeated-measures ANOVA on access time. In brief, I found that access time was only significantly improved (shortened) when cues conveyed item identity information. I found a significant main effect of Identity ($F(1,49 = 69.3$, $p < 0.0001$, pes = 0.59), a significant main effect of Reward ($F(1,49) = 4.10$, $p = 0.0483$, pes = 0.08) and a significant interaction between the two ($F(1,49) = 8.29$, $p = 0.006$, pes = 0.14). Paired-samples t-tests were performed to determine the effects driving this interaction. I found a significant difference in access time for Neutral cues and Reward-only cues ($t(49) = -2.655$, $p = 0.0107$) with access time slower for Reward-only cues (M = 1.30 seconds, SE = 0.0921 seconds) compared to Neutral cues (M = 1.17 seconds, SE = 0.0787 seconds). I also found that access time was slower for trials with Neutral cues compared to both Identity-only cue trials ($t(49) = 7.57$, $p < 0.0001$), and compared to combined Reward-Identity cue trials ($t(49) = 6.74$, $p < 0.0001$). I found no difference in access time when comparing cues that conveyed only item Identity information (M = 0.816 seconds, SE = 0.0626 seconds) to cues that conveyed both Identity and Reward information (M = 0.823 seconds, SE = 0.0672 seconds, $t(49) = -0.271$, $p = 0.788$).

I also explored whether the types of information contained within the cue had similar effects on the variability in time taken to retrieve the probed orientation from memory. To that end, I conducted a 2 x 2 repeated-measures ANOVA on access time variability (the standard deviation of access times). This revealed a significant main effect

of Identity (F(1,49) = 6.56, p = 0.0136, pes = 0.12) and no significant main effect of Reward (p = 0.835) or interaction between the two (p = 0.0718). This main effect of item Identity was driven by lower variability in access time for trials where cues conveyed Item identifying information (Identity-only cues, and Reward-Identity cues combined, M = 0.602 seconds, SE = 0.0583 seconds) compared to trials where cues contained no item identity information (Neutral cues and Reward-only cues combined, M = 0.693 seconds, SE = 0.0618 seconds).

To determine whether reward and item identity information in the retrocue had comparable effects on the accuracy of working memory recall, I implemented a 2 x 2 repeated-measures ANOVA on response error. Overall, I found that all types of information presented in the retrocue conferred behavioural benefits. I found significant main effects of Identity (F(1,49) = 14.57, p = 0.0004, pes = 0.23) and Reward (F(1,49) = 16.35, p = 0.0002, pes = 0.25), and a significant interaction between the two (F(1,49) = 7.28, p = 0.010, pes = 0.13). Post-hoc t-tests highlighted that response error was higher for Neutral-cue trials compared

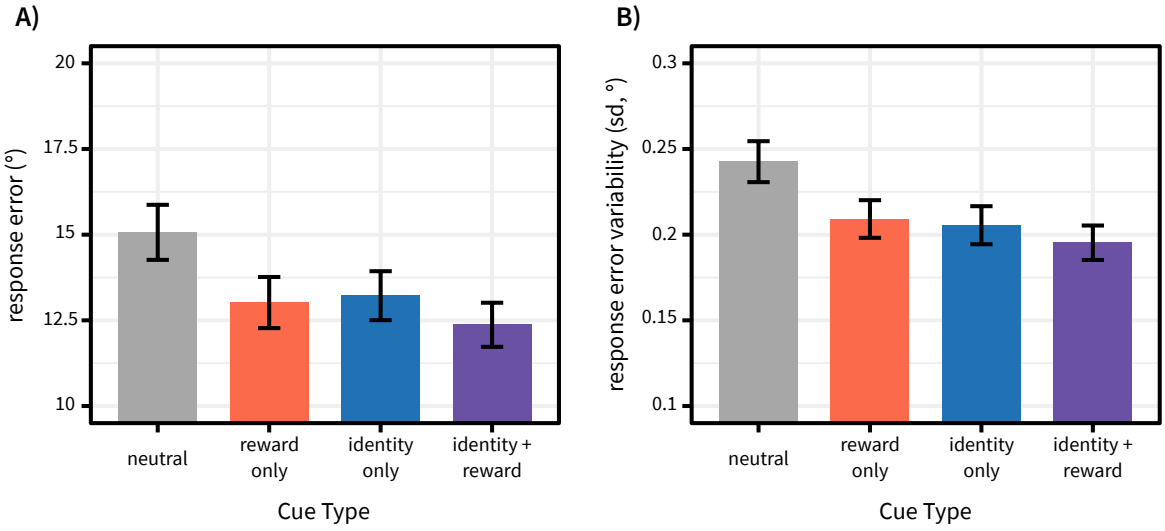

Figure 4.4 - Behavioural performance in Experiment 2. A) Response error as a function of cue type. B) Response error variability as a function of cue type. All bar heights represent the mean. All error bars reflect the standard error of the mean.

to Reward-only cue trials ($t(49) = 4.42$, $p < 0.001$), Identity-only cue trials ($t(49) = 4.33$, $p < 0.001$), and combined Reward-Identity-cue trials ($t(49) = 4.82$, $p < 0.001$). No significant difference in response error was found between trials with isolated Reward and Identity cues ($t(49)=-0.505$, $p = 0.616$), or between reward cue trials and trials with combined Reward-Identity cues ($t(49) = 1.80$, $p = 0.0787$). A significant difference was found between Identity-only cue trials, and combined Reward-Identity cue trials ($t(49) = 2.26$, $p = 0.0281$). This highlights that the combined Reward-Identity cue led to the smallest degrees of response error – reliably smaller than trials with isolated item identity cues, but not isolated reward-cues.

Previous studies have shown that attentional retrocues not only improve the average error when recalling information from memory but are also associated with reduced variability in error. To determine whether cues incorporating reward-related information are also associated with changes in response error variability, I conducted a 2x2 repeated-measures ANOVA on the standard deviation of response errors across trials. I found that all types of cue were associated with similar levels of improvement in response error variability. This found significant main effects of Identity ($F(1,49) = 13.89$, $p = 0.0005$, $pes = 0.22$), Reward ($F(1,49) = 12.34$, $p = 0.001$, $pes = 0.2$), and a significant interaction between them ($F(1,49) = 4.89$, $p = 0.0317$). Response error variability was higher in Neutral trials compared to all other cue types (all $p < 0.001$). No significant differences emerged between reward-only cue trials ($M = 0.209$ degrees, $SE = 0.0110$ degrees), identity-only cue

trials (M = 0.206 degrees, SE = 0.0111 degrees), or combined Reward-Identity cue trials (M = 0.195 degrees, SE = 0.0101 degrees, all $p > 0.05$).

4.8 Discussion

In two studies I explored whether reward can guide attentional prioritisation of the information held within working memory. Across both experiments, I presented participants with retrocues that conveyed information regarding the value of the items they had previously encoded into working memory. I found that reward modulates working-memory performance, adding reward-value to the list of dimensions of working memory that can be used to guide the prioritisation of internal representations.

