Investigating the Neural Mechanisms of Memory Retrieval

Devyn Eron Smith

Broadway, Virginia

B.S., Virginia Polytechnic Institute and State University, 2017

M.A., University of Virginia, 2021

A Dissertation presented to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of

Doctor of Philosophy

Department of Psychology

University of Virginia

April, 2024

Committee Members

Dr. Nicole M. Long (Chair)

Dr. Per B. Sederberg

Dr. Teague R. Henry

Dr. Zhihao Zhang

0.1 Acknowledgments

First, I would like to thank my advisor, Dr. Nicole Long for her mentorship and guidance the last five years. Thank you for pushing me and teaching me how to be a better scientist. I have learned so much from you and I am eternally grateful for all of your support. To my remaining committee members, Drs. Per Sederberg, Teague Henry, and Zhihao Zhang, thank you all for your insightful feedback and recommendations throughout this process.

To Isabelle Moore and Yuju Hong, thank you for being such wonderful friends and lab mates. My appreciation for you two always proofreading my emails/grants/dissertation, helping me with code, and doing escape rooms with me because I like them, cannot be overstated. I am grateful for and to you both.

To my mom, you always encouraged me to follow my dreams and I wouldn't be here without your support, so thank you. Thank you Shaye Waggy and Seth Reed for always being down to attend a sporting event or go to Philadelphia for a weekend to get a cheesesteak. You two have helped keep me sane. To my second family, the Wakemans, thank you for your support always.

Finally, to Briggs Wakeman (and Weenee and Arugula), I couldn't have done this without all your support. Thank you for the encouragement, jokes, and many adventures. I love you.

0.2 Abstract

We rely on our episodic memory, the memory for personally experienced events, to remember the last vacation we took or to remember what was for dinner last night. In this dissertation, I use electroencephalographic (EEG) recordings to measure the neural correlates of episodic memory retrieval. I collected EEG recordings across a series of recognition memory tasks in which participants studied words or images and then were tested on their memory for those items. In the first chapter, I use pattern classification analyses to measure neural evidence for a retrieval brain state, a whole-brain activity/connectivity pattern that is engaged when an individual attempts to access a stored representation. I find that greater temporal overlap - the distance in time between two experiences - leads to automatic induction of the retrieval state and impairs memory of past events. In the second chapter, I investigate memory and decision making processes during retrieval. I find distinct processes occur prior to and following a memory response that are modulated by successful retrieval. In the next chapter I measure post-response feedback signals using pattern classification analyses. I find that post-retrieval neural signals reflect an intrinsic reward signal in response to successful retrieval. In the final chapter, I show that the memory benefit for extrinsic reward following retrieval may be dependent on the strength of the memory that is retrieved.

List of tables

List of figures

Contents

Chapter 1

Introduction

Episodic memory is our memory for personally experienced events. For example, I can vividly remember my sixteenth birthday at my mom's house when I was gifted my dachshund, Weenee, or the beginning of my fourth year of graduate school at the Shenandoah Valley Animal Services Center where I adopted my cat, Arugula. Episodic memories are tied to a spatiotemporal context, i.e., when and where an event occurred (F. Wang & Diana, 2017). Successful episodic memory relies on at least two components: encoding – the formation of new memories – and retrieval – accessing a stored representation of a past experience. Successful retrieval is often accompanied by the reinstatement of an original event's spatiotemporal context. The goal of this dissertation is to investigate the neural mechanisms of memory retrieval.

One method for measuring episodic memory in the laboratory is through the recognition paradigm. In the recognition paradigm, a participant studies a list of items, typically words or images, and then completes a memory test in which

the goal is to recognize previously presented items. This paradigm allows us to measure the processes that occur when someone remembers having previously seen an item. Cognitive processes such as memory and decision making occur quickly, within ten to hundreds of milliseconds (Cohen, 2014a). To separately investigate these fast occurring processes, I utilize scalp electroencephalography (EEG) which allows the measurement of neural signals on the order of milliseconds. In scalp EEG, several electrodes are placed on the scalp of a participant which directly measure voltage potentials across many neurons. The raw voltage trace can be decomposed into a series of sine waves at different frequencies (Cohen, 2014a). Frequency is continuous, but is often separated into five measured frequency bands which include delta (1-3 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (13-30 Hz) and gamma (30-100 Hz). Frequency signals reflect synchronous neuronal firing and communication across the brain (Fries, 2005) and activity in these frequency bands support different cognitive processes.

We engage in memory retrieval often and for many reasons. For example, if I'm going on a trip out of the country, I need to remember where my passport is. Did I give my dog his medication before leaving the house? Do I have the ingredients at home to make cinnamon rolls? In order to remember if I have the ingredients at home, I might use memory cues to access the stored information, such as picturing myself in the kitchen or what I last baked. The ability to jump back in time and reconstruct or reinstate these specific events has been characterized as "mental time travel" (Tulving, 1972). Greater reactivation of previously experienced events during a memory test can predict retrieval accuracy (Gordon, Rissman, Kiani, & Wagner, 2014). However, memory retrieval is imperfect and accurate information is not always reactivated (Schacter & Addis, 2007). Understanding how stored in-

formation is accessed is critical to cognition, yet the neural mechanisms underlying memory retrieval are unclear. Therefore, the projects described in this dissertation encompass four aims, the broad goal of which is to elucidate the neural mechanisms that support retrieval. In Aim 1, I test the hypothesis that the retrieval brain state – a whole-brain configuration of activity/connectivity patterns – can be induced automatically on the basis of temporal contextual overlap between experiences. In Aim 2, I test the hypothesis that decision-making processes are engaged following successful memory retrieval. In Aim 3, I test the hypothesis that post-retrieval neural signals reflect intrinsic reward in response to successful retrieval. Finally, in Aim 4, I test the hypothesis that extrinsic reward during retrieval practice reinforces the contents of retrieval and improves subsequent memory.

1.1 Memory retrieval brain state

A brain state is a whole-brain activity/connectivity pattern that is temporally sustained (Tang, Rothbart, & Posner, 2012; Long, 2023). The ability to form a representation of a new experience and to access a stored representation of a past experience involves the engagement of memory encoding and retrieval brain states, respectively. The brain must flexibly switch between these states to effectively learn new information and retrieve existing information. Electrophysiological work in rodents suggests that encoding and retrieval are neurally dissociable brain states that recruit distinct neural substrates, meaning they cannot be engaged in simultaneously (Hasselmo, Bodelon, & Wyble, 2002). Therefore, the ability to effectively engage or disengage the two states is crucial for successfully perceiving and converting new information into memories while also maintaining the integrity of past

information.

When one experience overlaps with another (e.g. encountering an acquaintance at both a coffee shop and the grocery store) interference can occur due to a tradeoff between encoding the present event and retrieving the past event. Competition or interference between overlapping memories can lead to forgetting (McGeoch, 1942; Anderson, 2003; Kuhl, Rissman, Chun, & Wagner, 2011). Temporal information - 'when' something occurred - is a defining feature of episodic memory (Tulving, 1993) and impacts how events are encoded and retrieved. Two events that occur closer together in time and/or space are more likely to be recalled together (Kahana, 1996; Manning, Polyn, Baltuch, Litt, & Kahana, 2011). According to Retrieved Context Theory, items are bound to their spatiotemporal (i.e. when and where) context during memory formation and memory retrieval is driven by reinstatement of the study context (Howard & Kahana, 2002; Polyn, Natu, Cohen, & Norman, 2005). The finding that participants tend to remember experiences in the order in which the initially occurred (Kahana, 1996) provides evidence in support of a context representation that drifts slowly over time. Comparison of activity patterns between study and test items provides support for Retrieved Context Theory in that the shorter the temporal distance or greater temporal overlap (i.e. proximity in time) between two items at study, the greater the similarity between the study pattern of one item and the test pattern of the other item (Manning et al., 2011; El-Kalliny et al., 2019). The interpretation is that this increased similarity reflects greater overlap in temporal context as events occur nearer to one another. However, the influence of temporal overlap between events on memory brain states is unclear.

A retrieval state, or mode, is a tonically maintained state that is entered when there is need to engage episodic retrieval – that is, intentionally accessing a

stored representation of a past experience situated within a spatiotemporal context (Tulving, 1983). Although memory retrieval is typically considered in relation to attempting to recall a past event during the test phase of an experiment, in reality, memory retrieval could occur while trying to form new memories. Specifically, retrieval could occur whenever there is temporal overlap between experiences. If (1) items are bound to the context in which they occur (2) neighboring items are bound to overlapping temporal context representations and (3) context is used to cue retrieval, items that occur close together in time may retrieve each other. In **Chapter 2**, I directly investigate the role of temporal overlap in the trade-off between memory states using scalp EEG.

1.2 Intrinsic feedback signal following successful memory retrieval

What motivates us to engage in retrieval? Successful retrieval may be intrinsically rewarding such that the act of correctly remembering which Oreo flavors I have tried engages the reward system. Remembering which flavors I've tried informs my decision making. If I know I've tried all the flavors except the new Churro, maybe that's what I'll decide to buy. Or if I know I liked the Peanut Butter Chocolate Pie ones the best, maybe that's what I end up getting. Prior work has linked the reward system – in particular, the striatum – to successful memory retrieval, whereby the striatum shows greater activity when a subject correctly recognizes an old stimulus (hit) compared to when the subject correctly rejects a new stimulus (correct rejection; CR, Achim & Lepage, 2005; Fliessbach, Weis, Klaver, Elger, & Weber, 2006; Spaniol et al., 2009; H. Kim, 2013; Clos, Schwarze,

Gluth, Bunzeck, & Sommer, 2015). As both hits and CRs constitute accurate trials, the dissociation in striatal activity for hits compared to CRs suggests that the signal change is not driven by dissociations between correct and incorrect trials, but rather is a response to successful retrieval. As this striatal response occurs absent of explicit rewards (e.g. monetary incentives or positive feedback), it may signal that successful retrieval is intrinsically rewarding (Satterthwaite et al., 2012). As I utilize scalp EEG to enable differentiation of memory vs. decision-making processes engaged during retrieval, I cannot directly measure striatal activity. Instead, I can leverage reward/feedback based spectral signals that have been identified within the cognitive control literature.

Theta power (4-8 Hz activity in EEG) is associated with episodic memory (Klimesch, Doppelmayr, Schimke, & Ripper, 1997; Nyhus & Curran, 2010; Herweg, Solomon, & Kahana, 2020) and supports feedback based learning (Cohen, 2014b). In particular, during the test-phase of a memory experiment, theta power is greater for hits vs CRs (Burgess & Gruzelier, 1997; Duzel, Neufang, & Heinze, 2005; Nyhus & Curran, 2010). In feedback based learning, theta power increases following incorrect relative to correct responses (Mazaheri, Nieuwenhuis, Dijk, & Jensen, 2009) and negative relative to positive outcomes (Cavanagh & Frank, 2014). Taken together, these findings point to a potential internally driven feedback signal following successful retrieval which may be indexable via theta power. In **Chapters 3** and **4**, I will address the role of theta power in successful memory retrieval.

The standard approach to investigating memory reinforcement is through the use of extrinsic reward (i.e. monetary compensation), however, providing an explicit reward for every successfully remembered experience does not reflect real-world scenarios and may, in fact, change intrinsic reward processing (Hidi, 2016).

However, although an intrinsic reward signal is more likely to reflect real-world processing, reliance on such a signal limits an experimenter's ability to manipulate reward signaling and control what is reinforced. Therefore, although extrinsic reward may not reflect real-world experiences, their application allows me to directly test how rewards following memory retrieval modulates memory processing.

1.3 Extrinsic feedback following memory retrieval

Establishing the way in which reward reinforces the contents of retrieval will provide insights into how memory can be improved through extrinsic rewards. The positive relationship between motivation and memory – as motivation increases, successful memory increases (Dickerson & Adcock, 2018) – suggests that manipulating motivation may be a viable approach for improving memory. Individuals can use reward to prioritize the storage of information. Evidence suggests that associating a study item with a potential reward (e.g. monetary compensation) – to be received if the item is remembered at test – impacts the likelihood that the item is later remembered (Loftus & Wickens, 1970; Adcock, Thangavel, Whitfield-Gabrieli, Knutson, & Gabrieli, 2006; Marini, Marzi, & Viggiano, 2011), with higher potential rewards leading to better subsequent memory. Collectively, these findings point to a clear role of study-phase anticipatory reward on memory enhancement.

Less straightforward is the role of *test phase* extrinsic reward on subsequent memory. Memory reinforcement may be better accomplished through direct reward of what is retrieved, rather than through the manipulation of potential for future reward during study. Prior behavioral work that has investigated test phase extrinsic reward have found mixed results, such that test phase rewards either

improve memory (Shigemune, Tsukiura, Nouchi, Kambara, & Kawashima, 2017) or have no effect on memory (Castanheira, Lalla, Ocampo, Otto, & Sheldon, 2022). However, both studies still used anticipatory methods such that participants were aware *prior* to retrieval that there was a potential to receive a reward for remembering. Additionally, these studies measured the immediate influence of reward on memory performance rather than the influence of reward on later memory. Thus, it may be through post-retrieval mechanisms, rather than pre-retrieval anticipatory mechanisms, that reward reinforces the contents of retrieval. In **Chapter 4**, we set out to identify the extent to which test phase extrinsic reward reinforces the contents of retrieval and impacts subsequent memory.

1.4 Overview

In this dissertation, I will attempt to address outstanding questions about the neural mechanisms of memory retrieval. **Chapter 2** reveals that temporal overlap between similar but non-identical stimuli induces automatic retrieval. **Chapter 3** investigates decision making mechanisms following successful retrieval. **Chapter 4** extends these findings to investigate whether post-retrieval decision making signals relate to feedback. **Chapter 4** determines the extent to which test-phase extrinsic reward modulates subsequent memory.

Chapter 2

Temporal context modulates encoding and retrieval of overlapping events

Devyn E. Smith, Isabelle L. Moore & Nicole M. Long
The Journal of Neuroscience, 42(14), 3000–3010

2.1 Abstract

Overlap between events can lead to interference due to a tradeoff between encoding the present event and retrieving the past event. Temporal context information – 'when' something occurred, a defining feature of episodic memory – can cue retrieval of a past event. However, the influence of temporal overlap, or proximity in time, on the mechanisms of interference are unclear. Here, by identifying brain states using scalp electroencephalography (EEG) from male and female human subjects, we show the extent to which temporal overlap promotes interference and

induces retrieval. In this experiment, subjects were explicitly directed to either encode the present event or retrieve a past, overlapping event while perceptual input was held constant. We find that the degree of temporal overlap between events leads to selective interference. Specifically, greater temporal overlap between two events leads to impaired memory for the past event selectively when the top-down goal is to encode the present event. Using pattern classification analyses to measure neural evidence for a retrieval state, we find that greater temporal overlap leads to automatic retrieval of a past event, independent of top-down goals. Critically, the retrieval evidence we observe likely reflects a general retrieval mode, rather than retrieval success or effort. Collectively, our findings provide insight into the role of temporal overlap on interference and memory formation.

2.2 Introduction

Overlap between events leads to interference and impairs memory for those events (McGeoch, 1942; Anderson, 2003). For example, at a conference you may talk to a colleague whom you had previously met over Zoom. Later you may have difficulty remembering either the original Zoom meeting or the subsequent conference conversation. The overlap between these events (e.g. the colleague) promotes retrieval of the past event (the meeting on Zoom) while you are trying to encode the present event (your conversation; Kuhl, Shah, DuBrow, & Wagner, 2010). As retrieval and encoding recruit distinct neural substrates and cannot be engaged in simultaneously (Hasselmo et al., 2002), retrieving the past comes at the expense of encoding the present (Long & Kuhl, 2019). Although overlap is a critical factor in retrieval-mediated interference, two events may overlap along many dimensions

and to varying degrees. Temporal overlap, or proximity in time, has been shown to enhance inference (Zeithamova & Preston, 2017), but it is unclear how temporal overlap contributes to interference. The aim of this study is to investigate the extent to which temporal overlap induces retrieval and, in turn, impacts interference.