I show that reward information presented during working-memory maintenance can result in item-specific benefits. Focusing on forced-choice trials in Experiment 1 - where participants were instructed to report a specific orientation from memory with an associated value – I found that participants are better when reporting the most valuable item in an array, compared to reporting the corresponding less valuable item. Not only were they better in their objective performance (reduced response error), but participants were also subjectively more confident in the responses they gave (see Figure 4.2C) – although there was no significant change in the quality of their introspections. Overall this suggests that even in the absence of choice, where reward was not relevant for successful task performance, reward can guide attention to prioritise memoranda based on their value. This is consistent with another study using an analogous design that tested the effect of reward when the value of items across the array was uneven (Grogan et al., 2022, experiment 4), where they also find reduced response error for the more valuable item in the array in the unequal reward condition. However, while they consistently find slower

reaction times to more valuable items (across many replications), I find that access time – the time taken to report retrieving the item and starting the dial phase – is faster for the more valuable item. This difference in reaction time effect could be attributable to aspects of task design and data collection. In experiment 1, participants were instructed to take as long as they needed to retrieve the item from memory before clicking on a coloured square to choose which item to report. This instruction emphasises is thus a proxy for how long it takes to retrieve it from working memory. In the experiments by Grogan and colleagues, they focus on the full time taken to click and confirm the report of the working-memory item, rather than looking at the onset time of the movement itself (while their analyses do suggest that RT slowing is not simply due to deliberation and fine-tuning towards the end of the response). Further, our effect of reward value on RT in forced-choice trials is small – a mean difference of ~ 30ms – in an online sample with varying hardware and browsers across participants. A replication of this effect in an independent, in-person sample with consistent hardware would help verify if this is true, or due to chance due to imprecision in data collection. Given that in Experiment 2 of this chapter I show reward-related slowing (in the reward-only cues vs neutral cues) on working memory performance (consistent with Grogan et al., 2022), further investigations would be essential to see if this reduced access time is true.

One key issue in both tasks examining unequal reward priorities is the issue of baseline. When reward value is imbalanced between trials, one can compare performance under different levels of reward-based prioritisation. Equal value trials themselves are not

strictly a baseline as although reward is balanced, it is still present, and thus not a truly neutral control. As a result, it is unclear whether the associated reduction in performance (increased error) of low value items vs high value items seen amongst various studies (e.g. those here, and also Grogan et al., 2022; Brissenden et al., 2022; Manga et al., 2020) are due to an absolute de-prioritisation of the less valuable item (dropping the less valuable item), rather than just a relative prioritisation of the most valuable item (enhancing the more valuable). Comparing high-value to low-value is not the same as comparing high-value to no-value. This is naturally a tricky obstacle to overcome as strategically participants may decide to disengage when items have no value if their over-arching goal is to accumulate financial reward.

Important questions arise after demonstrating that reward can be used to guide internally directed attention: What mechanisms (behavioural or neural) support reward-guided attentional selection? Do these differ from the implementations of other types of guided attention, such as those supported by identity cues, feature cues, or temporal attention? How does reward intersect with ways of guiding attention within working memory?

These findings suggest that reward and identity-based retrocues have different mechanisms by which they act upon working memories. In Experiment 2, I factorially manipulate the information content of retrocues to explore isolated effects of reward and identity information, and their combination. Importantly, I found differing effects of reward and identity information when presented in isolated cues – while both resulted in similar

improvements in response error (reducing it compared to neutral cues), only identity cues resulted in shorter times to access (or retrieve) an item from working memory. This pattern suggests a shift in the speed and accuracy priority in the case of reward-based prioritisation, with participants sacrificing speed for higher accuracy (consistent with e.g. Grogan et al., 2022). In the case of identity retrocues, no such trade-off was observed, with participants becoming both faster and more precise. This behavioural dissociation in effects suggests two different mechanisms by which reward and identity cues alter working-memory utilisation. It is likely that identity cues (100% valid with respect to the item they cue), result in item-specific shifts in the focus of attention. Along with behavioural evidence suggesting item-specific attentional prioritisation, neurophysiological investigations have also shown signals that reflect item-specific processing of attended, or prioritised, memoranda (Mok et al., 2016; Myers et al., 2015; Wallis et al., 2013)

Reward could have two possible mechanisms of action – non-specific effects linked to general motivational changes associated with the availability of reward, and item-specific effects associated with the reward bound to specific memoranda. Across these experiments, I found evidence for both of these mechanisms. In Experiment 1, I found differences in performance when reporting the most or least valuable item on a trial, with lower response error when reporting the more valuable item on a trial. This suggests some difference in prioritisation when items have unequal associated rewards – an item-specific effect. In Experiment 2 I found that cues conveying reward information alone (unlinked to specific representations in working memory) facilitate performance similarly to cues that

indicate the identity of the to-be-reported item. In the absence of identity information, these isolated reward cues did not afford the prioritisation of a specific item, nor did they permit the early retrieval in a manner appropriate for successful task performance. This suggests their benefit arises due to generalised enhancements at the prospect of reward, rather than item specific retrieval – indeed, access time was not improved (shortened) by isolated reward cues. Thus, I provide evidence that reward information presented during the maintenance of working-memories can have both general and item-specific effects, depending on the structure and appearance of the utilised (retro)cues.

One interesting consideration is the mechanisms by which reward can induce general enhancements of working-memory performance. Working memory is a multi-faceted cognitive function, typically broken down into three key parts: encoding, maintenance and retrieval. In our experiments, reward cues were presented after encoding (during maintenance) to attempt to overcome biases in encoding associated with reward value. Thus, reward could have generalised effects on both the maintenance or retrieval of memories – some studies have argued that reward cannot alter the priority state of information already encoded into memory (e.g. Klink et al., 2017; Brissenden et al., 2020), while others have shown that reward post-cues can result in comparable effects on WM performance to reward-related pre-cues (Grogan et al., 2022), further arguing that reward effects are not solely linked to variable encoding precision, but instead linked to changes in retrieval processes . However, reward could be used to help protect items from decay over time, from interference between items, or help maintain them more actively or faithfully

during maintenance periods. Retrocues that do not convey reward information have been shown to have similar effects on working memory (Myers et al., 2018; Pertzov et al., 2013), guided either by retrocues (e.g. Kuo et al., 2009; Myers et al., 2015; van Ede et al., 2019) or by temporal structures of the task (e.g. van Ede et al., 2017), has been shown to result in improved performance along with changes in neural activity associated with orienting towards the attended representation. In Experiment 1, it is hard to determine what is causing the differences in performance between low- and high-value items due to non-significant differences compared to the "baseline" condition where items had equal reward. To better understand the contributions of general reward value (e.g. the magnitude of the items themselves) and the difference in value between items, a more parametric experimental design may be more appropriate. Focusing on forced-choice trials but allowing values to vary more continuously may provide greater insights. In the design of Experiment 1, the total array value was fixed. However, by allowing the array value to vary across trials, one can have similar value differences where the objective value of both items is different (e.g. 70:60 trials vs 40:30 trials). Multiple regression approaches could then be used to model the effect of the probed item value and the difference between trials to more systematically disentangle the contributions of reward availability and the reward difference between items to look at general and item-specific reward effects.

In Experiment 2, I showed that both reward and identity information in isolation can improve response error. However, reward-only cues did not facilitate (shorten) access time – the time taken to retrieve an item from the short-term store. This suggests that the two

types of cues may act differently – identity-based cues facilitating the early retrieval of a task relevant item (thus shorter access time), whereas isolated reward cues do not afford such early access. Instead, they necessarily must have different influences after their appearance. This behavioural difference in the effect of reward- and identity-based retrocues provides further evidence that item-based retrocues facilitate performance through the early selection, or prioritisation, of task-relevant information held in working memory.

Despite the difference in behavioural consequences of reward- and identity-based retrocues, both types of information had comparable effects on response error. Given that the former must operate through generalised mechanisms, and the latter affords item-specific retrieval, one must consider how two different mechanisms could facilitate similar behaviours. Internally-directed attention to visual working memories is typically associated with increased activity across a distributed network of fronto-parietal regions (e.g. Nee & Jonides, 2009; Nobre et al., 2004; Wallis et al., 2013) and occipital sensory regions (Kuo et al., 2014; Sligte et al., 2009). However, reward, and learning about rewards, is associated with phasic dopaminergic activity in the striatum (Schultz et al., 1997; Waelti et al., 2001), and with other cortical areas like the Anterior Cingulate Cortex (Behrens et al., 2007; Rushworth & Behrens, 2008) representing the value of choices (or outcomes) in reward-guided learning and action. Thus, reward- and identity-based cues may be supported by different networks of cortical (and subcortical) regions. Comparing these two cue types with neurophysiological recordings could help better understand both the commonalities

and differences in neural implementation of reward-guided and identity-guided attentional orienting to working memories.

Overall, across two experiments I show that reward-value information presented during working-memory maintenance can guide internally-directed selective attention. I provide evidence of both general reward effects, and item-specific effects associated with differences in reward value between memoranda. Thus, I bring reward into the list of dimensions of attention that can support improved working-memory guided behaviours.

5 - Discussion

5.1 General Overview

To use memories effectively in guiding behaviour, one should know their quality and value to judge whether they should be guiding what we do. This thesis has explored topics within working memory with this goal in mind. Specifically, I set out to explore ideas in two main themes. The first theme explored uncertainty in working memories – whether selective attention to working-memory contents changes the way that we introspect upon them, and whether and how we can learn about the quality of our working memories. The second theme relates specifically to how reward interacts with attention. I explore whether reward-guided attention can act retrospectively upon already-encoded objects in working memory. I test the consequences of reward-related retrocues upon working-memory performance. Here I will discuss the main findings of these studies and their implications for what we know about working memory.

5.2 Summary of experimental findings

In Chapter 2, I report the results of a study looking at how the attentional prioritisation of working-memory contents changes the way that we introspect upon them. I show that retrocues confer benefits not only in objective performance but also in the quality of introspection into memory quality. Confidence judgements became more accurate when cues afforded the prioritisation of specific working-memory representations during

maintenance. Further, I provide novel evidence that subjective confidence in working-memory performance is flexibly adjusted across trials based on feedback concerning recent accuracy of confidence judgements. Overall, the experiments in Chapter 2 highlight that the benefits of attentional retrocues might not only be implicit but also change the introspective quality of mental contents.

In Chapter 3, I extend upon this work using EEG to look at the neural representation of uncertainty at feedback across two working-memory tasks. In the first experiment, I experimentally manipulate uncertainty by varying working-memory load while keeping constant the number of items presented in the visual array. In the second, I measure confidence through the use of a retrospective confidence judgement on performance and examine how this endogenous variability in confidence is related to the feedback-evoked response. I extend previous findings of a feedback-related negativity in perceptual tasks to the working-memory domain. I show, across both experiments, differences in the feedback-evoked response as a function of feedback valence (contrasting poor vs. good performance). Further, I use model-based analyses to show that the feedback response is parametrically sensitive to aspects of performance – namely working-memory error, uncertainty (both manipulated and measured), and confidence error (how well confidence matches actual performance).

In Chapter 4, I sought to test whether reward can guide the prioritisation of working-memory contents. Earlier studies have tested the consequence of previously learned reward associations on working-memory performance (e.g., associating spatial locations

with reward). Instead, I adapted a retrocue design to assess the impact of reward-guided attention on items already encoded into memory – finding profound effects of reward on choice and performance. I show that reward can have both general and item-specific effects on working-memory performance. Reward information on its own (not tied to a specific item) is sufficient to induce benefits in response error. I also show item-specific effects, with more valuable items being recalled more accurately from memory than less valuable items. I show dissociable effects of reward and identity information during working-memory maintenance. Identity-based retrocues conferred benefits to both access time (faster working-memory retrieval) and response error (lower response error). In contrast, isolated reward cues conferred benefits in response error – with a similar degree of improvement to the typical identity-based retrocue – but no speeding of access time. Indeed, reward in isolation was associated with a slower time to retrieve an object from memory. This highlights that reward may operate upon working-memory contents via different mechanisms compared to typical identity-based retrocues.

5.3 Uncertainty in working memories

In the introduction to this thesis, I argue it is important to understand working-memory uncertainty to be able to use memories adaptively, and accurately, to guide behaviour. Certainty in one's memories may encourage one to use memory to guide action – whereas uncertainty in memory quality may encourage us to rely more on external sources of information. When accurate performance in a task relies upon using memories to allow us

to integrate information over time, understanding the faithfulness of our memory store is important.

In my research here, I extend upon previous findings that working-memory confidence tracks performance (Geurts et al., 2022; Honig et al., 2020; Rademaker et al., 2012; Samaha et al., 2019). For the first time, I show that orienting attention to a specific memorandum confers not only objective performance benefits but also benefits to the quality of introspection. An important additional novel contribution of this research is the demonstration that recent confidence history shapes current confidence judgements. I show that confidence is adapted across trials in a way that reflects the recent quality of introspection - trials in which performance was better than expected led to a subsequent increase in confidence on the next trial.

This is important because it shows that alongside a flexible working memory, we have flexible insight into these memories. Attention and task performance wax and wane over time rather than being constant. An introspective mechanism that evaluates performance or uncertainty in memories should also fluctuate over time to be accurate. Rather than having a general belief about working-memory quality – which would manifest as stereotyped confidence judgements across all trials – we find variable confidence judgements that are correlated with working-memory performance across trials, and that appear to integrate recent performance history into their creation.

5.4 What processes are evaluated when making confidence judgements?

While the introspective judgements made on working-memory contents are clearly flexible, I believe two important questions still remain largely unanswered. Firstly, what aspects of working-memory processes are such introspective judgements based on? Outside of this thesis, other researchers have shown that working-memory confidence is sensitive to working-memory load (Rademaker et al., 2012). However, these working-memory load manipulations alter the working-memory array – e.g. presenting three versus six items at encoding. As a result, this type of load manipulation affects both working-memory encoding (more items are encoded) and maintenance (more items are maintained). This makes it difficult to disentangle the unique contribution of these distinct processes when making confidence judgements.

Alternative task designs could get around this issue and allow us to examine their relative contributions. For example, in Chapter 3 I use a working-memory load manipulation that keeps the encoding array consistent across trials – the same number of items are presented in for all levels of working-memory load. Using such a task manipulation would enable us to look more carefully at how changing maintenance processes influences confidence while keeping encoding consistent. Another alternative would be to leverage the ability to change sensory parameters of Gabor stimuli without varying load. Gabor stimuli are heavily adaptable to precisely control the amount of evidence available for a particular orientation. In this way, one could affect the fidelity of working-memory encoding (for a fixed number of items) to parametrically investigate how

changing the encoding of information into working memory changes confidence. New task designs are needed to better understand the possible unique contributions that different working-memory processes might have towards our estimation of uncertainty.

5.5 The use of confidence

The second important question is whether confidence itself is functional – do we evaluate working-memory uncertainty in a way that is useful for behaviour beyond just being able to report our confidence?

In the perceptual domain, there is evidence that uncertainty (in sensory evidence or the decision process) is integrated into perceptual decisions. Certainly, there is a broad literature showing that sensory evidence quality is represented and integrated into perceptual decisions (e.g. Kiani et al., 2014; van Bergen et al., 2015). It is important to understand whether the evaluation of working-memory uncertainty – explicitly measured through confidence judgements – is also integrated into working-memory guided decisions, and associated with changes in behaviour. For example, monkeys making perceptual decisions appear to opt out of the task (in favour of a smaller but certain reward) when sensory evidence is weak (Kiani & Shadlen, 2009). This suggests that an evaluation of sensory quality can guide their decision to engage with the task for a potentially higher reward, or opt out of a certain reward. In this instance, the evaluation of uncertainty is functional in that high uncertainty is associated with a large change in behaviour. This type of task design is easily adaptable to human working-memory research. In a task where

participants are rewarded based on the quality of working-memory recall and given the option to opt out in favour of a smaller certain reward, one could examine how uncertainty at different stages influences the decision to act based upon memory. For example, one could examine how changing encoding parameters, working-memory load, or maintenance-period parameters like duration, influences uncertainty in a way that affects the propensity to opt out.