Temporal information is a hallmark of episodic memory (Tulving, 1993) and is well known to impact how events are encoded and retrieved. The closer two events are in time and/or space the more likely they are to be recalled together (Kahana, 1996; Manning et al., 2011) and the greater their neural similarity (Manns, Howard, & Eichenbaum, 2007; Folkerts, Rutishauser, & Howard, 2018). Retrieved context theory (Howard & Kahana, 2002; Sederberg, Howard, & Kahana, 2008; Polyn, Norman, & Kahana, 2009; Lohnas & Kahana, 2014) provides an account for these effects whereby spatiotemporal context – an amalgamation of external stimuli and internal states – is bound, via the hippocampus, to the present experience (Eichenbaum, 2004; F. Wang & Diana, 2017; Long & Kahana, 2019; A. Yonelinas, Ranganath, Ekstrom, & Wiltgen, 2019) and is later used by the hippocampus as a cue to retrieve past experiences (Long et al., 2017). Comparison of activity patterns between study and test items – a recalled word or recognition probe – provides support for retrieved context theory in that the shorter the temporal distance between two items at study, the greater the pattern similarity between the study pattern of one item and the test pattern of the other item (Manning et al., 2011; Howard, Viskontas, Shankar, & Fried, 2012; El-Kalliny et al., 2019). Although contextually-mediated retrieval is typically considered in relation to the test phase of an experiment, in principle contextually-mediated retrieval should occur whenever there is a contextual overlap between items. Such retrieval may occur automatically, or independent from top-down demands (S. M. Smith, Handy,

Hernandez, & Jacoby, 2018). Therefore, we hypothesized that overlap in temporal context between two events produces retrieval during study and in turn promotes interference.

Here, we report a human scalp electroencephalography (EEG) study in which subjects studied two sets of object images in which the second set categorically overlapped with the first set. During study of the second set of object images, subjects were explicitly instructed to either encode the second (present) object or retrieve the first (past) object. These instructions were intended to bias subjects toward either an encoding or retrieval state. A retrieval state or *mode* is a tonically maintained mental set that is entered when there is need to engage episodic retrieval (Tulving, 1983; Rugg & Wilding, 2000). Our critical manipulation was the temporal distance between the first and second object, whereby the shorter the temporal distance between two objects, the greater their temporal contextual overlap. Following study, subjects completed a recognition task to probe their memory for all previously-presented objects. To the extent that temporal contextual overlap influences interference, we should find that temporal distance modulates memory performance for the first and/or second objects. To the extent that temporal contextual overlap promotes retrieval, we should find that subjects are biased toward a retrieval state during second objects that are presented near in time to a categorically overlapping first object.

2.3 Methods

Subjects

Forty (34 female; age range = 18-37, mean age = 20.3 years) right-handed, native English speakers from the University of Virginia community participated. This sample size is based on our previous work in which we enrolled 40 participants (Long & Kuhl, 2019). All subjects had normal or corrected-to-normal vision. Informed consent was obtained in accordance with the University of Virginia Institutional Review Board for Social and Behavioral Research and subjects were compensated for their participation. Three subjects were excluded from the final dataset: one who previously completed a behavioral version of the task, one who had poor task performance (recognition accuracy below three standard deviations of the mean of the full dataset), and one due to technical issues resulting in poor signal quality throughout the majority of the session. Thus, data are reported for the remaining 37 subjects. The raw, de-identified data and the associated experimental and analysis codes used in this study will be made available via the Long Term Memory laboratory website upon publication.

Mnemonic State Task Experimental Design

Stimuli consisted of 576 object pictures, drawn from an image database with multiple exemplars per object category (Konkle, Brady, Alvarez, & Oliva, 2010). From this database, we chose 144 unique object categories and 4 exemplars from each category. For each subject, one exemplar in a set of four served as a List 1 object,

one as a List 2 object, and the two remaining exemplars served as lures for the recognition phase. Object condition assignment was randomly generated for each subject.

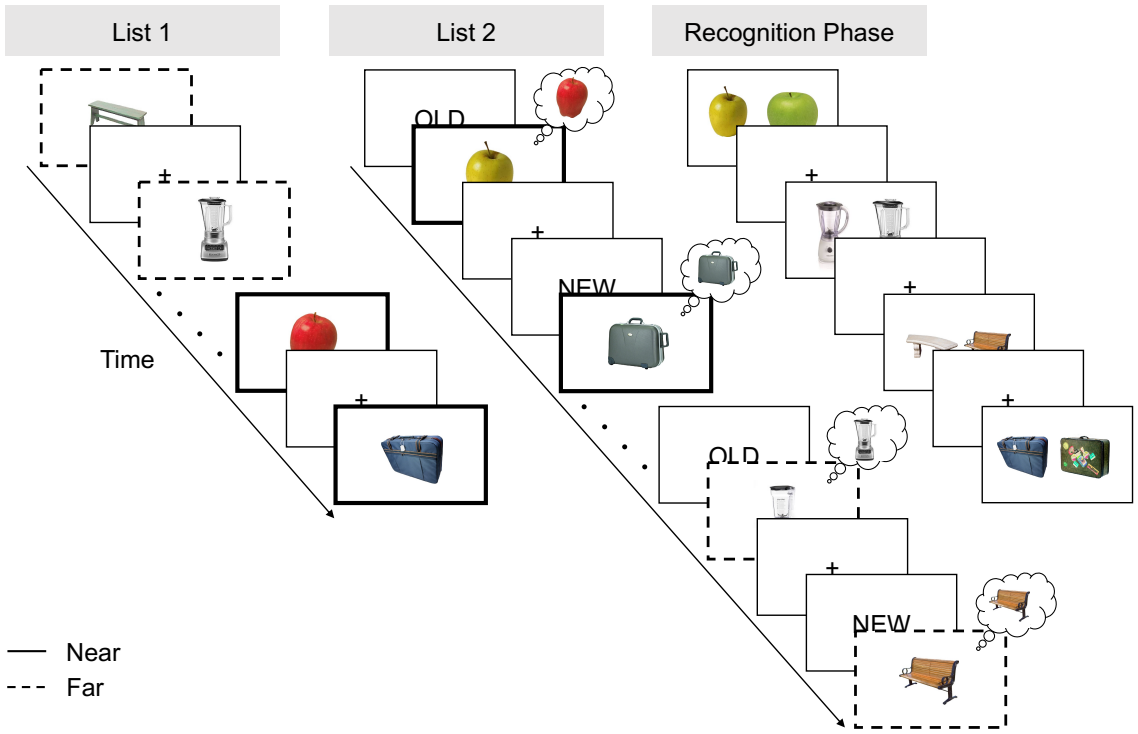

Figure 2.1: Task Design. During List 1, subjects studied individual objects (e.g. bench, apple). During List 2, subjects saw novel objects that were from the same categories as the objects shown in List 1 (e.g., a new bench, a new apple). Preceding each List 2 object was an "OLD" instruction cue or "NEW" instruction cue. The "OLD" cue signaled that subjects were to *retrieve* the corresponding object from List 1 (e.g., the old apple). The "NEW" cue signaled that subjects were to *encode* the current object (e.g. the new bench). Each run of the experiment contained a List 1 and List 2; object categories (e.g., bench) were not repeated across runs. List 1 and List 2 objects separated by fewer than 18 intervening objects were coded as *near* and List 1 and List 2 objects separated by 18 or more intervening objects were coded as *far* (see Methods). Lines around the boxes are shown for illustrative purposes and were not present during the actual experiment. After eight runs, subjects completed a two alternative forced choice recognition test that tested memory for each List 1 and List 2 object. On each trial, a previously presented object, either from List 1 or List 2, was shown alongside a novel lure from the same category. The subject's task was to choose the previously presented object. List 1 and List 2 objects were never presented together.

General Overview. In each of eight runs, subjects viewed two lists containing 18 object images. For the first list, each object was new (List 1 objects). For the second list (List 2 objects), each object was again new, but was categorically related

to an object from the first list. For example, if List 1 contained an image of a bench, List 2 would contain an image of a different bench (Figure 2.1). During List 1, subjects were instructed to encode each new object. During List 2, however, each trial contained an instruction to either encode the current object (e.g., the new bench) or to retrieve the corresponding object from List 1 (the old bench). The critical manipulation was the distance between the corresponding List 1 and List 2 objects. We divided each list of 18 objects into thirds according to serial position (first [1-6], middle [7-12], and last [13-18]). The objects in the first third of List 1 were "paired" with the objects in the last third of List 2. For example, if List 1 contained an image of a bench in serial position 1, List 2 would contain an image of a different bench in serial position 13-18. The objects in the middle third of List 1 were paired with the objects in the middle third of List 2. The objects in the last third of List 1 were paired with the objects in the first third of List 2. We coded List 1 and List 2 objects as *near* and *far* based on the lag, or difference in serial position, between the two objects in a pair. List 1 and List 2 objects separated by fewer than 18 intervening objects were coded as *near*; List 1 and List 2 objects separated by 18 or more intervening objects were coded as *far*. Following eight runs, subjects completed a two-alternative forced-choice recognition test that separately assessed memory for List 1 and List 2 objects.

List 1. On each trial, subjects saw a single object presented for 2000 ms followed by a 1000 ms inter-stimulus interval (ISI). Subjects were instructed to study the presented object in anticipation for a later memory test.

List 2. On each trial, subjects saw a cue word, either "OLD" or "NEW" for 2000 ms. The cue was followed by presentation of an object for 2000 ms, which was followed by a 1000 ms ISI. All objects in List 2 were non-identical exemplars drawn

from the same category as the objects presented in the immediately preceding List 1. That is, if a subject saw a bench and an apple during List 1, a different bench and a different apple would be presented during List 2. On trials with a "NEW" instruction (encode trials), subjects were to encode the presented object. On trials with an "OLD" instruction (retrieve trials), subjects tried to retrieve the categorically related object from the preceding List 1. Importantly, this design prevented subjects from completely ignoring List 2 objects following "OLD" instructions in that they could only identify the to-be-retrieved object category by processing the List 2 object.

Subjects completed eight runs with two lists in each run (List 1, List 2). Subjects viewed 18 objects per list, yielding a total of 288 object stimuli from 144 unique object categories. Subjects did not make a behavioral response during either List 1 or 2. Following the eight runs, subjects completed a two-alternative forced choice recognition test.

Recognition Phase. Following the eight runs, subjects completed the recognition phase. On each trial, subjects saw two exemplars from the same object category (e.g. two benches; Figure 2.1). One object had previously been encountered either during List 1 or 2. The other object was a lure and had not been presented during the experiment. Because both test probes were from the same object category, subjects could not rely on familiarity or gist-level information to make their response (Brainerd & Reyna, 2002). Trials were self-paced and subjects selected (via button press) the previously presented object. Trials were separated by a 1000 ms ISI. There were a total of 288 recognition trials (corresponding to the 288 total List 1 and 2 objects presented in the experiment). Note: List 1 and List 2 objects never appeared in the same trial together, thus subjects never had to choose between two

previously presented objects. List 1 and List 2 objects were presented randomly throughout the test phase.

EEG Data Acquisition and Preprocessing

EEG recordings were collected using a BrainVision system and an ActiCap equipped with 64 Ag/AgCl active electrodes positioned according to the extended 10-20 system. All electrodes were digitized at a sampling rate of 1000 Hz and were referenced to electrode FCz. Offline, electrodes were later converted to an average reference. Impedances of all electrodes were kept below 50 kΩ. Electrodes that demonstrated high impedance or poor contact with the scalp were excluded from the average reference. Bad electrodes were determined by voltage thresholding (see below).

Custom Python codes were used to process the EEG data. We applied a high pass filter at 0.1 Hz, followed by a notch filter at 60 Hz and harmonics of 60 Hz to each subject's raw EEG data. We then performed three preprocessing steps (Nolan, Whelan, & Reilly, 2010) to identify electrodes with severe artifacts. First, we calculated the mean correlation between each electrode and all other electrodes as electrodes should be moderately correlated with other electrodes due to volume conduction. We z-scored these means across electrodes and rejected electrodes with z-scores less than -3. Second, we calculated the variance for each electrode as electrodes with very high or low variance across a session are likely dominated by noise or have poor contact with the scalp. We then z-scored variance across electrodes and rejected electrodes with a $|z| >= 3$. Finally, we expect many electrical signals to be autocorrelated, but signals generated by the brain versus noise are

likely to have different forms of autocorrelation. Therefore, we calculated the Hurst exponent, a measure of long-range autocorrelation, for each electrode and rejected electrodes with a $|z| >= 3$. Electrodes marked as bad by this procedure were excluded from the average re-reference. We then calculated the average voltage across all remaining electrodes at each time sample and re-referenced the data by subtracting the average voltage from the filtered EEG data. We used wavelet-enhanced independent component analysis (Castellanos & Makarov, 2006) to remove artifacts from eyeblinks and saccades.

EEG Data Analysis

We applied the Morlet wavelet transform (wave number 6) to the entire EEG time series across electrodes, for each of 46 logarithmically spaced frequencies (2-100 Hz; Long & Kahana, 2015). After log-transforming the power, we downsampled the data by taking a moving average across 100 ms time intervals from either 4000 ms preceding to 4000 ms following object presentation during List 1 and List 2 or 0 ms preceding to 1000 ms following probe presentation for the recognition data. For each phase, we slid the window every 25 ms, resulting in 317 and 37 time intervals, respectively (80 and 10 non-overlapping). Power values were then z-transformed by subtracting the mean and dividing by the standard deviation power. Mean and standard deviation power were calculated across all List 1 and List 2 objects or all recognition events, across time points for each frequency, which is analogous to performing a pre-stimulus baseline correction. Z-transforming or baseline correcting spectral power is a necessary step to both reduce the $1/f$ shape of the power spectrum – lower frequencies inherently have more power than higher frequencies – and to perform parametric statistics on the data (Cohen, 2014a).

General Linear Model

Trial-specific signals during List 2 were estimated using the General Linear Model (GLM) implemented via the sklearn linear model module in Python. We ran a separate GLM for each trial in which the trial was modeled as the regressor of interest and all other trials were combined into a single nuisance regressor (Mumford, Turner, Ashby, & Poldrack, 2012). Serial position (1-36, corresponding to List 1 [1-18] and List 2 [19-36]) was included as a single parametric regressor in each GLM to account for serial position effects. This parametric regressor predicts recognition memory accuracy, such that memory declines as serial position increases (M = -0.0108, SD = 0.0196, t_{36} = -3.3116, p = 0.0021), and is consistent with other approaches for modeling a continuous variable (e.g. Long, Öztekin, & Badre, 2010; Spitzer, Gloel, Schmidt, & Blankenburg, 2014; Tuladhar et al., 2007). We fit trial-specific GLMs to the z-scored spectral power for each time point and frequency for each electrode in order to generate trial-level beta values. These beta values were used in all subsequent analyses.

Pattern Classification Analyses

Pattern classification analyses were performed using penalized (L2) logistic regression (penalty parameter = 1), implemented via the sklearn linear model module in Python. Before pattern classification analyses were performed on the List 2 data, an additional round of z-scoring was performed across features (electrodes and frequencies) to eliminate trial-level differences in spectral power (Kuhl & Chun, 2014; Long & Kuhl, 2018). Therefore, mean univariate activity was matched precisely across all conditions and trial types. Classifier performance was assessed in two

ways. "Classification accuracy" represented a binary coding of whether the classifier successfully guessed the instruction condition. We used classification accuracy for general assessment of classifier performance (i.e., whether encode/retrieve instructions could be decoded). "Classifier evidence" was a continuous value reflecting the logit-transformed probability that the classifier assigned the correct instruction for each trial. Classifier evidence was used as a trial-specific, continuous measure of mnemonic state information, which was used to assess the degree of retrieval evidence present on *near* and *far* trials. The logic of using both classifier accuracy and classifier evidence is that although accuracy indicates how well the classifier can distinguish encode vs. retrieve trials, accuracy may obscure differences in conditions upon which the classifier was not directly trained, e.g. the distance (*near*, *far*) between objects. As an example, the classifier may correctly label both *near* encode and *far* encode trials as "encode," however, it may have less confidence on the *near* compared to *far* trials, reflecting relatively greater retrieval state evidence on *near* trials.

We trained within-subject classifiers to discriminate List 2 encode vs. retrieve trials based on a feature space comprised of all 63 electrodes $\times$ 46 logarithmically spaced frequencies ranging from 2 to 100 Hz. For each subject, we used leave-one-run-out cross validated classification in which the classifier was trained to discriminate encode from retrieve instructions for seven of the eight runs and tested on the held-out run. For classification analyses in which we assessed classifier accuracy, we averaged beta values over the 2000 ms stimulus interval. For analyses measuring classifier evidence, we averaged beta values over four separate 500-ms time intervals across the 2000 ms stimulus interval. We assessed classifier evidence as a function of instruction (encode, retrieve), temporal distance (*near*, *far*), and/or

retrieval status (success, failure; see below).

To measure the ability of the classifier to generalize across temporal distance, we trained and tested two separate classifiers to distinguish List 2 encode/retrieve trials. One classifier was trained on *near* trials and tested on *far* trials, the other classifier was trained on *far* trials and tested on *near* trials. As there was a slight imbalance in the number of encode and retrieve trials within each distance, we subsampled trials from the condition with the greater number of trials to match the condition with fewer trials. We repeated this procedure for 100 iterations and averaged the resulting classification accuracy values across the 100 iterations.

Retrieval Status Analysis

Because we did not explicitly measure retrieval success during the List 2 trials, we generated 'retrieval success' and 'retrieval failure' templates based on the recognition phase data. Specifically, we extracted stimulus-locked z-scored spectral power across 63 electrodes and 46 frequencies separately for hits (trials in which participants selected the target) and misses (trials in which participants selected the categorically-related lure). We extracted z-power 500 - 800 ms post-stimulus, as this interval has routinely been linked with retrieval success (Friedman & Johnson Jr., 2000; Voss & Paller, 2008; Johnson, Price, & Leiker, 2015). We averaged z-power across all subjects to generate a single 'retrieval success' template and a single 'retrieval failure' template.

After having generated the success/failure templates, we applied these templates to the List 2 data. Because we were interested in whether or not the corresponding List 1 object was retrieved at any point within the List 2 trial, we used the

trial-level beta values averaged across the stimulus interval (2000 ms). We correlated trial-level beta values with the success and failure templates using a Pearson correlation. Each trial was assigned a label based on its correlation with the success and failure templates. A trial that was more positively correlated with the success template was labeled 'retrieval success' or 1, and a trial that was more positively correlated with the failure template was labeled 'retrieval failure' or 0.

We calculated the average label as a function of distance (*near*, *far*) and instruction (encode, retrieve). An average label value of 0.5 means that a given condition was no more likely to be labeled 'retrieval success' than 'retrieval failure.' An average label value greater than 0.5 means that a given condition was more likely to be labeled 'retrieval success' than 'retrieval failure.'

Statistical Analyses

We used repeated measures ANOVAs and paired-sample *t*-tests to assess the effect of instruction (encode, retrieve) and temporal distance (*near*, *far*) on behavioral memory performance.

We used paired-sample *t*-tests to compare classification accuracy across subjects to chance decoding accuracy, as determined by permutation procedures. Namely, for each subject we shuffled the condition labels of interest (e.g., encode and retrieve for the List 2 instruction classifier) and then calculated classification accuracy. We repeated this procedure 1000 times for each subject and then averaged the 1000 shuffled accuracy values for each subject. These mean values were used as subject-specific empirically derived measures of chance accuracy.

We used repeated measures ANOVAs and paired-sample *t*-tests to assess the interaction between instruction (encode, retrieve), temporal distance (*near*, *far*), and time interval on retrieval evidence.

2.4 Results

Influence of Temporal Contextual Overlap on Interference

We first sought to replicate the finding that subjects are able to shift between encoding and retrieval states in a goal directed manner (Long & Kuhl, 2019), by testing whether instructions influenced performance on the recognition task. Although encode/retrieve instructions only appeared during List 2, we also considered whether memory for List 1 objects was influenced by List 2 instructions (e.g., whether memory for the old bench was influenced by whether the new bench was associated with an encode vs. retrieve instruction). A two-way, repeated measures ANOVA with factors of list (1, 2) and instruction (encode, retrieve) revealed a list by instruction interaction ($F_{1,36}$ = 6.045, p = 0.0189, η_p^2 = 0.14, Figure 2.2A). This interaction was driven by numerically greater recognition for List 2 objects presented with an encode (M = 82.88%, SD = 8.51%) relative to a retrieve instruction (M = 80.52%, SD = 7.79%; difference between List 2 encode vs retrieve: t_{36} = 2.1072, p = 0.0421, Bonferroni corrected α = 0.025, Cohen's d = 0.2938) and numerically greater recognition for List 1 objects presented with a retrieve (M = 84.27%, SD = 7.7%) relative to an encode instruction (M = 83.3%, SD = 7.03%; difference between List 1 encode vs retrieve: t_{36} = -1.7542, p = 0.0879, Bonferroni corrected α = 0.025, Cohen's d = 0.1324).

To further demonstrate the impact that encode vs. retrieve instructions have on memory behavior, we conducted an analysis of recognition phase reaction times. If subjects are able to shift between encoding and retrieval states, we would expect to find a list by instruction interaction such that memory responses are slowed for List 2 objects associated with a retrieve instruction and List 1 objects associated with an encode instruction. We assessed reaction times from correct trials only. A two-way, repeated measures ANOVA with factors of list (1, 2) and instruction (encode, retrieve) revealed a significant main effect of list ($F_{1,36}$ = 24.84, $p < 0.0001$, η_p^2 = 0.41) driven by faster reaction times for List 1 compared to List 2 objects. There was a main effect of instruction ($F_{1,36}$ = 7.27, p = 0.0106, η_p^2 = 0.17) driven by faster reaction times for encode compared to retrieve instructions. There was a significant interaction between list and instruction ($F_{1,36}$ = 12.9, p = 0.0010, η_p^2 = 0.26, Figure 2.2B). This interaction was driven by faster reaction times for List 2 objects presented with an encode (M = 1.6456, SD = 0.4405) relative to a retrieve instruction (M = 1.7904, SD = 0.3954; difference between List 2 encode vs retrieve: t_{36} = -4.248, p = 0.0001, Bonferroni corrected α = 0.025, Cohen's d = 0.346).

We next assessed the relationship between List 1 and List 2 object memory on a pair-by-pair basis to investigate the encoding-retrieval tradeoff. We isolated cases in which either the List 1 object was remembered and the associated List 2 object was forgotten (L1R-L2F) or the List 1 object was forgotten and the associated List 2 object was remembered (L1F-L2R). To the extent that retrieval of List 1 objects trades off with encoding of List 2 objects, the proportion of L1R-L2F should be greater for retrieve compared to encode instructions and the proportion of L1F-L2R should be greater for encode compared to retrieve instructions. To test this hypothesis, we ran a 2×2 repeated measures ANOVA with factors of instruction (encode, retrieve) and

condition (L1R-L2F, L1F-L2R) and proportion as the dependent variable. There was no main effect of instruction ($F_{1,36}$ = 0.471, p = 0.497, η_p^2 = 0.01) and the main effect of condition did not reach significance ($F_{1,36}$ = 3.923, p = 0.0553, η_p^2 = 0.10). There was a significant interaction between condition and instruction ($F_{1,36}$ = 6.045, p = 0.0189, η_p^2 = 0.14, Figure 2.2C). This interaction was driven by a numerically greater proportion of L1F-L2R items when the instruction was to encode (M = 0.1288, SD = 0.0507) compared to retrieve (M = 0.1152, SD = 0.0569; difference between L1F-L2R encode vs retrieve: t_{36} = 2.1733, p = 0.0364, Bonferroni corrected α = 0.025, Cohen's d = 0.2507) and a numerically greater proportion of L1R-L2F items when the instruction was to retrieve (M = 0.1528, SD = 0.0597) compared to encode (M = 0.1329, SD = 0.0599; difference between L1R-L2F encode vs retrieve: t_{36} = -2.0211, p = 0.0508, Bonferroni corrected α = 0.025, Cohen's d = 0.3325). Together, these results support the interpretation that encoding and retrieval processes tradeoff.

Having replicated our previous finding that instructions to encode and retrieve modulate behavior, we next sought to test the effect of temporal distance on recognition accuracy, specifically for List 1 objects, as shorter temporal distance may impair List 1 memory specifically for encode trials. The intuition is that automatically retrieved *near* List 1 objects may be inhibited or suppressed by virtue of being goal-irrelevant during encode trials. This outcome would be analogous to the inhibition that is thought to occur during retrieval induced forgetting (Anderson, Bjork, & Bjork, 1994; Anderson, 2003).

We assessed whether the distance between objects, as well as the instruction given during List 2, influenced recognition memory List 1 objects (Figure 2.3). A two-way, repeated measures ANOVA with factors of instruction (encode, retrieve) and distance (*near*, *far*), revealed a significant main effect of distance ($F_{1,36}$ = 4.916,

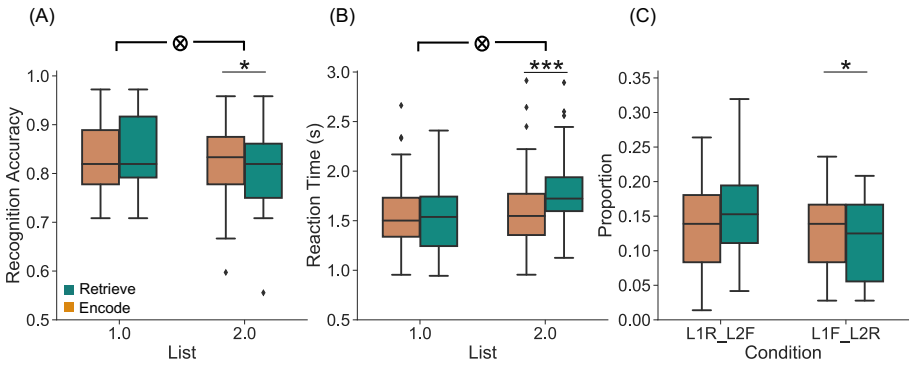

Figure 2.2: Influence of Mnemonic Instructions on Memory Behavior. (A) We assessed recognition accuracy as a function of list (1, 2) and instruction (encode, orange; retrieve, teal). We find a significant interaction between list and instruction (p = 0.0189) driven by greater accuracy for List 2 objects presented with an encode compared to a retrieve instruction and numerically greater accuracy for List 1 objects presented with a retrieve compared to an encode instruction. **(B)** We assessed reaction times as a function of list and instruction. We find a significant interaction between list and instruction (p = 0.0010) driven by faster reaction times for List 2 objects presented with an encode compared to a retrieve instruction. **(C)** We assessed the relationship between List 1 and List 2 object memory on a pair-by-pair basis for cases where either the List 1 object was remembered and the associated List 2 object was forgotten (L1R-L2F) or the List 1 object was forgotten and the associated List 2 object was remembered (L1F-L2R) separately for encode and retrieve instructions. There was a significant interaction between condition and instruction (p = 0.0189) driven by a greater proportion of L1F-L2R items for encode compared to retrieve trials and a numerically greater proportion of L1R-L2F items for retrieve compared to encode trials. * p < 0.05, *** p < 0.001, uncorrected.

p = 0.0330, η_p^2 = 0.12) driven by greater recognition accuracy for *far* compared to *near* objects. The main effect of instruction did not reach significance ($F_{1,36}$ = 3.769, p = 0.0601, η_p^2 = 0.09). There was a significant interaction between instruction and distance ($F_{1,36}$ = 4.381, p = 0.0435, η_p^2 = 0.11), driven by greater accuracy for *near* retrieve trials (M = 83.88%, SD = 11.11%) relative to *near* encode trials (M = 80.9%, SD = 9.17%; difference between *near* encode vs *near* retrieve: t_{36} = -2.6225, p = 0.0127, Bonferroni corrected α = 0.0167, Cohen's d = 0.2964). Notably, recognition accuracy on *near* encode trials was significantly worse compared to both *far* encode trials (t_{36} = -3.3417, p = 0.0020, Bonferroni corrected α = 0.0167, Cohen's d = 0.5561) and *far* retrieve trials (t_{36} = -3.1204, p = 0.0035, Bonferroni corrected α = 0.0167, Cohen's d = 0.4653).

We observed decreased recognition accuracy for List 1 *near* objects when subjects attempted to encode the List 2 object compared to when they attempted to retrieve the *near* List 1 object. In fact, *near* List 1 objects paired with the encode instruction are remembered worse than all other List 1 objects, strongly suggesting that bottom-up or automatic retrieval of the *near* List 1 object, when coupled with the top-down demand to encode the List 2 object, leads to suppression of the *near* List 1 object.

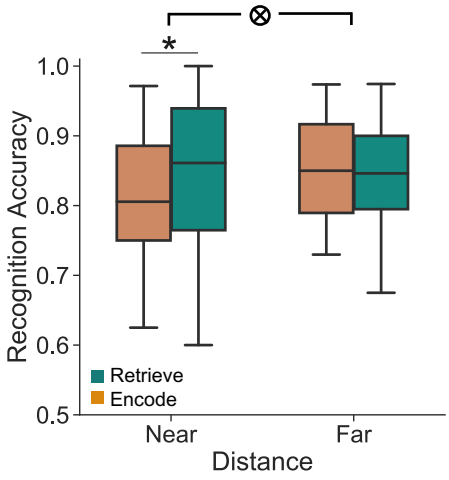

Figure 2.3: List 1 Recognition Accuracy by Instruction and Distance. We assessed recognition accuracy for List 1 objects as a function of instruction (encode, orange; retrieve, teal) and distance (*near*, *far*). We find a significant interaction between instruction and distance ($p = 0.0435$) driven by greater accuracy for *near* retrieve trials compared to *near* encode trials. * $p < 0.05$, uncorrected.

Influence of Temporal Contextual Overlap on Retrieval State

Our first goal was to replicate our previous finding that a pattern classifier trained on spectral signals can distinguish encode and retrieve trials (Long & Kuhl, 2019). We conducted a multivariate pattern classification analysis in which we trained a classifier to discriminate encode vs. retrieve List 2 trials based on a feature space comprised of all 63 electrodes and 46 frequencies ranging from 2 to 100 Hz. For this analysis, we averaged beta values over the 2000 ms stimulus interval. Us-

ing within-subject, leave-one-run-out classifiers, mean classification accuracy was 53.32% (SD = 7.75%), which was significantly greater than chance, as determined by permutation tests (t_{36} = 2.5595, p = 0.0148, Cohen's d = 0.6043; Figure 2.4A).

We next sought to investigate the effect of temporal overlap on retrieval. If greater temporal contextual overlap between two events promotes retrieval, we would expect to find greater evidence for a retrieval state on *near* compared to *far* trials. Moreover, to the extent that this retrieval occurs automatically, we would expect to find greater evidence for a retrieval state early in the stimulus interval. Although temporal distance could interact with instruction – evidence for a retrieval state may be particularly strong for *near* retrieve trials – given that temporal distance did not enhance memory for *near* List 1 objects on retrieve trials or impact memory for List 2 objects, we do not anticipate an interaction between temporal distance and instruction.

To investigate the effect of temporal distance on retrieval state evidence over time, we trained classifiers to discriminate encode vs. retrieve trials using the average betas from four 500 ms time intervals across the 2000 ms stimulus interval. We conducted a repeated measures ANOVA in which true (non-permuted) retrieval evidence was the dependent variable and with factors of instruction (encode, retrieve), distance (*near*, *far*) and time interval (four 500 ms time intervals). We find a significant two-way interaction between distance and time interval ($F_{3,108}$ = 5.355, p = 0.0018, η_p^2 = 0.13) whereby retrieval evidence is greater for *near* compared to *far* trials during the first two 500 ms time intervals (*near* vs. *far*: 0-500, t_{36} = 2.4899, p = 0.0175, Cohen's d = 0.598; 500-1000, t_{36} = 4.159, p = 0.0002, Cohen's d = 0.9056; Bonferroni corrected α = 0.0125). Retrieval evidence does not differ during the second two 500 ms time intervals (*near* vs. *far*: 1000-1500, t_{36} = -1.2887, p = 0.2057,

Cohen's d = 0.2772; 1500-2000, t_{36} = 0.9867, p = 0.3304, Cohen's d = 0.2492; Bonferroni corrected α = 0.0125). We also find a main effect of distance ($F_{1,36}$ = 8.649, p = 0.0057, η_p^2 = 0.19; Figure 2.4B), with greater retrieval evidence for *near* compared to *far* trials. We find a significant two-way interaction between instruction and time interval ($F_{3,108}$ = 5.041, p = 0.0026, η_p^2 = 0.12), whereby the largest differences in retrieval evidence between retrieve and encode trials occur during the last two 500 ms time intervals (encode vs. retrieve: 0-500, t_{36} = -1.4215, p = 0.1638, Cohen's d = 0.3759; 500-1000, t_{36} = -1.9205, p = 0.0628, Cohen's d = 0.4996; 1000-1500, t_{36} = -4.2349, p = 0.0002, Cohen's d = 1.2395; 1500-2000, t_{36} = -4.4573, p = 0.0001, Cohen's d = 1.2841; Bonferroni corrected α = 0.0125). We find a significant main effect of instruction ($F_{1,36}$ = 22.31, $p < 0.0001$, η_p^2 = 0.38; Figure 2.4C), consistent with the results of the classifier trained on the full 2000 ms interval above. The two-way interaction between instruction and distance was not significant ($F_{1,36}$ = 1.932, p = 0.173, η_p^2 = 0.05) nor was the three-way interaction between instruction, distance, and time interval ($F_{3,108}$ = 0.239, p = 0.869, η_p^2 = 0.0066; Figure 2.4D). Together, these results suggest that greater temporal contextual overlap induces automatic retrieval independent of the actual instruction to either encode or retrieve.