Some studies have shown that subjective uncertainty estimates in working memory are indeed functional and can be used to guide decision making. For example, when people are able to choose what to report from a memory array, they generally report the best remembered items (e.g. Adam, Vogel & Awh, 2017; Fougnie, Suchow & Alvarez, 2012; Suchow, Fougnie, Alvarez, 2016). When participants can choose the order in which they report items from a memory array (when required to report a feature from all encoded items), response distributions get flatter (more like a uniform guessing distribution) further along in the response sequence. When required to report items in an order generated by the task (forced-order), response distribution decline did not follow the same pattern, suggesting that participants really can determine which they remember best, and report those first (Adam et al., 2017). Further, when attentional state is manipulated through retrocues to cause changes in the quality of the remembered item, uncertainty judgements ("which item do you remember best") change across time to reflect updated states of remembering (Suchow et al., 2016). Altogether, this suggests that individuals have

knowledge about the quality of their memories that can change over time based on currently held information in a way that is functional and can guide behaviour.

5.6 Neural representations of working-memory uncertainty

Another important line of research is to understand how the brain represents uncertainty. Neurobiologically derived models of working memory, such as conjunctive coding models (Bays, 2014; Schneegans & Bays, 2017), can represent uncertainty as an inherent part of the working-memory code. In such models, feature combinations are represented by the activity of populations of neurons that encode them – for example, a population of neurons would respond maximally to a blue bar oriented at 30 degrees. Feature values are then read out (e.g. during working-memory retrieval) based on a decoding of the population activity. In such models, uncertainty can naturally be represented by the width of the distribution of feature values that a population of neurons encodes. With precise encoding, population tuning curves would be narrow, reflecting high certainty in the memorised feature.

In visual perceptual decisions, it has been proposed that sensory certainty is represented by population activity across the visual cortex (van Bergen et al., 2015). Within working memory, it has also been proposed that uncertainty is integrated into rewarded working-memory decisions (Honig et al., 2020) and that this working-memory uncertainty is maintained alongside their related representations (Yoo et al., 2021). If uncertainty representations are maintained and used in working-memory guided decisions, what is the neural implementation of this uncertainty? Recently, others have attempted to understand

working-memory uncertainty by extending such fMRI methods into working memory (Li et al., 2021). They show that the decoded feature value correlates with working-memory performance and that the uncertainty of the decoder correlated with self-reported confidence in the reported feature. Together, this suggests that uncertainty may be represented inherently as part of the working-memory representation itself.

However, visual working memory is a dynamic construct – interactions between working memory and attention highlight this. Using attentional cues, we can shift the prioritisation of the objects we hold in this store. As a result of how dynamic it is, fMRI may not fully capture all aspects of activity that represent our working-memory representations. Neuroimaging methods that operate on faster timescales – namely EEG and MEG – can help understand how the temporal dynamics of working-memory processes relate to uncertainty in a way that fMRI simply could not. By extending analytical methods used in fMRI at the faster timescales possible in EEG and MEG, one could examine how different aspects of faster brain activity (and uncertainty in these dynamics) relate to the uncertainty in our memories.

In Chapter 3, I explore how uncertainty is represented in the neural response to feedback. The FRN has commonly been explored in a variety of perceptual tasks, initially as the categorical difference in feedback processing between incorrect and correct trials (Hajcak et al., 2006; Miltner et al., 1997). What is thought to drive the FRN component when contrasting these two trial types has changed. More recent studies have suggested that the FRN is not a change in the processing of feedback relating to incorrect performance, but

instead relating to correct performance (Becker et al., 2014; Holroyd et al., 2008; Nieuwenhuis et al., 2005). Indeed, in my study – investigating the feedback-related negativity in the context of feedback on working-memory performance – I find modulations of the ERP are most prominent for trials where feedback indicated good performance. The findings of error modulation of feedback on correct trials are consistent with active processing changes when receiving feedback that is reflective of good performance.

However, I find that although the feedback-evoked response covaries with working-memory error only in correct trials, the feedback ERP is sensitive to subjective uncertainty (when explicitly measured across trials) in both types of feedback valence. In that experiment (Experiment 2, Chapter 3), however, the correctness of a trial is determined based on how subjective uncertainty related to objective performance, rather than on objective performance alone. Indeed, the feedback-evoked response is sensitive to many different types of task parameters (Cohen et al., 2007; Holroyd et al., 2008; Pfabigan et al., 2011; Yeung & Sanfey, 2004) and to the information that is emphasised at feedback (Nieuwenhuis et al., 2004). In our task, feedback valence is determined by the relationship between error and confidence, thus while error may not be as controllable (under the control of natural variability in working-memory processes), confidence can be adapted across trials to achieve the best possible feedback. This emphasis on confidence could be driving our findings. It is important to understand, through future research, how feedback on working-memory performance is altered by the types of information presented at feedback (e.g. error, confidence), the way it is presented (categorical or continuous

information), the types of task (e.g. change-detection vs free-recall), and how it may facilitate changes in performance across trials.

5.7 Reward and working memory

In Chapter 4, I explore reward as a task feature that can guide the prioritisation of working-memory contents. Much of the previous research has focused on how learned associations bias the encoding of information into working memory (e.g. Infanti et al., 2015; Klink et al., 2017; Lin & Fougnie, 2022; Wallis et al., 2015). These studies emphasise that reward can bias the transformation of information from external input to internal representation. In my research, I instead focus on how reward can bias the prioritisation of information already held in working memory.

Researchers had previously argued that reward could not influence working-memory maintenance, only the encoding of information (Klink et al., 2017). In their task, participants learned to associate colours with rewards, and these colours could be presented as placeholders for working-memory items during different working-memory processes. They find performance benefits when reward is presented at encoding, but a non-significant effect when reward information was presented during maintenance. In contrast, across two experiments, we find novel evidence of reward-guided attention to working-memory contents that confers performance benefits. Indeed, benefits to response error are comparable to identity-based retrocues, and benefits still arise for more valuable

items even when reward was made task-irrelevant. These findings provide new evidence that reward can bias the prioritisation of information held in working memory.

This leads to the natural next step – how does reward act upon working memories? There are myriad ideas of how retrocues may confer working-memory performance benefits, such as enhanced protection from decay over time (Pertzov et al., 2013), more effective orthogonalization from memoranda that are less relevant for task performance (Myers et al., 2018), or an elevation to a more functional state that better guides behaviour (Griffin & Nobre, 2003; Rerko et al., 2014). All of these are item-specific effects that could be equally applicable to reward-guided internal attention. However, we also find that reward boosts working-memory performance when not tied to specific items, and others have found effects not linked to item value, but the net value of working-memory arrays (Wallis et al., 2015). This suggests that reward can not only work in tandem with item-based attentional effects but can also have more generalised motivational benefits on performance. By presenting reward cues after items have already been encoded, we are able to disentangle this from increased attention at encoding. Instead, it is likely that generalised reward effects are derived from more effective maintenance or retrieval processes.

Future research could thus attempt to disentangle the behavioural and neural bases of these generalised and item-specific effects. Behaviourally, I show that reward can result in prioritisation of memoranda even when irrelevant to the task at hand – when participants were unable to choose which item from an array to report, they were better at reporting the

more valuable of the presented items. In this first experiment of Chapter 4, I constrained the net value of the working-memory array to be fixed across trials in order to control for generalised effects of working memory – thus performance benefits could be attributed to item-specific prioritisation based on reward differences. This was necessary to examine item-specific effects. Future studies could investigate this forced-choice block type alone, but allow the absolute value of items to vary parametrically across trials while still maintaining similar value differences between items. Regression-based analyses could then be used – with sufficient trial numbers – to evaluate statistically the relative contributions of generalised effects (due to the absolute value of the item or the array) and item-specific effects (based on relative differences between items).

There also remain open questions regarding the neural implementation of reward-guided attention to working memory. Behavioural studies have shown that the value of different stimuli in the working-memory array is associated with changes in performance. I show that the value of memoranda, presented in the form of a retrocue, is associated with changes in performance. Attentional selection of working-memory contents has been associated with changes in neural activity recorded at the scalp in both ERPs (e.g. the N2pc, Kuo et al., 2009) and modulations of ongoing oscillations in the alpha frequency (8 – 14 Hz, e.g., Mok et al., 2016; Myers et al., 2015; Poch et al., 2014; van Ede et al., 2017). Little is known about how reward affects such indices of attentional selection. In visual search, the value of target stimuli has been shown to influence the magnitude of the N2pc (Kiss et al., 2009) - with an earlier onset and larger magnitude N2pc for high-reward versus low-reward. When

memorising a series of words associated with different values, frontal theta activity (4 – 8 Hz) has been associated with later performance when recalling high-reward words (Gruber et al., 2013). How reward influences cortical dynamics during working-memory encoding and maintenance is unclear. However, there are many markers of attentional selection that could be investigated at these stages to better understand its implementation. Future research could combine pre- and retro-cues containing reward information to investigate their consequences on neural activity, and the information content therein.

5.8 Conclusion

Working memory is an important cognitive construct that underpins many different types of goal-directed behaviour. In order to know when we should use working memories to guide action, we should have insight into the quality of our memories and understand their relative value to decide which memories are most appropriate to utilise. In this thesis, I engage with these themes – I show that subjective confidence in working memories tracks working-memory performance, and is sensitive to attentional changes. I show that it is dynamic across trials and tracks the quality of insight. I show neural modulations of feedback that might support learning about performance in such tasks. Moreover, I show that we can dynamically assign value retrospectively to items held in memory to optimise the receipt of reward based on working-memory performance. Together, these findings attempt to integrate ideas from decision-making into the working-memory context. It is

important to understand how, beyond simply retrieving them from their internal store, we can effectively use working memories to guide the actions that we take.

References

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79–97. https://doi.org/10.1016/j.cogpsych.2017.07.001

Amiez, C., Joseph, J. P., & Procyk, E. (2005). Anterior cingulate error-related activity is modulated by predicted reward. *European Journal of Neuroscience*, *21*(12), 3447–3452. https://doi.org/10.1111/j.1460-9568.2005.04170.x

Amiez, C., Joseph, J. P., & Procyk, E. (2006). Reward encoding in the monkey anterior cingulate cortex. *Cerebral Cortex*, *16*(7), 1040–1055. https://doi.org/10.1093/cercor/bhj046

Anderson, B. A. (2013). A value-driven mechanism of attentional selection. *Journal of Vision*, *13*(3), 1–16. https://doi.org/10.1167/13.3.7

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10367–10371. https://doi.org/10.1073/pnas.1104047108

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

Baker, T. E., & Holroyd, C. B. (2011). Dissociated roles of the anterior cingulate cortex in reward and conflict processing as revealed by the feedback error-related negativity and N200. *Biological Psychology*, *87*(1), 25–34. https://doi.org/10.1016/j.biopsycho.2011.01.010

Baruni, J. K., Lau, B., & Salzman, C. D. (2015). Reward expectation differentially modulates attentional behavior and activity in visual area V4. *Nature Neuroscience*, *18*(11), 1656–1663. https://doi.org/10.1038/nn.4141

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

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

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

Becker, M. P. I., Nitsch, A. M., Miltner, W. H. R., & Straube, T. (2014). A single-trial estimation of the feedback-related negativity and its relation to BOLD responses in a time-

estimation task. *Journal of Neuroscience*, *34*(8), 3005–3012. https://doi.org/10.1523/JNEUROSCI.3684-13.2014

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 245–249. https://doi.org/10.1038/nature07538

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Bichot, N. P., Rossi, A. F., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, *308*(5721), 529–534. https://doi.org/10.1126/science.1109676

Boldt, A., Schiffer, A. M., Waszak, F., & Yeung, N. (2019). Confidence Predictions Affect Performance Confidence and Neural Preparation in Perceptual Decision Making. *Scientific Reports*, *9*(1), 1–17. https://doi.org/10.1038/s41598-019-40681-9

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, *35*(8), 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Bourgeois, A., Chelazzi, L., & Vuilleumier, P. (2016). How motivation and reward learning modulate selective attention. In *Progress in Brain Research* (1st ed., Vol. 229). Elsevier B.V. https://doi.org/10.1016/bs.pbr.2016.06.004

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 1–34. https://doi.org/10.1167/11.5.4

Braun, E. K., Wimmer, G. E., & Shohamy, D. (2018). Retroactive and graded prioritization of memory by reward. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-07280-0

Brissenden, J. A., Adkins, T. J., Ting Hsu, Y., & Lee, T. G. (2021). Reward influences the allocation but not the availability of resources in visual working memory. *BioRxiv*, 2021.06.08.447414. https://doi.org/10.1101/2021.06.08.447414

Camara, E., Manohar, S., & Husain, M. (2013). Past rewards capture spatial attention and action choices. *Experimental Brain Research*, *230*(3), 291–300. https://doi.org/10.1007/s00221-013-3654-6

Chelazzi, L., Perlato, A., Santandrea, E., & Della Libera, C. (2013). Rewards teach visual selective attention. *Vision Research*, *85*, 58–72. https://doi.org/10.1016/j.visres.2012.12.005

Cho, Y. T., Moujaes, F., Schleifer, C. H., Starc, M., Ji, J. L., Santamauro, N., Adkinson, B., Kolobaric, A., Flynn, M., Krystal, J. H., Murray, J. D., Repovs, G., & Anticevic, A. (2022). Reward and loss incentives improve spatial working memory by shaping trial-by-trial posterior frontoparietal signals. *NeuroImage*, *254*. https://doi.org/10.1016/j.neuroimage.2022.119139

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *NeuroImage*, *35*(2), 968–978. https://doi.org/10.1016/j.neuroimage.2006.11.056

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(4), 87–185. https://doi.org/10.1017/S0140525X01003922

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204. https://doi.org/10.1016/j.conb.2006.03.006

Della Libera, C., & Chelazzi, L. (2006). Visual selective attention and the effects of monetary rewards. *Psychological Science*, *17*(3), 222–227. https://doi.org/10.1111/j.1467-9280.2006.01689.x

Della Libera, C., & Chelazzi, L. (2009). Learning to attend and to ignore is a matter of gains and losses. *Psychological Science*, *20*(6), 778–784. https://doi.org/10.1111/j.1467-9280.2009.02360.x

Dell'Acqua, R., Sessa, P., Toffanin, P., Luria, R., & Jolicœur, P. (2010). Orienting attention to objects in visual short-term memory. *Neuropsychologia*, *48*(2), 419–428. https://doi.org/10.1016/j.neuropsychologia.2009.09.033

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(43), 11090–11095. https://doi.org/10.1073/pnas.1717720115