Retrieval State Mechanisms

We have found an increase in retrieval state evidence when objects appear closer together in time. Although our hypothesis is that this dissociation reflects greater instantiation of a retrieval state, the classifier may be indexing retrieval success or retrieval effort as opposed to a general retrieval state or mode (Rugg & Wilding, 2000). Specifically, by virtue of the shorter temporal distance, retrieval success might be greater for *near* compared to *far* objects. Likewise, by virtue of the longer

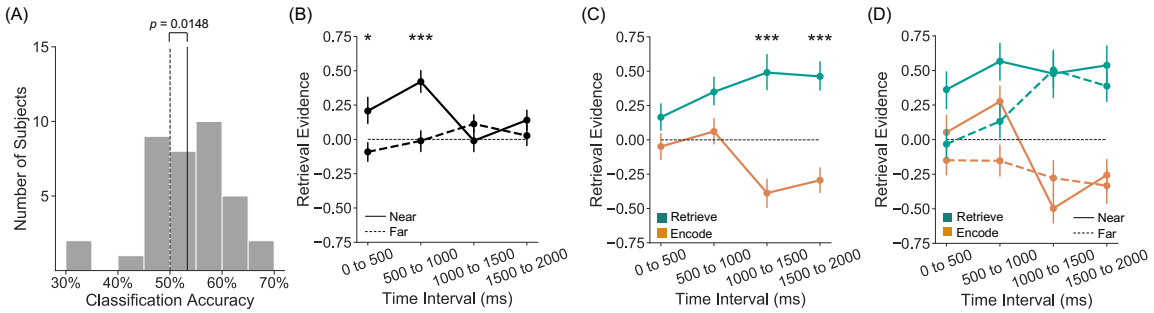

Figure 2.4: Retrieval State Evidence. We trained an L2-logistic regression classifier to discriminate encode vs. retrieve trials during List 2. The classifier was trained and tested on averaged beta values across 63 electrodes and 46 frequencies. **(A)** The classifier was trained on average beta values across the 2000 ms stimulus interval. Mean classification accuracy across all subjects (solid vertical line) is shown along with a histogram of classification accuracies for individual subjects (gray bars) and mean classification accuracy for permuted data across all subjects (dashed vertical line). Mean classification accuracy for permuted data ranged from 49.7% to 50.27% across individual subjects (1000 permutations per subject). Mean classification accuracy was 53.32%, which differed significantly from chance (p = 0.0148). **(B-D)** We trained and tested four classifiers on four 500 ms time intervals within the 2000 ms stimulus interval. **(B)** When we average retrieval evidence over instruction, we find a significant interaction between distance and time interval (p = 0.0018) driven by greater retrieval evidence on *near* compared to *far* trials early in the stimulus interval. **(C)** When we average retrieval evidence over distance, we find a significant interaction between instruction and time interval (p = 0.0026) driven by greater retrieval evidence on retrieve compared to encode trials late in the stimulus interval. **(D)** We do not find a three-way interaction between instruction, distance, and time (p = 0.869). Error bars denote SEM. * p < 0.05, *** p < 0.001, uncorrected.

temporal distance, retrieval might be more effortful for *far* compared to *near* objects. In our previous classification analysis, the classifier was trained using data from both *near* and *far* trials, meaning that the dissociation between encode/retrieve trials, and consequently, *near/far* trials, could be based on information exclusively from either *near* or *far* trials. Put another way, the classifier may have learned to distinguish either encode and 'retrieval success' (i.e. *near* retrieve) trials or encode and 'retrieval effort' (i.e. *far* retrieve) trials. Therefore, to demonstrate that a general retrieval state or mode underlies the dissociation between *near* and *far* trials, we trained two separate classifiers to distinguish encode/retrieve using only *near* or only *far* trials, and tested the classifiers on the other held-out distance (*far* or *near*) trials. The logic is that to the extent that the dissociation between encode/retrieve

is supported by the same mechanism on both *near* and *far* trials, classifiers trained on one distance should generalize – reflected by above chance (50%) performance – to the other distance. To the extent that the dissociation between encode/retrieve is driven either by retrieval success or retrieval effort, the classifiers should fail to generalize to the other distance.

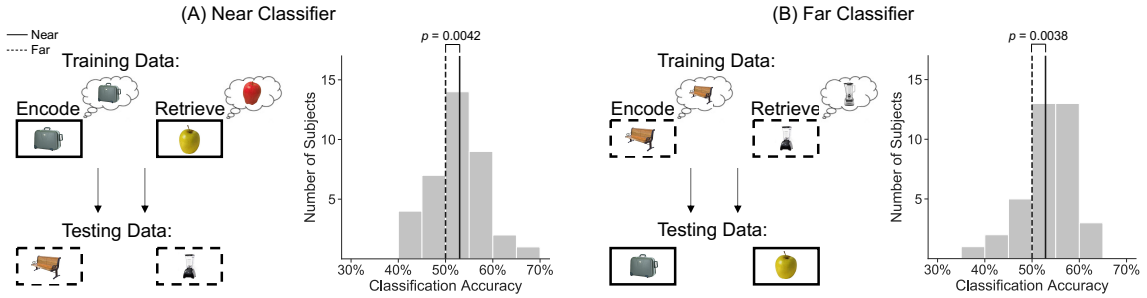

Figure 2.5: Cross Distance Mnemonic State Decoding. We trained two L2-logistic regression classifiers to discriminate encode vs. retrieve based on average beta values for the 2000 ms stimulus interval with 63 electrodes and 46 frequencies used as features. For each classifier we show mean classification accuracy across all subjects (solid vertical line) along with a histogram of classification accuracies for individual subjects (gray bars) and mean classification accuracy for permuted data across all subjects (dashed vertical line). **(A)** We trained the classifier on only List 2 *near* trials and tested the classifier on List 2 *far* trials. Mean classification accuracy for permuted data ranged from 49.73% to 50.40% across individual subjects (1000 permutations per subject). Mean classification accuracy was 52.98%, which was significantly greater than chance performance (p = 0.0042). **(B)** We trained the classifier on only List 2 *far* trials and tested the classifier on List 2 *near* trials. Mean classification accuracy for permuted data ranged from 49.27% to 50.46% across individual subjects (1000 permutations per subject). Mean classification accuracy was 52.85%, which was significantly greater than chance performance (p = 0.0038).

We conducted a multivariate pattern classification analysis in which we trained a classifier on only *near* or *far* trials to discriminate encode vs. retrieve trials. We averaged beta values across the stimulus interval (2000 ms) and used leave-one-run-out cross-validated classification. First, we trained a classifier to distinguish encode vs. retrieve List 2 *near* trials and tested the classifier on the List 2 *far* trials (Figure 2.5A). Mean classification accuracy was 52.98% (SD = 5.77%), which was significantly greater than chance performance (t_{36} = 3.0602, p = 0.0042, Cohen's d = 0.7225; Figure 2.5A). Next, we trained a classifier to distinguish encode vs. retrieve

List 2 *far* trials and tested the classifier on List 2 *near* trials (Figure 2.5B). Mean classification accuracy was 52.85% (SD = 5.48%), which was significantly above chance (t_{36} = 3.0933, p = 0.0038, Cohen's d = 0.7377; Figure 2.5B). The ability of these classifiers to generalize across distance suggests that neural signals during encode and retrieve trials are similar across temporal distance.

The cross-distance decoding analysis suggests that a general retrieval mode is present during both *near* and *far* trials. However, it is possible that the dissociation we observe between *near* and *far* trials in our prior analysis of retrieval state evidence is still driven in some part by retrieval success. Namely, greater retrieval state evidence may specifically be tracking *near* - success trials.

To adjudicate between the possibilities that elevated retrieval evidence on *near* trials is due to a retrieval mode vs. retrieval success, it is necessary to account for retrieval success during each List 2 trial. By design, there are no behavioral responses made during List 2 trials in order to equate the behavioral output across instructions. Therefore, we do not have a direct measure of retrieval success. However, we can generate a proxy of retrieval success by leveraging the recognition phase data. Specifically, we created a 'retrieval success' and a 'retrieval failure' template (Figure 2.6A) across all subjects and assigned a retrieval 'status' label to each List 2 trial of either 'retrieval success' (1) or 'retrieval failure' (0) based on how well a given trial correlated with each template (see Methods).

To validate our proxy of retrieval success, we first assessed whether temporal overlap impacts retrieval success. Given that retrieval success should be more likely for *near* compared to *far* objects, we predicted that *near* trials should be labeled 'retrieval success' more often than *far* trials, reflected by an average label

value closer to 1. We conducted a two-way, repeated measures ANOVA with factors of instruction (encode, retrieve) and distance (*near*, *far*) and the average retrieval status label as the dependent variable (Figure 2.6B). We find a significant main effect of distance ($F_{1,36}$ = 11.32, p = 0.0018, η_p^2 = 0.24) driven by greater assignment of 'retrieval success' for *near* compared to *far* trials. We find no main effect of instruction ($F_{1,36}$ = 0.104, p = 0.749, η_p^2 = 0.0029) and no interaction between instruction and distance ($F_{1,36}$ = 0.351, p = 0.557, η_p^2 = 0.0097).

Having established that our proxy for retrieval success matches our predictions, we next sought to test whether retrieval state evidence differs as a function of retrieval success. If the output of a classifier trained on all List 2 trials purely reflects a retrieval mode, *near* trials should show greater retrieval state evidence than *far* trials regardless of retrieval success. If the classifier purely reflects retrieval success, retrieval success trials should show greater retrieval state evidence than retrieval failure trials regardless of distance. We conducted a repeated measures ANOVA in which true (non-permuted) retrieval evidence was the dependent variable with factors of retrieval status (success, failure), distance (*near*, *far*) and time interval (four 500 ms time intervals). We find a significant main effect of distance ($F_{1,36}$ = 7.564, p = 0.0093, η_p^2 = 0.17; Figure 2.6C) driven by greater retrieval evidence on *near* compared to *far* trials. We find a significant two-way interaction between distance and time interval ($F_{3,108}$ = 5.853, p = 0.0010, η_p^2 = 0.14) whereby retrieval evidence is greater for *near* compared to *far* trials for the first two 500 ms time intervals (*near* vs. *far*: 0-500, t_{36} = 2.579, p = 0.0141, Cohen's d = 0.6058; 500-1000, t_{36} = 3.973, p = 0.0003, Cohen's d = 0.8835; Bonferroni corrected α = 0.0125). Retrieval evidence does not differ during the second two 500 ms time intervals (*near* vs. *far*: 1000-1500, t_{36} = -1.5492, p = 0.1301, Cohen's d = 0.3336; 1500-2000, t_{36} = 0.8985, p = 0.3749,

Cohen's d = 0.2254; Bonferroni corrected α = 0.0125). The three-way interaction between retrieval status, distance, and time interval was not significant ($F_{3,108}$ = 0.703, p = 0.552, η_p^2 = 0.02). Bayes factor analysis revealed that a model without the three-way interaction term is preferred to a model with the three-way interaction by a factor of 13.1333. Together, these results suggest that although retrieval may be more successful on *near* compared to *far* trials, retrieval success does not influence the dissociation in retrieval evidence between *near* and *far* trials.

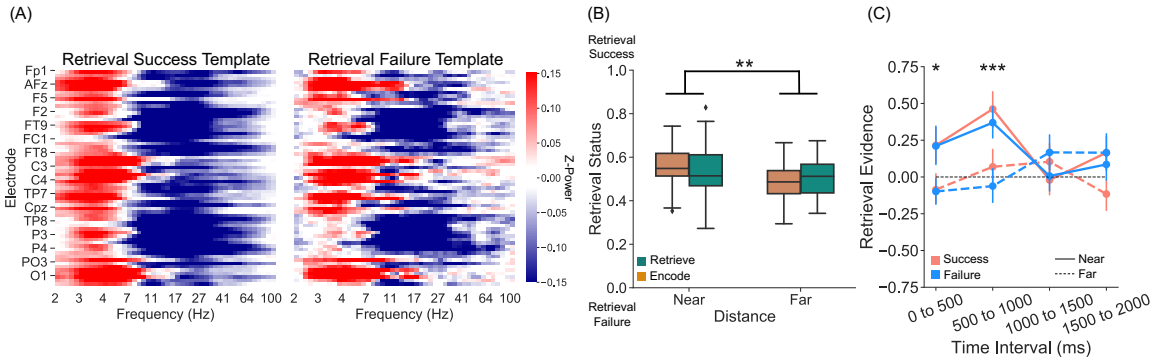

Figure 2.6: Impact of Retrieval Success on Retrieval State Evidence. (A) We derived retrieval success and retrieval failure templates from hit and miss trials during the recognition phase. Each panel shows an across-subject electrode-frequency spectrogram of z-power during retrieval success (hits; left) and retrieval failure (misses; right) in which red indicates z-power increases and blue indicates z-power decreases. **(B)** We assessed average retrieval status label as a function of instruction (encode, orange; retrieve, teal) and distance (*near*, *far*). We find a significant main effect of distance (p = 0.0018) driven by greater assignment of 'retrieval success' for *near* compared to *far* trials. **(C)** We assessed retrieval state evidence as a function of distance (*near*, solid; *far*, dashed) and retrieval status (success, red; failure, blue). We find a significant interaction between distance and time interval (p = 0.0010) driven by greater retrieval evidence on *near* compared to *far* trials early in the stimulus interval. Error bars denote SEM. * p < 0.05, ** p < 0.01, *** p < 0.001, uncorrected.

2.5 Discussion

Here we show that temporal contextual overlap between events selectively increases interference and induces automatic retrieval. We used scalp EEG to measure memory brain states in a task during which subjects were explicitly instructed

to either encode the present event or retrieve a past, overlapping event. We find behavioral evidence that temporal overlap selectively leads to interference for past events when the top-down goal is to encode the present event. We find neural evidence that temporal overlap induces automatic retrieval independent from top-down demands to encode or retrieve. Critically, our neural results suggest that the retrieval state we observe is likely the result of a general retrieval *mode* (Rugg & Wilding, 2000), rather than a reflection of retrieval success or effort. Collectively, these findings demonstrate a link between temporal context, interference, and memory brain states.

We find that greater temporal overlap between events leads to a selective memory deficit for a past event when the top-down demand is to encode the present event. Overlap between events can lead to both proactive interference, in which learning about a past event impairs memory for the present, and retroactive interference, in which learning about a present event impairs memory for the past (Underwood, 1948; Crowder, 1976). Here we find that greater temporal overlap between two events leads to an increase in retroactive interference; however, this increase is selective for conditions in which subjects' top-down goal is to encode the currently presented stimulus. This result has striking similarity with retrieval induced forgetting (Anderson et al., 1994; Anderson & Spellman, 1995). In paradigms that produce retrieval induced forgetting, subjects retrieve a target (e.g. strawberry) based on a word stem (e.g. s_____) and a cue (e.g. food) that is associated with other non-targets (e.g. tomato). Researchers theorize that cue driven retrieval of the non-target leads to suppression or inhibition which impairs subsequent memory for the non-target (*c.f.* Perfect et al., 2004). As the strength, typically framed in terms of semantic overlap, between non-target and cue increases, there is an in-

crease in memory impairment, putatively due to stronger inhibition (Anderson et al., 1994). We extend these findings by showing that temporal overlap can likewise impair memory for non-targets, suggesting that greater temporal overlap may lead to inhibition of automatically retrieved items that are not goal relevant.