Ditterich, J., Mazurek, M. E., & Shadlen, M. N. (2003). Microstimulation of visual cortex affects the speed of perceptual decisions. *Nature Neuroscience*, *6*(8), 891–898. https://doi.org/10.1038/nn1094

Doallo, S., Patai, E. Z., & Nobre, A. C. (2013). Reward associations magnify memory-based biases on perception. *Journal of Cognitive Neuroscience*, *25*(2), 245–257. https://doi.org/10.1162/jocn_a_00314

Donner, T. H., Siegel, M., Fries, P., & Engel, A. K. (2009). Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology : CB*, *19*(18), 1581–1585. https://doi.org/10.1016/j.cub.2009.07.066

Eimer, M. (1996). The N2pc component as an indicator of attentional selectivity. *Electroencephalography and Clinical Neurophysiology*, *99*(3), 225–234. https://doi.org/10.1016/s0921-884x(96)95711-2

Engelmann, J. B., Damaraju, E., Padmala, S., & Pessoa, L. (2009). Combined effects of attention and motivation on visual task performance: Transient and sustained motivational effects. *Frontiers in Human Neuroscience*, *3*(MAR), 1–17. https://doi.org/10.3389/neuro.09.004.2009

Engelmann, J. B., & Pessoa, L. (2007). Motivation Sharpens Exogenous Spatial Attention. *Emotion*, *7*(3), 668–674. https://doi.org/10.1037/1528-3542.7.3.668

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*. https://doi.org/10.1038/ncomms2237

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*(5563), 2279–2282. https://doi.org/10.1126/science.1066893

Geurts, L. S., Cooke, J. R. H., van Bergen, R. S., & Jehee, J. F. M. (2022). Subjective confidence reflects representation of Bayesian probability in cortex. *Nature Human Behaviour*, *6*(2), 294–305. https://doi.org/10.1038/s41562-021-01247-w

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic Updating of Working Memory Resources for Visual Objects. *Journal of Neuroscience*, *31*(23), 8502–8511. https://doi.org/10.1523/JNEUROSCI.0208-11.2011

Gould, I. C., Nobre, A. C., Wyart, V., & Rushworth, M. F. S. (2012). Effects of decision variables and intraparietal stimulation on sensorimotor oscillatory activity in the human brain. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(40), 13805–13818. https://doi.org/10.1523/JNEUROSCI.2200-12.2012

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python . In *Frontiers in Neuroscience* (Vol. 7, p. 267). https://www.frontiersin.org/article/10.3389/fnins.2013.00267

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

Grogan, J. P., Randhawa, G., Kim, M., & Manohar, S. G. (2022). Motivation improves working memory by two processes: Prioritisation and retrieval thresholds. *Cognitive Psychology*, *135*. https://doi.org/10.1016/j.cogpsych.2022.101472

Gruber, M. J., Watrous, A. J., Ekstrom, A. D., Ranganath, C., & Otten, L. J. (2013). Expected reward modulates encoding-related theta activity before an event. *NeuroImage*, *64*(1), 68–74. https://doi.org/10.1016/j.neuroimage.2012.07.064

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, *71*(2), 148–154. https://doi.org/10.1016/j.biopsycho.2005.04.001

Hajonides, J. E., van Ede, F., Stokes, M. G., & Nobre, A. C. (2020). Comparing the prioritization of items and feature-dimensions in visual working memory. *Journal of Vision*, *20*(8), 1–12. https://doi.org/10.1167/JOV.20.8.25

Hanks, T. D., Ditterich, J., & Shadlen, M. N. (2006). Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience*, *9*(5), 682–689. https://doi.org/10.1038/nn1683

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage*, *84*, 159–168. https://doi.org/10.1016/j.neuroimage.2013.08.028

Hernández, a, Zainos, a, & Romo, R. (2000). Neuronal correlates of sensory discrimination in the somatosensory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(11), 6191–6196. https://doi.org/10.1073/pnas.120018597

Ho, T. C., Brown, S., Abuyo, N. A., Ku, E. H. J., & Serences, J. T. (2012). Perceptual consequences of feature-based attentional enhancement and suppression. *Journal of Vision*, *12*(8), 1–17. https://doi.org/10.1167/12.8.15

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*(5), 688–697. https://doi.org/10.1111/j.1469-8986.2008.00668.x

Honig, M., Ma, W. J., & Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8391–8397. https://doi.org/10.1073/pnas.1918143117

Huk, A. C., & Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *25*(45), 10420–10436. https://doi.org/10.1523/JNEUROSCI.4684-04.2005

Hunt, L. T., & Hayden, B. Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. *Nature Reviews Neuroscience*, *18*(3), 172–182. https://doi.org/10.1038/nrn.2017.7

Husain, M., Shapiro, K., Martin, J., & Kennard, C. (1997). Abnormal temporal dynamics of visual attention in spatial neglect patients. *Nature*, *385*(6612), 154–156. https://doi.org/10.1038/385154a0

Infanti, E., Hickey, C., & Turatto, M. (2015). Reward associations impact both iconic and visual working memory. *Vision Research*, *107*, 22–29. https://doi.org/10.1016/j.visres.2014.11.008

Ito, S., Stuphorn, V., Brown, J. W., & Schall, J. D. (2003). Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science (New York, N.Y.)*, *302*(5642), 120–122. https://doi.org/10.1126/science.1087847

Kennerley, S. W., & Wallis, J. D. (2009). Reward-dependent modulation of working memory in lateral prefrontal cortex. *Journal of Neuroscience*, *29*(10), 3259–3270. https://doi.org/10.1523/JNEUROSCI.5353-08.2009

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Kiss, M., Driver, J., & Eimer, M. (2009). Reward priority of visual target singletons modulates event-related potential signatures of attentional selection. *Psychological Science*, *20*(2), 245–251. https://doi.org/10.1111/j.1467-9280.2009.02281.x

Klink, P. C., Jeurissen, D., Theeuwes, J., Denys, D., & Roelfsema, P. R. (2017). Working memory accuracy for multiple targets is driven by reward expectation and stimulus contrast with different time-courses. *Scientific Reports*, *7*(1), 1–13. https://doi.org/10.1038/s41598-017-08608-4

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science*, *335*(6077), 95–98. https://doi.org/10.1126/science.1216930

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., & Ono, T. (2001). Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature*, *412*(6846), 546–549. https://doi.org/10.1038/35087595

Kuo, B.-C., Rao, A., Lepsien, J., & Nobre, A. C. (2009). Searching for targets within the spatial layout of visual short-term memory. *The Journal of Neuroscience : The Official Journal*

of the Society for Neuroscience, *29*(25), 8032–8038. https://doi.org/10.1523/JNEUROSCI.0952-09.2009

Kuo, B.-C., Stokes, M. G., Murray, A. M., & Nobre, A. C. (2014). Attention Biases Visual Activity in Visual Short-term Memory. *Journal of Cognitive Neuroscience*, *26*(7), 1377–1389. https://doi.org/10.1162/jocn_a_00577

Kuo, B.-C., Stokes, M. G., & Nobre, A. C. (2012). Attention modulates maintenance of representations in visual short-term memory. *Journal of Cognitive Neuroscience*, *24*(1), 51–60. https://doi.org/10.1162/jocn_a_00087

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Landman, R., Spekreijse, H., & Lamme, V. a F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, *43*(2), 149–164.

Lauer, T., Cornelissen, T. H. W., Draschkow, D., Willenbockel, V., & Võ, M. L. H. (2018). The role of scene summary statistics in object recognition. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-32991-1

Li, H.-H., Sprague, T. C., Yoo, A. H., Ma, W. J., & Curtis, C. E. (2021). Joint representation of working memory and uncertainty in human cortex. *Neuron*, 1–14. https://doi.org/10.1016/j.neuron.2021.08.022

Lin, Y., & Fougnie, D. (2022). No evidence that the retro-cue benefit requires reallocation of memory resources. *Cognition*, *229*(June 2021), 105230. https://doi.org/10.1016/j.cognition.2022.105230

Ling, S., Liu, T., & Carrasco, M. (2009). How spatial and feature-based attention affect the gain and tuning of population responses. *Vision Research*, *49*(10), 1194–1204. https://doi.org/10.1016/j.visres.2008.05.025

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Luft, C. D. B. (2014). Learning from feedback: The neural mechanisms of feedback processing facilitating better performance. *Behavioural Brain Research*, *261*, 356–368. https://doi.org/10.1016/j.bbr.2013.12.043

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Manga, A., Vakli, P., & Vidnyánszky, Z. (2020). The influence of anticipated monetary incentives on visual working memory performance in healthy younger and older adults. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-65723-5

Matsumoto, M., Matsumoto, K., Abe, H., & Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience*, *10*(5), 647–656. https://doi.org/10.1038/nn1890

Matthey, L., Bays, P. M., & Dayan, P. (2015). A Probabilistic Palimpsest Model of Visual Short-term Memory. *PLOS Computational Biology*, *11*(1), e1004003. https://doi.org/10.1371/journal.pcbi.1004003

McAdams, C. J., & Maunsell, J. H. R. (2000). Attention to both space and feature modulates neuronal responses in macaque area V4. *Journal of Neurophysiology*, *83*(3), 1751–1755. https://doi.org/10.1152/jn.2000.83.3.1751

Mesulam, M.-M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology*, *28*(5), 597–613. https://doi.org/10.1002/ana.410280502

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-Related Brain Potentials Following Incorrect Feedback in a Time-Estimation Task: Evidence for a "Generic" Neural System for Error Detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798. https://doi.org/10.1162/jocn.1997.9.6.788

Mok, R. M., Myers, N. E., Wallis, G., & Nobre, A. C. (2016). Behavioral and Neural Markers of Flexible Attention over Working Memory in Aging. *Cerebral Cortex*, *26*(4), 1831–1842. https://doi.org/10.1093/cercor/bhw011

Myers, N. E., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2018). Benefits of flexible prioritization in working memory can arise without costs. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(3), 398–411. https://doi.org/10.1037/xhp0000449

Myers, N. E., Walther, L., Wallis, G., Stokes, M. G., & Nobre, A. C. (2015). Temporal Dynamics of Attention during Encoding versus Maintenance of Working Memory: Complementary Views from Event-related Potentials and Alpha-band Oscillations. *Journal of Cognitive Neuroscience*, *27*, 492–508. https://doi.org/10.1162/jocn_a_00727

Mystakidou, M., & van den Berg, R. (2020). More motivated but equally good: No effect of gamification on visual working memory performance. *BioRxiv*, 1–24. https://doi.org/10.1101/2020.01.12.903203

Nee, D. E., & Jonides, J. (2009). Common and distinct neural correlates of perceptual and memorial selection. *NeuroImage*, *45*(3), 963–975. https://doi.org/10.1016/j.neuroimage.2009.01.005

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, *341*(6237), 52–54. https://doi.org/10.1038/341052a0

Nieuwenhuis, S., Slagter, H. A., Von Alting Geusau, N. J., Heslenfeld, D. J., & Holroyd, C. B. (2005). Knowing good from bad: Differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, *21*(11), 3161–3168. https://doi.org/10.1111/j.1460-9568.2005.04152.x

Nieuwenhuis, S., Yeung, N., Holroyd, C. B., Schurger, A., & Cohen, J. D. (2004). Sensitivity of electrophysiological activity from medial frontal cortex to utilitarian and performance feedback. *Cerebral Cortex*, *14*(7), 741–747. https://doi.org/10.1093/cercor/bhh034

Niklaus, M., Nobre, A. C., & van Ede, F. (2017). Feature-based attentional weighting and spreading in visual working memory. *Scientific Reports*, *7*(June 2016), 42384. https://doi.org/10.1038/srep42384

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, *12*(7), 265–272. https://doi.org/10.1016/j.tics.2008.03.006

Nobre, A. C., Coull, J. T., Maquet, P., Frith, C. D., Vandenberghe, R., & Mesulam, M. M. (2004). Orienting Attention to Locations in Perceptual Versus Mental Representations. *Journal of Cognitive Neuroscience*, *16*(3), 363–373. https://doi.org/10.1162/089892904322926700

Nobre, A. C., Sebestyen, G. N., Gitelman, D. R., Mesulam, M. M., Frackowiak, R. S. J., & Frith, C. D. (1997). Functional localization of the system for visuospatial attention using positron emission tomography. *Brain*, *120*(3), 515–533. https://doi.org/10.1093/brain/120.3.515

Nobre, A. C., & Stokes, M. G. (2020). Memory and attention: the back and forth. In *The Cognitive Neurosciences* (pp. 291–300). MIT Press.

Oberauer, K. (2009). Design for a Working Memory. In *Psychology of Learning and Motivation - Advances in Research and Theory* (1st ed., Vol. 51, Issue 09, pp. 45–100). Elsevier Inc. https://doi.org/10.1016/S0079-7421(09)51002-X

Padoa-Schioppa, C., & Assad, J. a. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226. https://doi.org/10.1038/nature04676

Park, Y. E., Sy, J. L., Hong, S. W., & Tong, F. (2017). Reprioritization of Features of Multidimensional Objects Stored in Visual Working Memory. *Psychological Science*, *28*(12), 1773–1785. https://doi.org/10.1177/0956797617719949

Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning and Memory*, *24*(1), 65–69. https://doi.org/10.1101/lm.042978.116

Peck, C. J., Jangraw, D. C., Suzuki, M., Efem, R., & Gottlieb, J. (2009). Reward modulates attention independently of action value in posterior parietal cortex. *Journal of Neuroscience*, *29*(36), 11182–11191. https://doi.org/10.1523/JNEUROSCI.1929-09.2009

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1224–1231. https://doi.org/10.1037/a0030947

Pfabigan, D. M., Alexopoulos, J., Bauer, H., & Sailer, U. (2011). Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials. *Psychophysiology*, *48*(5), 656–664. https://doi.org/10.1111/j.1469-8986.2010.01136.x

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*(6741), 233–238. https://doi.org/10.1038/22268

Poch, C., Campo, P., & Barnes, G. R. (2014). Modulation of alpha and gamma oscillations related to retrospectively orienting attention within working memory. *European Journal of Neuroscience*, *40*(February), 2399–2405. https://doi.org/10.1111/ejn.12589

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*, 3–25. https://doi.org/10.1080/00335558008248231

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*.