Although in our study we find that temporal overlap is detrimental to later memory, there is evidence that temporal overlap between events can facilitate behavior. Participants are better at associative inference tasks when associated events are studied close together in time (Zeithamova & Preston, 2017). Events presented close together in time are often recalled together (temporally clustered, Kahana, 1996; Long & Kahana, 2015) and overall recall performance increases as more events are temporally clustered (Sederberg, Miller, Howard, & Kahana, 2010; Healey, Crutchley, & Kahana, 2014). Temporal overlap may promote the integration of two separate events (Schlichting & Preston, 2015; Richter, Chanales, & Kuhl, 2016), which enables those events to cue one another during a memory test. It is possible that in our study the explicit instruction to encode interrupts or prevents integration leading to worse memory for the past event. Follow-up studies investigating the influence of temporal overlap in the absence of explicit instructions to encode or retrieve are needed to test this possibility.

We find induction of a retrieval state early in the stimulus interval when objects are closer together in time. We anticipated that greater temporal overlap would lead to increased retrieval on the basis of retrieved context theory. According to retrieved context theory (Howard & Kahana, 2002; Sederberg et al., 2008; Polyn et al., 2009; Lohnas & Kahana, 2014), spatiotemporal context is bound to items during study and used as a retrieval cue during test (Long et al., 2017), enabling items with overlapping spatiotemporal contexts to cue retrieval of one another (Manning et

al., 2011). Consistent with retrieved context theory, we find more retrieval state evidence for objects with greater temporal overlap (*near* compared to *far* objects). Our observation of elevated retrieval state evidence on *near* trials even when the instruction is to encode the present (or, conversely, when the instruction is to *not* retrieve the past), suggests that the retrieval we observe is the result of a bottom-up or stimulus driven property of the object (e.g. its temporal contextual overlap with a past object) rather than the result of top-down or goal driven demands. The dissociation in retrieval evidence as a function of temporal overlap may reflect the engagement of an automatic retrieval process, given that automatic retrieval is thought to be a fast, bottom-up process that can occur without top-down control (Moscovitch, 1994). That the largest retrieval state evidence dissociations between *near* and *far* trials occur within the first 1000 ms following stimulus onset is consistent with this interpretation. Collectively, these findings indicate that memory brain states can be impacted by both bottom-up and top-down influences.

We interpret the retrieval state effects that we observe as reflecting a general retrieval *mode* rather than serial position effects, retrieval success, or retrieval effort. By design, *near* and *far* objects occurred in systematically distinct serial positions (primacy and recency, respectively). To address this inherent confound, we fit a trial-level GLM to the z-transformed spectral power and included serial position as a parametric regressor based on a logistic-regression model fit of the behavioral data. We used this approach to limit the potential contribution of serial position to the observed retrieval state effects as distinct neural signals are recruited across primacy and non-primacy positions (Sederberg et al., 2006). Given that the GLM cannot completely eliminate serial position effects, lingering primacy-related signals could contribute to the observed dissociation in retrieval evidence between

near and *far* trials. However, we note that as the pattern classifier is trained on data across all serial positions, we expect such a contribution to be limited.

The dissociation between *near* and *far* trials could be the result of other retrieval processes rather than a more general retrieval mode (Tulving, 1985; Rugg & Wilding, 2000). *Retrieval* as it stands is a broad concept and can encompass multiple different 'sub-processes.' We consider a retrieval state or mode as a content-independent process. Although typically retrieval mode has been considered within the framework of goal-directed or intentional remembering, we expect that a retrieval mode can also be engaged automatically based on bottom-up inputs (as demonstrated in the current study) and may align or be synonymous with the internal axis of attention (Chun, Golomb, & Turk-Browne, 2011). A retrieval mode is thought to be distinct from retrieval 'orientation' in which specific cues or features are used to guide memory (Herron & Wilding, 2004; Hornberger, Rugg, & Henson, 2006a, 2006b). Finally, both retrieval mode and orientation are separate from retrieval success and retrieval effort. After directing attention internally and orienting to particular cues to guide retrieval, an individual will either bring to mind the target item (retrieval success) or fail to bring to mind the target item, leading to effortful retrieval.

The retrieval process that we observe in the current study likely reflects a retrieval state given that a pattern classifier can distinguish encoding and retrieval across both *near* and *far* trials and that retrieval state evidence does not differ as a function of retrieval success. If the processes underlying *near* and *far* trials were entirely the product of retrieval success and retrieval effort, respectively, the cross-distance pattern classifier would be unable to distinguish encoding and retrieval across these trials. This is not to say that there are not potential differences in

terms of retrieval success or effort between *near* and *far* trials, only that there exist shared mechanisms which enables the pattern classifier to generalize across these trials. Although we cannot rule out the potential influence of retrieval effort, the interpretation that elevated retrieval evidence on *near* compared to *far* trials reflects *decreased* retrieval effort would be inconsistent with our findings of greater retrieval evidence on retrieve compared to encode trials, given that one would expect more retrieval effort for retrieve trials. By leveraging the recognition phase data, we indexed retrieval success across *near* and *far* trials and found that retrieval state evidence is modulated by distance, but not retrieval success. It is important to note that our index of retrieval success is more likely to capture recollection-based as opposed to familiarity-based retrieval processes, though given the strong categorical overlap between the object pairs, we would anticipate high levels of familiarity for all objects regardless of temporal overlap. As the content of retrieval varies on every trial, it is unlikely that retrieval orientation differs systematically across *near* and *far* trials. Thus the account best supported by these findings is that the dissociation in retrieval state evidence reflects a general retrieval mode. These results present an exciting avenue for future work to further dissociate these different retrieval sub-processes via multivariate methods and to more generally relate memory retrieval to internal attention.

Our results add to a growing body of work demonstrating the presence of neurally dissociable mnemonic states (Hasselmo et al., 2002; Hasselmo, 2005). Like other brain states (e.g. Kay & Frank, 2019), mnemonic states likely reflect sustained brain activity configurations. The shift between encoding and retrieval can occur on the order of milliseconds via theta oscillations which drive rapid shifts in entorhinal-hippocampal connectivity (Hasselmo et al., 2002). However, these states

may operate along slower timescales and be mediated by acetylcholine (Hasselmo & McGaughy, 2004; Meeter, Murre, & Talamini, 2004). Mnemonic states predict subsequent memory (Long & Kuhl, 2019), impact the cortical location of stimulus representations (Long & Kuhl, 2021), and can influence behavior and decision making (Duncan, Sadanand, & Davachi, 2012; Duncan & Shohamy, 2016; Patil & Duncan, 2018). Memory encoding and retrieval may reflect two states along a continuum within the broader framework of external and internal attention, respectively (Chun et al., 2011). Here we show that mnemonic states in the cortex persist for several hundred milliseconds and are influenced by bottom-up stimulus properties, in addition to explicit top-down demands. We expect that mnemonic states fluctuate based on both stimuli and goals – to the extent that events overlap, there is the potential for automatic retrieval and a shift into a retrieval state. Tracking mnemonic state fluctuations will be critical for understanding both how these states are induced and how these states in turn impact behavior.

In summary, we show that temporal overlap between events induces retrieval and selectively impairs memory performance. These findings are consistent with theoretical models which propose that temporal information can cue retrieval (Howard & Kahana, 2002) and behavioral findings that retrieving non-goal relevant information can lead to memory impairments (Anderson et al., 1994). More broadly, these findings point to a role for bottom-up stimulus features in driving mnemonic brain states.

Chapter 3

Response-locked theta dissociations reveal potential feedback signal following successful retrieval

Devyn E. Smith, Justin R. Wheelock & Nicole M. Long
Under Review

3.1 Abstract

Successful memory retrieval relies on memory processes to access an internal representation and decision processes to evaluate and respond to the accessed representation, both of which are supported by fluctuations in theta (4-8Hz) activity. However, the extent to which decision making processes are engaged following a memory response is unclear. Here, we recorded scalp electroencephalography

(EEG) while human participants performed a recognition memory task. We focused on response-locked data, allowing us to investigate the processes that occur prior to and following a memory response. We replicate previous work and find that prior to a memory response theta power is greater for identification of previously studied items (hits) relative to rejection of novel lures (correct rejections; CRs). Following the memory response, the theta power dissociation 'flips' whereby theta power is greater for CRs relative to hits. We find that the post-response 'flip' is more robust for hits that are committed quickly, potentially reflecting a positive feedback signal for strongly remembered experiences. Our findings suggest that there are potentially distinct processes occurring before and after a memory response that are modulated by successful memory retrieval.

3.2 Introduction

Successful remembering is dependent on both memory processes to access stored representations and decision processes to evaluate and respond to the accessed representation. How these processes unfold over time, and the underlying neural mechanisms, are critically important to memory success, yet our understanding of these processes remains limited. In particular, convergent evidence from scalp electroencephalography (EEG) studies in both the memory literature (Klimesch et al., 1997; Nyhus & Curran, 2010) and the decision making literature (Frank et al., 2015; Pinner & Cavanagh, 2017; Senftleben & Scherbaum, 2021) suggest that frontocentral theta (4-8 Hz activity in EEG) supports both memory and decision making processes. However, the extent to which theta dissociations in a memory task reflect decision making processes is unclear, likely due to limited investigation

of EEG signals *following* a memory response. The aim of this study is to investigate the neural correlates leading up to and following a memory response.

It is well established that successful remembering is characterized by electrophysiological changes around 300 to 800 ms following stimulus onset during a memory test (Friedman & Johnson Jr., 2000; Rugg & Wilding, 2000; Voss & Paller, 2008; Addante, Ranganath, & Yonelinas, 2012). Specifically, two event-related potentials (ERPs) distinguish successful remembering of a target or studied item (hit) from successful rejection of a lure or non-studied item (correct rejection, CR). The FN400, a negative going frontal ERP component thought to reflect familiarity (Rugg et al., 1998; Mecklinger, 2000; Curran & Hancock, 2007) is more negative for CRs than hits around 300 to 500 ms after stimulus onset (Curran, 2000; Curran & Cleary, 2003; Curran, 2004) and the LPC, a late positive going parietal ERP component thought to reflect recollection (Friedman & Johnson Jr., 2000; Mecklinger, 2000) is more positive for hits than CRs around 400 to 800 ms after stimulus onset (Curran, 2000; Friedman & Johnson Jr., 2000; Curran & Cleary, 2003; Curran, 2004). Similarly, theta power is greater for hits compared to CRs, most often around 500 to 1000 ms after stimulus onset (Burgess & Gruzelier, 1997; Klimesch, Doppelmayr, Schwaiger, Winkler, & Gruber, 2000; Düzel et al., 2003). According to the drift diffusion model, recognition memory is supported by a mechanism whereby evidence accumulates over time until a threshold is reached and a decision is made (Ratcliff & McKoon, 2008). Given evidence that theta power is positively correlated with reaction times, such that theta power increases until a response is made (Jacobs, Hwang, Curran, & Kahana, 2006), theta power prior to a response may reflect evidence accumulation and/or reinstatement (Nyhus & Curran, 2010; Herweg et al., 2020; Kota, Rugg, & Lega, 2020; Guan, Ma, Chen, Luo, & He, 2023). However,

these effects are related to representation access prior to a response and as such, do not elucidate the processes that may unfold after a memory response is made.

Parallel findings from the decision making and cognitive control literature have revealed that both ERPs and theta power track errors and negative feedback signals prior to and following a response (Luu, Tucker, & Makeig, 2004; Trujillo & Allen, 2007; Cavanagh, Frank, Klein, & Allen, 2010; Cavanagh & Frank, 2014; Luft, 2014). Specifically, the error-related negativity (ERN), a negative going fronto-central ERP component that reflects decision conflict or error monitoring (Frank, Woroch, & Curran, 2005; L. Wang, Gu, Zhao, & Chen, 2020), is more negative following incorrect compared to correct responses (Gehring, Goss, Coles, Meyer, & Donchin, 1993; Cavanagh, Zambrano-Vazquez, & Allen, 2012). Similarly, across cognitive control tasks such as the Stroop task, the flanker task, and go/no-go tasks, theta power is greater following incorrect compared to correct responses and following negative relative to positive outcomes (Mazaheri et al., 2009; Cohen, 2014b; Cavanagh & Frank, 2014). Together, these findings suggest that post-response theta power may reflect a feedback signal or monitoring process.

Although there is evidence for post-retrieval monitoring during recognition memory (Rugg, Henson, & Robb, 2003; Hill, Horne, Koen, & Rugg, 2021), these signals are often measured following access of the representation, but prior to the memory response itself. Further complicating interpretation is that reaction times are generally faster for hits than CRs (Uncapher, Boyd-Meredith, Chow, Rissman, & Wagner, 2015; Weidemann & Kahana, 2016), meaning that stimulus-locked hit vs. CR dissociations may include both pre- and post-response related processes. That is, insofar as an evaluative or updating decision making process is engaged after a memory response has been made, stimulus-locked comparisons may contrast post-

hit evaluative decision making processing with pre-CR memory retrieval related processing. There is ERP evidence that post-retrieval monitoring processes are supported by a late old/new effect over right frontal cortex (Wilding & Rugg, 1996; Johansson & Mecklinger, 2003), characterized by a positive voltage deflection that is greater for hits than CRs (Hayama, Johnson, & Rugg, 2008). However, the majority of such post-retrieval monitoring signals occur after the putative representation has been accessed, but before a behavioral response is made (Woodruff, Uncapher, & Rugg, 2006; Cruse & Wilding, 2009, 2011), leaving open the question of whether decision making mechanisms are engaged following a memory response. Greater conflict between memory decisions in a recognition task – created via differential payoff rates for correct old vs. new responses – leads to a greater post-response ERN (Curran, DeBuse, & Leynes, 2007), suggesting that control or monitoring processes may be engaged after a memory response is made.

Our hypothesis is that distinct processes occur prior to and following a memory response. To test our hypothesis, we conducted a human scalp EEG recognition memory study in which we specifically assessed response-locked theta power and provided no explicit feedback to participants. By investigating response-locked signals, we can separately assess pre- and post-response related processing during both hits and CRs. We expected to replicate prior work and find greater theta power for hits compared to CRs preceding the response (equivalent to the established stimulus-locked effects, e.g. Burgess & Gruzelier, 1997; Düzel et al., 2003; Nyhus & Curran, 2010). To the extent that distinct, and potentially decision making related, processes are engaged following a memory response, we expected to find a post-response theta pattern that differed from the pre-response hit vs. CR effect. First, if there are neither memory nor decision making related signals following a memory

response, theta power following both hits and CRs should return to baseline. Alternatively, because both hits and CRs are correct responses, theta power may be similarly decreased for both response types following a memory response. Finally, given that the reward system – in particular, the striatum – has previously been linked to successful memory retrieval, whereby the striatum shows greater activity for hits compared to CRs (Spaniol et al., 2009; Clos et al., 2015) successful retrieval may be intrinsically rewarding (Satterthwaite et al., 2012; Speer, Bhanji, & Delgado, 2014). Therefore, theta power may dissociate post-response hits from CRs such that theta power would be greater for CRs compared to hits following the memory response, reflecting a positive feedback signal for hits. The direct comparison of two classes of responses that are both accurate, but differ in terms of successful retrieval, enables adjudication between these alternative hypotheses.

3.3 Methods

Participants

Forty (30 female; age range = 18-42, mean age = 21.9 years) native English speakers from the University of Virginia community participated. Our sample size of N = 40 was selected based on prior scalp EEG studies conducted in our lab (D. E. Smith, Moore, & Long, 2022; Moore & Long, 2024). All participants had normal or corrected-to-normal vision. Informed consent was obtained in accordance with University of Virginia Institutional Review Board for Social and Behavioral Research and participants were compensated for their participation. Two participants were excluded from the final dataset: one for technical difficulties that

resulted in a subset of test items being presented twice during the test phase and one who failed to comply with task instructions. Thus data are reported for the remaining 38 participants. All raw, de-identified data and the associated experimental and analysis codes used in this study will be made available via the Long Term Memory Lab Website upon publication.

Recognition Task Experimental Design

Stimuli consisted of 1602 words, drawn from the Toronto Noun Pool (Friendly, Franklin, Hoffman, & Rubin, 1982). From this set, 288 words were randomly selected for each participant. Of these words, 192 were presented in both the study and test phase while the remaining 96 served as lures in the test phase.

Study Phase. In each of 12 runs, participants viewed a list containing 16 words, yielding a total of 192 trials. During each trial, participants saw a single word presented for 2000 ms followed by a 1000 ms inter-stimulus interval (ISI; Figure 3.1A). As we did not want to bias participants to the semantic features of the study items (Moore & Long, 2024), we did not include an encoding task. Instead, participants were instructed to study the presented word in anticipation for a later memory test and did not make any behavioral responses. An earlier motivation of this work was to investigate how semantic associations among study items influence memory formation mechanisms. To that end, each list was split evenly into two parts containing 8 words ("first associates" and "second associates," respectively) separated by a brief 2000 ms delay. Semantic association strength was determined using Word Association Space values (WAS; Nelson, Zhang, & McKinney, 2001); 'strong' semantic associates had a WAS value of 0.4 or greater and 'weak' semantic

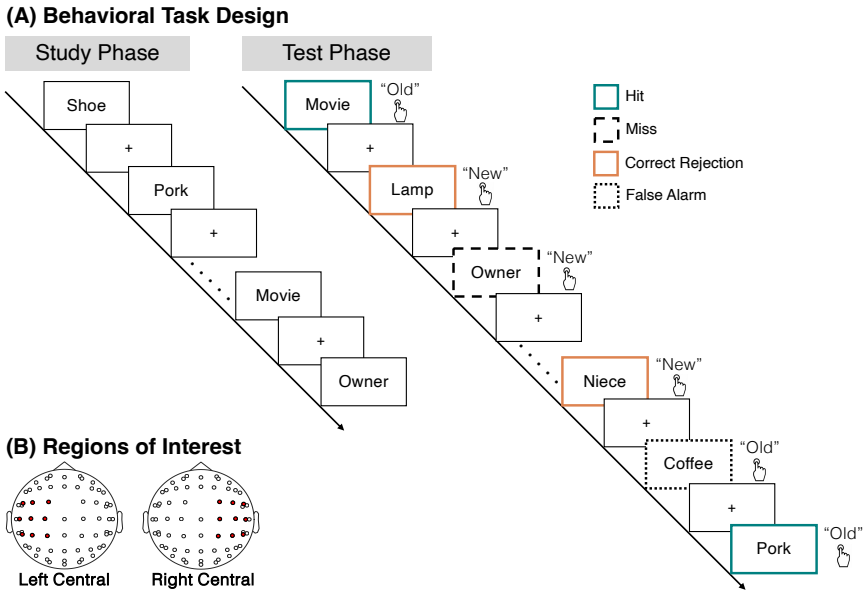

Figure 3.1: Task design. (A) During the study phase, participants studied individual words in anticipation of a later memory test and made no behavioral responses. After 12 runs of 16 item word lists, participants completed a recognition test phase. On each trial, participants saw either a target, a word that was presented during the study phase, or a lure, a word that was not presented during the study phase. Participants' task was to make an old or new judgement for each word. There were one of four possible response types, hits (teal; an 'old' response to a target), correct rejections (orange; a 'new' response to a lure), misses (dashed black lines; a 'new' response to a target), and false alarms (dotted black lines; an 'old' response to a lure). Lines and colors around the boxes are shown for illustrative purposes and were not present during the actual experiment. **(B)** We analyzed two regions of interest (ROIs), a left central ROI (FC5, FC1, C3, CP5, CP1, FC3, C1, C5, CP3) and a right central ROI (CP6, CP2, C4, FC6, FC2, CP4, C6, C2, FC4).

associates had a WAS value less than 0.4 (Long & Kahana, 2017). Half of the first and second associates were strongly semantically associated and half of the first and second associates were weakly semantically associated. Both strong and weak semantic associates were weakly semantically associated to all other study words. Word lists were generated for each participant by randomly drawing a word from the pool of 1602 words and selecting either a strong or weak associate from the word pool and then removing the selected word, selected associate, and all other strong semantic associates of the selected word, from the pool. We iteratively repeated this process until a total of 192 words were selected.

Test Phase. Following the 12 study runs, participants completed the recognition test phase. On each trial, participants viewed a word which had either been presented during the study phase (target) or had not been presented (lure; Figure 3.1A). Participants' task was to make an old or new judgment for each word by pressing one of two buttons ("d" or "k"). Response mappings were counterbalanced across participants. Test trials were self-paced and responses could occur anytime after the stimulus onset. Participants received no feedback on the accuracy of their responses. Test trials were separated by a 1000 ms ISI. There were a total of 288 test trials with all 192 study words presented along with 96 novel lures, half of which were semantically associated to a study word. As key findings are unchanged when accounting for semantic associations, we do not consider them further.

EEG Data Acquisition and Preprocessing

All acquisition and preprocessing methods are based on our previous work (D. E. Smith et al., 2022); for clarity we use the same text as previously reported. EEG recordings were collected using a BrainVision system and an ActiCap equipped with 64 Ag/AgCl active electrodes positioned according to the extended 10-20 system. All electrodes were digitized at a sampling rate of 1000 Hz and were referenced to electrode FCz. Offline, electrodes were later converted to an average reference. Impedances of all electrodes were kept below 50kΩ. Electrodes that demonstrated high impedance or poor contact with the scalp were excluded from the average reference. Bad electrodes were determined by voltage thresholding (see below).

Custom python codes were used to process the EEG data. We applied a high

pass filter at 0.1 Hz, followed by a notch filter at 60 Hz and harmonics of 60 Hz to each participant's raw EEG data. We then performed three preprocessing steps (Nolan et al., 2010) to identify electrodes with severe artifacts. First, we calculated the mean correlation between each electrode and all other electrodes as electrodes should be moderately correlated with other electrodes due to volume conduction. We z-scored these means across electrodes and rejected electrodes with z-scores less than -3. Second, we calculated the variance for each electrode, as electrodes with very high or low variance across a session are likely dominated by noise or have poor contact with the scalp. We then z-scored variance across electrodes and rejected electrodes with a $|z| >= 3$. Finally, we expect many electrical signals to be autocorrelated, but signals generated by the brain versus noise are likely to have different forms of autocorrelation. Therefore, we calculated the Hurst exponent, a measure of long-range autocorrelation, for each electrode and rejected electrodes with a $|z| >= 3$. Electrodes marked as bad by this procedure were excluded from the average re-reference. We then calculated the average voltage across all remaining electrodes at each time sample and re-referenced the data by subtracting the average voltage from the filtered EEG data. We used wavelet-enhanced independent component analysis (Castellanos & Makarov, 2006) to remove artifacts from eyeblinks and saccades.

EEG Data Analysis

We applied the Morlet wavelet transform (wave number 6) to the entire EEG time series across electrodes, for each of 46 logarithmically spaced frequencies (2-100 Hz; Long & Kahana, 2015). Because we hypothesized distinct processes occur prior to and following a memory response, after log-transforming the power we

focused exclusively on test-phase data. We then downsampled the test-phase data by taking a moving average across 100 ms time intervals from -1000 to 3000 ms relative to the response and sliding the window every 25 ms, resulting in 157 time intervals (40 non-overlapping). Mean and standard deviation power were calculated across all trials and across time points for each frequency. Power values were then z-transformed by subtracting the mean and dividing by the standard deviation power. We focus exclusively on the theta band (4-8 Hz) for all analyses.

Regions of Interest

We examined theta power across two regions of interest (ROIs; Figure 3.1B), left central (FC5, FC1, C3, CP5, CP1, FC3, C1, C5, CP3) and right central (CP6, CP2, C4, FC6, FC2, CP4, C6, C2, FC4). We specifically focus on the frontocentral region as prior work has demonstrated that theta power in these regions dissociates both hits and correct rejections (Burgess & Gruzelier, 1997; Klimesch et al., 1997; Gruber, Tsivilis, Giabbiconi, & Müller, 2008; Nyhus & Curran, 2010) and positive and negative feedback (Cohen, Elger, & Ranganath, 2007; Marco-Pallarés et al., 2008; Mas-Herrero, Ripollés, HajiHosseini, Rodríguez-Fornells, & Marco-Pallarés, 2015).

Univariate Analyses

To test the effect of response type on theta power leading up to and following memory responses, our two conditions of interest were hits (correctly recognized targets) and correct rejections (CRs, correctly rejected lures). We compared theta power across hits and CRs separately for each ROI. For each participant, we cal-

culated z-transformed theta power across both ROIs in each of the two conditions, across 100 ms time intervals from 500 ms pre-response to 1000 ms post-response. For a direct comparison of pre-response and post-response theta power, we further averaged z-transformed theta power over the 500 ms pre-response and 500 ms post-response interval separately for hits and CRs. We selected 500 ms as our pre-response interval based on our prior work investigating contextually mediated retrieval processes in the hippocampus (Long et al., 2017).

Peak Analysis

To measure the center frequency (CF) of theta leading up to and following memory responses, we used fitting oscillations & one over f (FOOOF; Ostlund et al., 2022). To specifically measure periodic signals, for each participant, we fit the FOOOF model to every test trial and extracted all identified CFs within the theta band prior to and following the response for both ROIs. We compared the averaged CF over the 500 ms pre-response and 500 ms post-response interval between hits and CRs.

Statistical Analyses

We used a repeated measures ANOVA (rmANOVA) to assess the distribution of reaction times (RTs) for hits and CRs. For post hoc comparisons across RTs, we used false discovery rate (FDR; $p = .05$) correction (Benjamini & Hochberg, 1995) to correct for multiple comparisons. We used rmANOVAs and paired-sample *t*-tests to assess the effect of response type (hits, CRs) and time interval (pre-response, post-response) on theta power.

3.4 Results

Our first goal was to measure memory discrimination (d') to ensure that participants were following directions and able to discriminate between targets and lures. For each participant, we calculated d' by subtracting the normalized false alarm rate (the percentage of lures that were incorrectly identified as 'old') from the normalized hit rate (the percentage of targets that were correctly identified as 'old'). The average d' was 1.75 (SD = 0.58; Figure 3.2A), indicating that participants were able to successfully distinguish targets from lures.

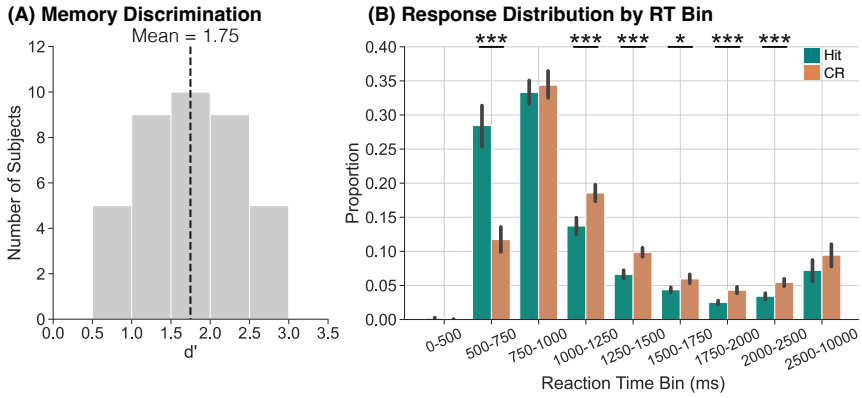

Figure 3.2: Memory discrimination and responses as a function of reaction time bin. **(A)** We used d' to assess memory discrimination. Participants were able to correctly discriminate between targets and lures. **(B)** We assessed the proportion of hits (teal) and CRs (orange) as a function of RT bin. The highest proportion of hits and CRs occurs in the 750-1000 ms bin and a significantly greater proportion of hits compared to CRs occur within the 500-750 ms bin. Error bars reflect standard error of the mean. *p < 0.05; ***p <.001 (FDR-corrected).

Having found that participants are able to discriminate targets and lures, we next assessed median reaction times (RTs) for hits and CRs. To the extent that RTs reliably differ between hits and CRs, stimulus-locked neural dissociations between these conditions may be driven by engagement of different processes, e.g. memory vs. decision making, rather than differential engagement of the same process. That is, if hits occur more quickly than CRs, a stimulus-locked comparison between the

two conditions could reflect a comparison between post-response processes for hits vs. pre-response processes for CRs. The average median RT for hits (M = 954.9, SD = 419.6) was significantly faster than for CRs (M = 1122.8, SD = 317.3, t_{36} = -4.427, p = 0.0001, d = 0.4514). This RT difference suggests that stimulus-locked theta dissociations could be driven by dissociations in pre- vs. post-response processes across hits and CRs.

Given the dissociation in median RT between hits and CRs, we next sought to compare the distribution of RTs across conditions to determine when relative to stimulus onset the majority of hits and CRs occur. We grouped RTs into nine bins selected to cover the full range of RTs with higher resolution in the faster (< 2500 ms) bins: 0-500 ms, 500-750 ms, 750-1000 ms, 1000-1250 ms, 1250-1500 ms, 1500-1750 ms, 1750-2000 ms, 2000-2500 ms, and 2500-10000 ms (Figure 3.2B). We calculated the proportion of hits and CRs that occurred within each RT bin for each participant and then averaged those proportions across all participants. We conducted a 2 × 9 rmANOVA with factors of response type (hit, CR) and RT bin. We do not find a significant main effect of response type ($F_{1,37}$ = 0.51, p = 0.479, η_p^2 = 0.01). We find a significant main effect of RT bin ($F_{8,296}$ = 74.48, $p < 0.0001$, η_p^2 = 0.67) and a significant interaction between response type and RT bin ($F_{8,296}$ = 22.36, $p <$ 0.0001, η_p^2 = 0.38). We report the results of post hoc t-tests comparing proportions of hits and CRs within each RT bin in Table 3.1 and highlight the key findings below. We find that the highest proportion of responses occurs in the 750-1000 ms RT bin for both hits and CRs. However, a significantly greater proportion of hits (M = 0.28, SD = 0.19) occur within the 500-750 ms RT bin compared to CRs (M = 0.12, SD = 0.11). Thus, if participants evaluate or update a representation after responding, neural activity observed within the 500-750 ms interval may reflect a comparison of

post-hit evaluation processes with pre-CR memory or retrieval related processes.

Table 3.1: Post hoc *t*-tests comparing the proportion of hits and CRs in each RT bin.

	Hits		CRs		Hits vs CRs		
RT Bin (ms)	Mean	SD	Mean	SD	t_{36}	p	d
0-500	0.002	0.006	0.0004	0.002	1.579	0.1229	0.3252
500-750	0.28	0.19	0.12	0.11	7.058	< **0.0001**	1.081
750-1000	0.33	0.11	0.34	0.13	-0.466	0.6438	0.0898
1000-1250	0.14	0.08	0.19	0.07	-4.198	**0.0002**	0.6536
1250-1500	0.07	0.04	0.10	0.04	-4.425	**0.0001**	0.7708
1500-1750	0.04	0.02	0.06	0.04	-2.549	**0.02**	0.4681
1750-2000	0.02	0.02	0.04	0.03	-4.012	**0.0003**	0.7581
2000-2500	0.03	0.03	0.05	0.03	-3.973	**0.0003**	0.6373
2500-10000	0.07	0.10	0.09	0.10	-2.009	0.052	0.2192

Note: bold values indicate tests that survive FDR correction.

Our hypothesis is that distinct memory and decision making processes occur preceding and following a memory response, meaning that we should find differential theta power engagement pre- and post-response. Specifically, we should replicate past findings of greater theta power for hits compared to CRs pre-response, reflecting memory related processing. To the extent that decision making processes are engaged following a response, we should either find decreased theta power for both hits and CRs – as both conditions are correct responses – or we may find a theta dissociation between hits and CRs. That is, to the extent that successful retrieval is intrinsically rewarding, post-response theta power should be decreased for hits compared to CRs. We conducted a $2 \times 2 \times 15$ rmANOVA with factors of response type (hit, CR), ROI (left central; LC, right central; RC) and time interval (-500 to 1000 ms in fifteen 100 ms intervals). We report the results of this ANOVA in Table 3.2 and highlight the key findings here. We find a significant interaction between response type and time interval ($F_{14,518} = 6.127$, $p < 0.0001$, $\eta_p^2 = 0.14$) which indicates that theta power dissociations between hits and CRs changes over time.

Table 3.2: Analysis of variance results for response type (hit, CR), ROI, and time interval (-500 to 1000 ms in fifteen 100 ms intervals) on theta power.

Effect	df	F	p	${\eta_p}^2$
Main effect of response type	(1,37)	0.172	0.681	0.005
Main effect of ROI	(1,37)	11.36	**0.002**	0.23
Main effect of time interval	(14,518)	21.59	< **0.0001**	0.37
Interaction of response type × ROI	(1,37)	0.218	0.643	0.006
Interaction of ROI × time interval	(14,518)	3.408	< **0.0001**	0.08
Interaction of response type × time interval	(14,518)	6.127	< **0.0001**	0.14
Interaction of response type × ROI × time interval	(14,518)	2.23	**0.006**	0.06

Note: bold values indicate p < 0.05.