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13), 1–13. https://doi.org/10.1167/12.13.21

Raymond, J. E., & O'Brien, J. L. (2009). Selective Visual Attention and Motivation. *Psychological Science*, *20*(8), 981–988. https://doi.org/10.1111/j.1467-9280.2009.02391.x

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849–860. https://doi.org/10.1037/0096-1523.18.3.849

Rerko, L., Souza, A. S., & Oberauer, K. (2014). Retro-cue benefits in working memory without sustained focal attention. *Memory & Cognition*, *42*, 712–728. https://doi.org/10.3758/s13421-013-0392-8

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *22*(21), 9475–9489. https://doi.org/10.1016/S0377-2217(02)00363-6

Romo, R., Hernández, A., Zainos, A., Brody, C. D., & Lemus, L. (2000). Sensing without Touching: Psychophysical Performance Based on Cortical Microstimulation. *Neuron*, *26*(1), 273–278. https://doi.org/10.1016/S0896-6273(00)81156-3

Romo, R., Hernández, A., Zainos, A., & Salinas, E. (1998). Somatosensory discrimination based on cortical microstimulation. *Nature*, *392*(6674), 387–390. https://doi.org/10.1038/32891

Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397. https://doi.org/10.1038/nn2066

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*(7), 631–632. https://doi.org/10.1038/nn876

Samaha, J., Switzky, M., & Postle, B. R. (2019). Confidence boosts serial dependence in orientation estimation. *Journal of Vision*, *19*(4), 1–13. https://doi.org/10.1167/19.4.25

Schneegans, S., & Bays, P. M. (2017). Neural Architecture for Feature Binding in Visual Working Memory. *The Journal of Neuroscience*, *37*(14), 3913–3925. https://doi.org/10.1523/JNEUROSCI.3493-16.2017

Scholl, J., Kolling, N., Nelissen, N., Wittmann, M. K., Harmer, C. J., & Rushworth, M. F. S. (2015). The good, the bad, and the irrelevant: Neural mechanisms of learning real and hypothetical rewards and effort. *Journal of Neuroscience*, *35*(32), 11233–11251. https://doi.org/10.1523/JNEUROSCI.0396-15.2015

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Serences, J. T. (2008). Value-Based Modulations in Human Visual Cortex. *Neuron*, *60*(6), 1169–1181. https://doi.org/10.1016/j.neuron.2008.10.051

Siegel, M., Donner, T. H., Oostenveld, R., Fries, P., & Engel, A. K. (2007). High-frequency activity in human visual cortex is modulated by visual motion strength. *Cerebral Cortex*, *17*(March), 732–741. https://doi.org/10.1093/cercor/bhk025

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *afex: Analysis of Factorial Experiments*.

Sligte, I. G., Scholte, H. S., & Lamme, V. a F. (2008). Are there multiple visual short-term memory stores? *PLoS ONE*, *3*(2), 2–10. https://doi.org/10.1371/journal.pone.0001699

Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2009). V4 activity predicts the strength of visual short-term memory representations. *Journal of Neuroscience*, *29*(23), 7432–7438. https://doi.org/10.1523/JNEUROSCI.0784-09.2009

Small, D. M., Gitelman, D., Simmons, K., Bloise, S. M., Parrish, T., & Mesulam, M. M. (2005). Monetary incentives enhance processing in brain regions mediating top-down control of attention. *Cerebral Cortex*, *15*(12), 1855–1865. https://doi.org/10.1093/cercor/bhi063

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, and Psychophysics*, *78*(7), 1839–1860. https://doi.org/10.3758/s13414-016-1108-5

Souza, A. S., Rerko, L., Lin, H.-Y., & Oberauer, K. (2014). Focused attention improves working memory: implications for flexible-resource and discrete-capacity models. *Attention, Perception & Psychophysics*, 2080–2102. https://doi.org/10.3758/s13414-014-0687-2

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*. *42*(6), 890–910.

Stanisor, L., Van Der Togt, C., Pennartz, C. M. A., & Roelfsema, P. R. (2013). A unified selection signal for attention and reward in primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(22), 9136–9141. https://doi.org/10.1073/pnas.1300117110

Suchow, J. W., Fougnie, D., & Alvarez, G. A. (2017). Looking inward and back: Real-time monitoring of visual working memories. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(4), 660–668. https://doi.org/10.1037/xlm0000320

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nature Reviews. Neuroscience*, *6*(5), 363–375. https://doi.org/10.1038/nrn1666

Tankelevitch, L., Spaak, E., Rushworth, M. F. S., & Stokes, M. G. (2020). Previously Reward-Associated Stimuli Capture Spatial Attention in the Absence of Changes in the Corresponding Sensory Representations as Measured with MEG. *Journal of Neuroscience*, *40*(26), 5033–5050. https://doi.org/10.1523/JNEUROSCI.1172-19.2020

Tosoni, A., Galati, G., Romani, G. L., & Corbetta, M. (2008). Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions. *Nature Neuroscience*, *11*(12), 1446–1453. https://doi.org/10.1038/nn.2221

Treue, S., & Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(June), 575–579. https://doi.org/10.1038/21176

van Bergen, R. S., Ji Ma, W., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730. https://doi.org/10.1038/nn.4150

van den Berg, R., Zou, Q., Li, Y., & Ma, W. J. (2023). No effect of monetary reward in a visual working memory task. *PLoS ONE*, *18*(1 January). https://doi.org/10.1371/journal.pone.0280257

van den Berg, R., Zou, Q., & Ma, W. J. (2019). Does monetary reward increase visual working memory performance? *BioRxiv*, 767343. https://doi.org/10.1101/767343

van Ede, F., Chekroud, S. R., & Nobre, A. C. (2019). Human gaze tracks attentional focusing in memorized visual space. *Nature Human Behaviour*, *3*(5), 462–470. https://doi.org/10.1038/s41562-019-0549-y

van Ede, F., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2018). Decoding the influence of anticipatory states on visual perception in the presence of temporal distractors. *Nature Communications*, *9*(1), 1449. https://doi.org/10.1038/s41467-018-03960-z

van Ede, F., Niklaus, M., & Nobre, A. C. (2017). Temporal expectations guide dynamic prioritization in visual working memory through attenuated α oscillations. *The Journal of Neuroscience*, *37*(2), 437–445. https://doi.org/10.1523/JNEUROSCI.2272-16.2017

Waelti, P., Dickinson, a, & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*(6842), 43–48. https://doi.org/10.1038/35083500

Wallis, G., Stokes, M., Cousijn, H., Woolrich, M., & Nobre, A. C. (2013). *Frontoparietal and cingulo-opercular networks play dissociable roles in control of working memory*. 2019–2034. https://doi.org/10.1162/jocn

Wallis, G., Stokes, M. G., Arnold, C., & Nobre, A. C. (2015). Reward boosts working memory encoding over a brief temporal window. *Visual Cognition*, *23*(1–2), 291–312. https://doi.org/10.1080/13506285.2015.1013168

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*, 1120–1135. https://doi.org/10.1167/4.12.11

Williams, S. M., & Goldman-Rakic, P. S. (1998). Widespread origin of the primate mesofrontal dopamine system. *Cerebral Cortex*, *8*(4), 321–345. https://doi.org/10.1093/cercor/8.4.321

Wittmann, B. C., Dolan, R. J., & Düzel, E. (2011). Behavioral specifications of reward-associated long-term memory enhancement in humans. *Learning and Memory*, *18*(5), 296–300. https://doi.org/10.1101/lm.1996811

Ye, C., Hu, Z., Ristaniemi, T., Gendron, M., & Liu, Q. (2016). Retro-dimension-cue benefit in visual working memory. *Scientific Reports*, *6*(October), 1–13. https://doi.org/10.1038/srep35573

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. *Psychological Review*, *111*(4), 931–959. https://doi.org/10.1037/0033-295X.111.4.931

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP correlates of feedback and reward processing in the presence and absence of response choice. *Cerebral Cortex*, *15*(5), 535–544. https://doi.org/10.1093/cercor/bhh153

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, *24*(28), 6258–6264. https://doi.org/10.1523/JNEUROSCI.4537-03.2004

Yoo, A. H., Acerbi, L., & Ma, W. J. (2021). Uncertainty is maintained and used in working memory. *Journal of Vision*, *21*(8), 1–19. https://doi.org/10.1167/jov.21.8.13

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

Zokaei, N., Board, A. G., Manohar, S. G., & Nobre, A. C. (2019). Modulation of the pupillary response by the content of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(45), 22802–22810. https://doi.org/10.1073/pnas.1909959116