Given the significant three-way interaction between response type, ROI, and time interval, we next performed follow-up post-hoc ANOVAs over time separately for each ROI. We conducted two 2 × 15 rmANOVAs with factors of response type (hit, CR) and time interval (-500 to 1000 ms in fifteen 100 ms intervals). We report the results of this ANOVA in Table 3.3 and highlight the key findings here. For both ROIs (Figure 3.3), we find a significant interaction between response type and time interval (LC: $F_{14,518} = 5.373$, $p < 0.0001$, $\eta_p^2 = 0.13$; RC: $F_{14,518} = 3.791$, $p < 0.0001$, $\eta_p^2 = 0.09$). These results demonstrate that theta power dissociations between hits and CRs vary across time interval, suggesting that differential processes may be engaged pre- vs. post-response that distinguish these response types.

Table 3.3: Analysis of variance results for response type (hit, CR) and time interval (-500 to 1000 ms in fifteen 100 ms intervals; pre-response, post-response) on theta power for the left and right central ROI.

		-500 to 1000 ms interval							Pre vs. post response interval					
		Left Central			Right Central				Left Central			Right Central		
Effect	df	F	p	${\eta_p}^2$	F	p	${\eta_p}^2$	df	F	p	${\eta_p}^2$	F	p	${\eta_p}^2$
Main effect of response type	(1,37)	0.382	0.54	0.01	0.007	0.934	0.0002	(1,37)	0.009	0.925	0.0002	0.871	0.357	0.02
Main effect of time interval	(14,518)	19.76	< **0.0001**	0.35	20.91	< **0.0001**	0.36	(1,37)	5.832	**0.0208**	0.14	0.708	0.406	0.02
Interaction of response type × time interval	(14,518)	5.373	< **0.0001**	0.13	3.791	< **0.0001**	0.09	(1,37)	16.18	**0.0003**	0.30	5.306	**0.027**	0.13

Note: bold values indicate p < 0.05.

To specifically test for a pre- vs. post-response dissociation in theta power for

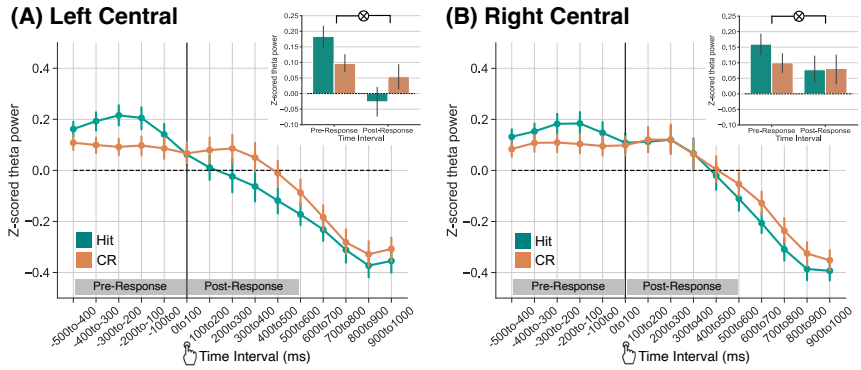

Figure 3.3: Theta power dissociations across hits and correct rejections preceding and following memory responses. Response-locked z-transformed theta power (4-8 Hz) for the left and right central ROIs. The solid vertical black line indicates when the response was made. Hits are shown in teal, correct rejections (CRs) are shown in orange. Error bars reflect standard error of the mean. **(A)** Over the left central ROI, we find a significant interaction between response type and time interval ($p < 0.001$) driven by greater pre-response theta power for hits than CRs and numerically greater post-response theta power for CRs than hits. **(B)** Over the right central ROI, we find a significant interaction between response type and time interval ($p = 0.027$) driven by greater theta power for hits than CRs during the pre-response time interval.

hits and CRs, we averaged signals within the 500 ms pre- and post-response time intervals (Figure 3.3, insets). We conducted two 2 × 2 rmANOVAs, one for each ROI, with factors of response type (hit, CR) and time interval (pre-response, post-response). We report the results of this ANOVA in Table 3.3 and highlight the key findings here. In both ROIs, we find a significant interaction between response type and time interval. In LC, this interaction was driven by greater theta power for hits (M = 0.18, SD = 0.21) relative to CRs (M = 0.10, SD = 0.18) in the pre-response time interval ($t_{36} = 2.559$, $p = 0.0147$, $d = 0.4459$) and numerically greater theta power for CRs (M = 0.05, SD = 0.25) relative to hits (M = -0.03, SD = 0.29) in the post-response time interval ($t_{36} = 2.021$, $p = 0.0506$, $d = 0.3019$). In RC, this interaction was driven by numerically greater theta power for hits (M = 0.16, SD = 0.21) relative to CRs (M = 0.10, SD = 0.19) in the pre-response time interval ($t_{36} = 1.911$, $p = 0.0637$, $d = 0.295$) and no difference in theta power for hits (M = 0.08, SD = 0.24) relative to CRs (M = 0.08, SD = 0.29) in the post-response time interval ($t_{36} = -0.1160$, p

= 0.9083, d = 0.015). We next tested the hemispheric specificity of this effect by directly comparing the post-response theta power dissociation across ROIs. We computed post-response difference scores (CRs minus hits) for each ROI. We find that the post-response theta power dissociation in LC (M = 0.081, SD = 0.25) does not significantly differ from the post-response theta power dissociation in RC (M = 0.004, SD = 0.21; t_{36} = 1.605, p = 0.1171) indicating that the effect is not specific to the left hemisphere. Together, these results indicate that theta power is modulated by successful retrieval leading up to and following memory responses.

As the theta band encompasses multiple frequencies (4-8Hz), we next tested the extent to which the center frequency of theta differed across response type and time interval. To specifically measure periodic signals, we fit the fitting oscillations & one over f (FOOOF) model to every test trial and extracted all identified center frequencies within the theta band prior to and following the response for the left and right central ROI. We compared the center frequency between hits and CRs pre- and post-response. We do not find any differences in center frequency for either ROI or time interval (LC, pre-response: t_{36} = -0.0976, p = 0.9228; LC, post-response: t_{36} = 0.2201, p = 0.8270; RC, pre-response: t_{36} = -0.3237, p = 0.748; RC, post-response t_{36} = -1.072, p = 0.2908). The center frequency for all conditions is generally around 5.6Hz.

Prior work (Herweg et al., 2020; Guan et al., 2023) suggests that the pre-response theta power dissociation that we observe specifically reflects evidence accumulation or reinstatement that ultimately supports recollection. The post-response theta power effects may likewise reflect a feedback signal in response to recollected content. Due to the current task design we cannot directly measure recollection and familiarity; however, we can leverage reaction time (RT) as a coarse assay of con-

fidence. The general assumption is that compared to trials with slow RTs, trials with fast RTs reflect faster evidence accumulation (Mulder & van Maanen, 2013; Shenhav, Straccia, Musslick, Cohen, & Botvinick, 2018; Rollwage et al., 2020) and greater confidence (Ratcliff, 1978; Ratcliff & Starns, 2009; Weidemann & Kahana, 2016). To divide the trials based on RT, we calculated the median RT across all correct responses (hits and CRs) for each participant, and labeled hits as either 'fast' (those below the median RT) or 'slow' (those above the median RT). To the extent that fast hits reflect strongly remembered experiences or the degree of evidence accumulation, we should find differential theta power engagement pre- and post-response. Specifically, we should find greater pre-response theta power for fast hits compared to CRs, as there is no experience to remember or reinstate during a CR. To the extent that post-response theta power reflects a feedback signal based on reinstated content, we should find decreased theta power for fast hits compared to CRs.

To specifically test for a pre- vs. post-response dissociation in theta power for fast hits, slow hits, and CRs, we averaged signals within the 500 ms pre- and post-response time intervals (Figure 3.4). We conducted two 2×3 rmANOVAs, one for each ROI, with factors of time interval (pre-response, post-response) and response type (fast hit, slow hit, CR). We report the results of this ANOVA in Table 3.4 and highlight the key findings here. In LC, we find a significant interaction between response type and time interval. This interaction was driven by a significant interaction between both fast hits and CRs ($F_{1,37} = 24.56$, $p < 0.0001$, $\eta_p^2 = 0.40$) and fast hits and slow hits ($F_{1,37} = 15.02$, $p = 0.0004$, $\eta_p^2 = 0.29$). The interaction between slow hits and CRs was not significant ($F_{1,37} = 2.892$, $p = 0.0974$, $\eta_p^2 = 0.07$). In the pre-response interval, theta power was significantly greater for fast hits (M = 0.22,

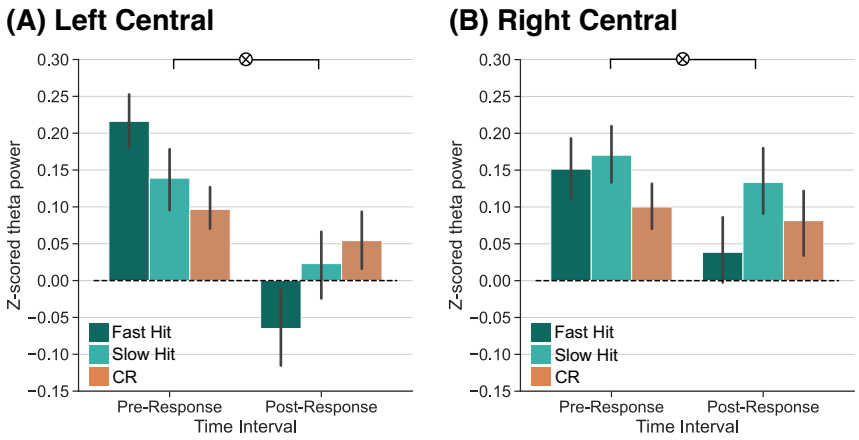

Figure 3.4: Theta power dissociations across fast hits, slow hits, and correct rejections preceding and following memory responses. Response-locked z-transformed theta power (4-8 Hz) for the left and right central ROIs. Fast hits are shown in dark teal, slow hits are shown in light teal and correct rejections (CRs) are shown in orange. Error bars reflect standard error of the mean. **(A)** Over the left central ROI, we find a significant interaction between response type and time interval ($p < 0.0001$) driven by a significant pre-post interaction between fast hits and CRs ($p < 0.0001$). **(B)** Over the right central ROI, we find a significant interaction between response type and time interval ($p = 0.0206$) driven by a significant pre-post interaction between fast hits and CRs ($p = 0.008$).

SD = 0.23) relative to CRs (M = 0.10, SD = 0.18; $t_{36} = 3.080$, $p = 0.0039$, $d = 0.5789$) and numerically greater for fast hits relative to slow hits (M = 0.14, SD = 0.25; $t_{36} = 1.908$, $p = 0.0642$, $d = 0.3256$). In the post-response interval, theta power was significantly greater for CRs (M = 0.05, SD = 0.25) and slow hits (M = 0.02, SD = 0.28) relative to fast hits (M = -0.07, SD = 0.33; CRs vs. fast hits: $t_{36} = 2.554$, $p = 0.0149$, $d = 0.4161$; slow hits vs. fast hits: $t_{36} = 2.606$, $p = 0.0131$, $d = 0.2918$).

In RC, we find a significant interaction between response type and time interval ($F_{2,74} = 4.095$, $p = 0.0206$, $\eta_p^2 = 0.10$). This interaction was driven by a significant interaction between fast hits and CRs ($F_{1,37} = 7.885$, $p = 0.008$, $\eta_p^2 = 0.18$), whereby theta power was numerically greater for fast hits (M = 0.15, SD = 0.26) relative to CRs (M = 0.10, SD = 0.19) in the pre-response time interval ($t_{36} = 1.274$, $p = 0.2107$, $d = 0.2278$) and numerically greater for CRs (M = 0.08, SD = 0.29) relative to fast hits (M = 0.04, SD = 0.27) in the post-response time interval ($t_{36} = 0.9688$, $p =$

0.3389, $d = 0.1549$). The interaction between slow hits and CRs was not significant ($F_{1,37} = 0.306$, $p = 0.583$, $\eta_p^2 = 0.008$) nor was the interaction between fast hits and slow hits ($F_{1,37} = 4.038$, $p = 0.0518$, $\eta_p^2 = 0.010$). Together, these results suggest that post-response theta power dissociations may reflect a positive feedback signal in response to strongly remembered experiences.

Table 3.4: Analysis of variance results for response type (fast hit, slow hit, CR) and time interval (pre-response, post-response) on theta power for the left and right central ROI.

		Left Central			Right Central		
Effect	df	F	p	$\eta_p{}^2$	F	p	$\eta_p{}^2$
Main effect of response type	(2,74)	0.019	0.981	0.0005	1.938	0.151	0.05
Main effect of time interval	(1,37)	7.459	**0.0096**	0.17	0.87	0.357	0.02
Interaction of response type × time interval	(2,74)	14.93	< **0.0001**	0.29	4.095	**0.0206**	0.10

Note: bold values indicate p < 0.05.

3.5 Discussion

The goal of the current study was to test the hypothesis that distinct processes occur prior to and following a memory response. We recorded scalp EEG while participants performed a recognition memory task and received no explicit feedback on their performance. Crucially, we focused our analyses on response-locked test-phase data to separate the processes occurring before and after a memory response. We show that reaction times (RTs) are faster for hits compared to correct rejections (CRs) and that a greater proportion of hits occur within 500-750 ms of stimulus onset compared to CRs. We replicate established findings (Nyhus & Curran, 2010) that preceding a memory response, theta power (4-8Hz) in frontocentral electrodes is greater for hits compared to CRs. We show that these pre-response theta power dissociations 'flip' in left central electrodes following the memory response. We

find that this post-response 'flip' is specific to hits committed quickly, potentially reflecting a positive feedback signal for strongly remembered experiences. Together, these findings suggest that there are potentially distinct memory and decision making processes engaged preceding and following a memory response that are modulated by successful retrieval.

We find faster RTs for hits compared to CRs, replicating past findings (Uncapher et al., 2015; Weidemann & Kahana, 2016). Faster RTs for hits may be driven by greater memory strength (Verde & Rotello, 2007; Wixted, 2007) and/or greater contextual reinstatement (Gordon et al., 2014; Hanczakowski, Zawadzka, & Macken, 2015). However, these RT differences indicate that traditional hit vs. CR comparisons of stimulus-locked data may capture both pre- and post-response related processes. We specifically find that a larger proportion of hits occur within 500-750 ms of stimulus onset compared to CRs meaning that if participants evaluate or update a representation after responding, and they respond at different times for hits relative to CRs, estimates of neural signals within this time window may reveal differences in post-hit evaluation processes following successful retrieval vs. memory/retrieval related processes related to correctly rejecting lures. Thus, our results highlight the importance of utilizing response-locked data to investigate the distinct processes leading up to and following a memory response.

Consistent with prior EEG work (Burgess & Gruzelier, 1997; Klimesch et al., 1997, 2000; Düzel et al., 2003; Nyhus & Curran, 2010), we find greater pre-response theta power for hits than CRs. EEG power in the theta frequency band has been proposed to coordinate cortical areas, supporting the ability to encode and retrieve contextual information about time and space that is central to episodic memory (Hasselmo & Stern, 2014). Successful memory retrieval depends on the access of

an internal representation of a past experience. This access may take the form of reinstatement, wherein encoded content is reconstructed during retrieval (Danker & Anderson, 2010), which may specifically be supported by coordination between cortical areas (Herweg et al., 2020). Our findings are generally consistent with the drift diffusion model (Ratcliff & McKoon, 2008; Ratcliff, Smith, Brown, & McKoon, 2016) whereby the pre-response theta power dissociations may reflect evidence accumulation. Insofar as pre-response theta power reflects reinstatement of a past experience (Kota et al., 2020; Guan et al., 2023), our finding of generally lower pre-response theta power for CRs is consistent with the evidence accumulation account as there is no experience to reinstate during a CR. Although we did not fit the drift diffusion model to the neural data, we would predict that greater theta power preceding a response would be associated with a higher or faster drift rate, the model parameter that reflects the strength of the rate of evidence accumulation (Ratcliff et al., 2016). This interpretation is consistent with our finding of robust pre-response theta power dissociations specifically between slow hits and fast hits. Fast RTs are thought to reflect rapid evidence accumulation (Mulder & van Maanen, 2013; Shenhav et al., 2018; Rollwage et al., 2020), thus greater pre-response theta power for fast hits compared to slow hits may reflect faster evidence accumulation for strongly remembered experiences.

Following a memory response, we find a 'flip' in theta power over the left central ROI, such that theta power is greater for CRs than hits. Participants did not receive feedback indicating that they were correct suggesting that the dissociation in theta power may occur in response to intrinsic feedback signals specifically following successful retrieval. Although prior work has demonstrated that post-retrieval monitoring processes are engaged following memory retrieval (Rugg et

al., 2003; Hill et al., 2021), the majority of these findings reflect signals that precede a behavioral response. Our interpretation is that the post-response theta power dissociation between hits and CRs may reflect a feedback signal. This interpretation is consistent with work from the cognitive control literature showing that theta power increases for incorrect compared to correct responses and following negative relative to positive outcomes (Cavanagh et al., 2010; Cavanagh & Frank, 2014). Prior work has proposed that frontal midline theta (FMT) is associated with reward processing – specifically that the FMT is larger following negative feedback or monetary loss (Cohen et al., 2007; Marco-Pallarés et al., 2008) – and is a mechanism for communication between brain regions (Glazer, Kelley, Pornpattananangkul, Mittal, & Nusslock, 2018). As both hits and CRs constitute correct responses, we may have anticipated decreased post-response theta power for both response types. However, only hits reflect successful retrieval. Thus, the post-response decrease in theta power for hits may reflect positive feedback specifically in response to successful retrieval.

We find that the pre- vs. post-response dissociation in theta power is specific to fast hits. To the extent that fast hits reflect highly confident responses (Ratcliff, 1978; Ratcliff & Starns, 2009; Weidemann & Kahana, 2016), the post-response theta power decrease for fast hits may reflect a positive feedback signal. Neuro-imaging work has repeatedly shown that reward-related regions (e.g. striatum) are more active during hits compared to CRs in the absence of explicit reward (Achim & Lepage, 2005; de Zubicaray, McMahon, Eastburn, Finnigan, & Humphreys, 2005; Henson, Hornberger, & Rugg, 2005; Fliessbach et al., 2006; Spaniol et al., 2009; Schwarze, Bingel, Badre, & Sommer, 2013; Clos et al., 2015). As both responses are correct, the dissociation in striatal activity for hits compared to CRs suggests

that the signal change is not driven by overall accuracy, but rather is a response to successful retrieval and indicates that successful retrieval may be intrinsically rewarding (Satterthwaite et al., 2012; Speer et al., 2014). Taken together, the post-response theta dissociation that we observe in the current study may represent a positive feedback signal in response to successful retrieval, though a direct test of this account is necessary to support this claim. The direct investigation of post-response test-phase feedback signals presents an exciting avenue for future work.

Our broad interpretation of the present findings is that the post-response theta power decrease reflects the last process in a cascade of processes that support "remembering," generally construed. An individual begins remembering by perceptually processing an externally presented stimulus and then engages in a memory search, internal attention, evidence accumulation, and/or matching process in an attempt to access a stored representation (Polyn & Kahana, 2007; Kahana, 2012; Ratcliff et al., 2016; D. E. Smith & Long, 2024) – these two processes need not be serial and instead an individual may iterate between them. If these processes are successful, an item – and possibly its context or other associated information – may be reinstated, as in Tulving's original proposal of ecphory (Tulving, 1983). Reinstatement is followed by a monitoring process in which the accessed representation is evaluated (Cruse & Wilding, 2009). A decision is then made in tandem with commission of a behavioral response. Finally, a decision making post-response evaluative or representation updating process is engaged. Our interpretation is that the frontocentral post-response theta dissociation reflects this final decision making step. Given the limitations in estimating reinstatement of verbal, as opposed to visual, stimuli (Brunec, Robin, Olsen, Moscovitch, & Barense, 2020), future work is needed to directly test this account and the timing and relationship

between reinstatement and post-decision signals.

A critical open question is how the observed RT distributions and pre- vs. post-response theta dissociations relate to the established processes of recollection and familiarity. Recollection is the retrieval of contextual details and familiarity is memory strength without detailed retrieval (A. P. Yonelinas, 2001a; Diana, Vilberg, & Reder, 2005; Gimbel & Brewer, 2011; Addante et al., 2012). There is mixed evidence as regards RTs for recollection and familiarity, with some evidence that recollection responses are faster than familiarity responses (Diana et al., 2005; Gimbel & Brewer, 2011; Herweg et al., 2016) and some evidence that recollection responses are slower than familiarity responses (Atkinson & Juola, 1974; Jacoby, 1991; Besson, Ceccaldi, Didic, & Barbeau, 2012). Due to our task design, we cannot disambiguate recollection from familiarity based responses, but we can leverage RT as a coarse assay of confidence. Our assumption is that compared to slow hits, fast hits reflect higher confident responses (Ratcliff, 1978; Ratcliff & Starns, 2009; Weidemann & Kahana, 2016). High confident responses may be supported by the recollection of contextual details (A. P. Yonelinas, 2001a, 2001b), in which case elevated pre-response theta power may reflect the reinstatement of contextual details during fast hits. Likewise, the decreased post-response theta power following fast hits may reflect a positive feedback signal in response specifically to recollected experiences. Future work will be needed to directly test this possibility.

Although we did not anticipate hemispheric differences, we consistently found influences of ROI on theta power. Overall, the effects that we observe are numerically stronger in the left central, relative to right central, ROI, although a direct test of the post-response theta dissociation for left vs. right ROIs indicated that the effect is not specific to the left hemisphere. It is unlikely that motor responses

contributed to the observed hemispheric effects given that responses were counterbalanced across participants and typically motor movements engage higher frequency bands (e.g. Crone et al., 1998). The hemispheric asymmetry may be driven by the stimuli used and/or intrinsic hemispheric connections (D. Wang, Buckner, & Liu, 2014). We used visually presented words in the current study and verbal stimuli are well known to recruit the left hemisphere (de Zubicaray, Miozzo, Johnson, Schiller, & McMahon, 2011; Vigneau et al., 2011; Price, 2012; Ries, Dronkers, & Knight, 2016), including during memory tasks (Kelley et al., 1998; H. Kim, 2011). Investigation of resting state data has shown that cortical networks have intrinsic within-hemisphere connections which may enable control over the specific functions or processes that are engaged (D. Wang et al., 2014). Future work will be needed to probe both the hemispheric asymmetry of this effect as well as the overall spatial specificity, given that we chose to focus exclusively on frontocentral regions given the convergence of memory and decision making literature on this area; however, other areas are known to also support memory retrieval (e.g. right frontal, Evans, Williams, & Wilding, 2015; left parietal, Jacobs et al., 2006). Our findings generally show a frontocentral theta pattern consistent with pre-response memory-related processing and post-response decision making related processing.

Together, our findings suggest that distinct processes occur prior to and following a memory response and, in particular, that decision making processes may follow successful retrieval. A direction for future research will be to directly investigate the extent to which the post-response theta dissociation reflects a feedback signal. More broadly, we contribute to a growing body of literature characterizing the role of theta activity in successful memory retrieval.

Chapter 4

Successful retrieval is followed by an intrinsic reward signal

Devyn E. Smith & Nicole M. Long

4.1 Abstract

Fluctuations in theta (4-8Hz) activity supports both successful retrieval and feedback-based learning. However, the extent to which theta power dissociations reflect a feedback signal in response to successful retrieval is unknown. Here, we recorded scalp electroencephalography (EEG) while human participants performed between-subjects recognition memory tasks in which we manipulated test phase goals. We replicate prior work and find that following a memory response, theta power decreases selectively for identification of previously studied items

(hits) relative to rejection of novel lures (correct rejections; CRs), regardless of task goals. We used an independently validated feedback classifier to measure positive feedback evidence as a function of task goals, responses (hits vs. CRs), and time. We find greater positive feedback evidence for hits following a response, regardless of the task goals. Together, these results suggest that successful retrieval is intrinsically rewarding.

4.2 Introduction

Theta power (4-8 Hz) in scalp electroencephalography (EEG) has been shown to dissociate both successful retrieval (Burgess & Gruzelier, 1997; Nyhus & Curran, 2010; D. E. Smith, Wheelock, & Long, 2024) and track feedback or outcomes across cognitive control tasks (Cavanagh & Frank, 2014). However, the extent to which theta power dissociations reflect a feedback signal in response to successful retrieval is unknown. Successful retrieval may be intrinsically rewarding (Speer et al., 2014) as evidenced by greater reward-related region (e.g. striatum) activity during successful item retrieval (hits) compared to correct identification of a new stimulus (correct rejections; CRs) in the absence of extrinsic reward (Spaniol et al., 2009; Clos et al., 2015). Alternatively, test phase reward signals may reflect goal attainment (Han, Huettel, Raposo, Adcock, & Dobbins, 2010) as in a typical recognition experiment, participants are asked to recognize old items, such that hits reflect goal attainment. The aim of this study was to identify whether test phase theta signals reflect intrinsic reward or goal attainment.

Theta has been shown to support both episodic memory retrieval (Nyhus & Curran, 2010) and feedback-based learning (Cohen, 2014a). Specifically, following

stimulus onset, theta power is greater for hits compared to CRs (Burgess & Gruzelier, 1997; Klimesch et al., 2000). Following successful retrieval, this dissociation in theta power 'flips' such that theta power is greater for CRs than hits (D. E. Smith et al., 2024). Given evidence that theta power increases following incorrect relative to correct responses and negative relative to positive outcomes (Mazaheri et al., 2009; Cavanagh & Frank, 2014), the post-response theta power dissociation between hits and CRs may reflect a feedback signal. Furthermore, the error-related negativity (ERN), an intrinsic feedback signal, reflects comparison processes or conflict monitoring (L. Wang et al., 2020) that arises from ongoing theta activity (Trujillo & Allen, 2007). Taken together, these findings suggest that post-response theta power may reflect an intrinsic feedback signal or monitoring process in response to successful retrieval.

Parallel findings from neuro-imaging work have repeatedly shown that reward-related regions are active during successful item retrieval (Achim & Lepage, 2005; de Zubicaray et al., 2005; Spaniol et al., 2009; Clos et al., 2015). Specifically, the striatum shows greater activity when a subject correctly recognizes an old stimulus compared to when the subject correctly rejects a new stimulus. As both hits and CRs constitute accurate trials, the dissociation in striatal activity for hits compared to CRs suggests that the signal change is not driven by dissociations between correct and incorrect trials, but rather is a response to successful retrieval. As this striatal response occurs absent of explicit rewards (e.g. monetary incentives or positive feedback), it may signal that successful retrieval is intrinsically rewarding (Satterthwaite et al., 2012).

Reward signals following successful retrieval may alternatively reflect the attainment of a task goal. Evidence from a recognition study in which participants

were explicitly rewarded for hits or CRs suggests that test phase reward signals may be driven by the task goal rather than successful retrieval (Han et al., 2010). When participants had the potential to earn a reward for CRs, striatal activity was greater in response to CRs than hits, suggesting that reward signals reflect goal attainment. However, presenting an extrinsic reward (e.g. money) changes a task from potentially intrinsically rewarding to extrinsically rewarding (Hidi, 2016), meaning that the introduction of explicit rewards for CRs may alter the natural structure whereby hits are intrinsically rewarding. Therefore it is unclear whether test phase reward signals are similarly modulated by goals in the absence of extrinsic rewards.

Our hypothesis is that post-response test phase reward signals reflect successful retrieval. Alternatively, test phase reward signals may reflect goal attainment (Han et al., 2010). In a typical recognition experiment, the participants' goal to identify old items is confounded with successful retrieval. To adjudicate between these hypotheses, we conducted two recognition memory experiments (E1, E2) in which we manipulated test phase goals. Participants goal was to either successfully retrieve study items (E1) or detect new items (E2). We recorded scalp EEG to separately measure memory vs. feedback related theta signals. Our expectation is that the task instructions will impact the memory and decision processes in which participants engage when responding to stimuli. Prior work (Brainerd, Bialer, Chang, & Upadhyay, 2021) provides two hypotheses for how task goals can impact behavior, encoding specificity and the Fuzzy Trace Theory (FTT) noncompensatory gist principle. According to the encoding specificity account, memory decisions are made based on the match between the task goal (e.g. recognize old) and the probe (e.g. target) whereas the FTT noncompensatory gist principle suggests that general

or gist level information is used to reject items. We used cross-study classification to measure test phase feedback evidence. To the extent that feedback signals reflect successful retrieval, the same post-response signals should dissociate hits and CRs regardless of memory goals.

4.3 Methods

Participants

Seventy six native English speakers from the University of Virginia community participated, with thirty eight participants enrolled in each experiment (E1: 28 female; age range = 18-32, mean age = 20.47 years; E2: 26 female; age range = 18-32, mean age = 20.5 years). All participants had normal or corrected-to-normal vision. Informed consent was obtained in accordance with University of Virginia Institutional Review Board for Social and Behavioral Research and participants were compensated for their participation. Our sample size was determined *a priori* based on behavioral pilot data (E1, N = 5; E2, N = 3) described in the pre-registration report of this study (https://osf.io/tfq9u). A total of four participants (two each from E1 and E2) were excluded from the final dataset due to EEG event markers not being recorded. Thus data are reported for the remaining seventy two participants. All raw, de-identified data and the associated experimental and analysis codes used in this study will be made available via the Long Term Memory Lab Website upon publication. These data have previously been reported (D. E. Smith & Long, 2024); all of the analyses and results described here are novel.

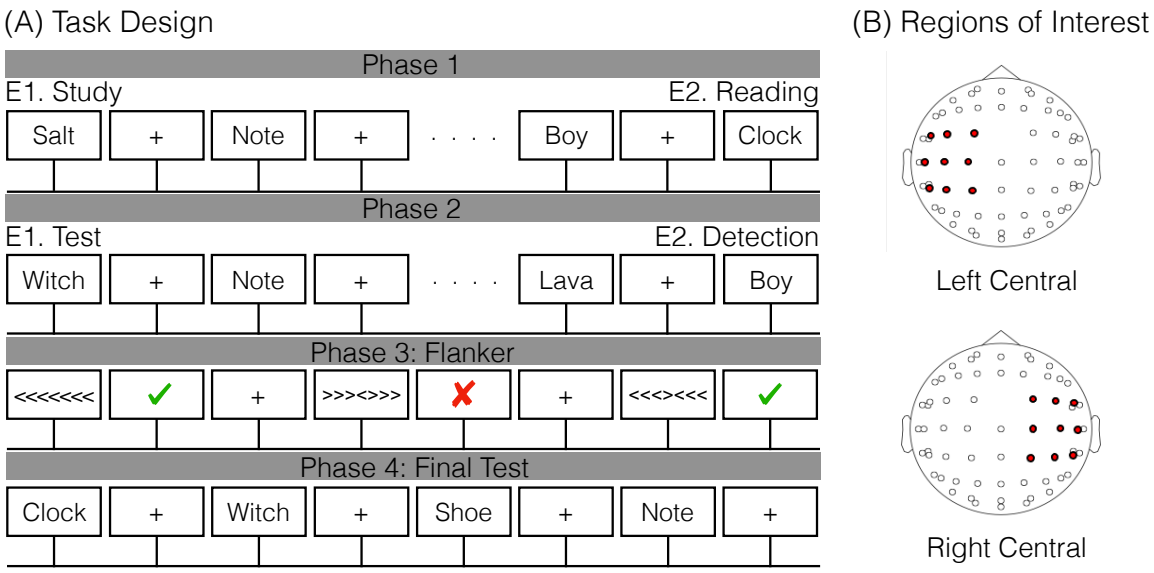

Figure 4.1: Task design and regions of interest. (A) The Phase 3 flanker task was divided into a practice subset of three runs and a main subset of six runs. The practice subset was completed prior to Phase 1 to determine response duration during the main subset (see Methods). In E1 Phase 1, participants studied individual words in anticipation of a later memory test. In E2 Phase 1, participants read the words silently. In E1 Phase 2, participants completed a recognition test and made old or new judgements using a confidence rating scale from 1 to 4, with 1 being definitely new and 4 being definitely old. In E2 Phase 2, participants completed a detection phase in which the goal was to detect new words that were not presented in Phase 1. Participants made old or new judgements without the use of a confidence rating scale. All participants then completed Phase 3, a flanker task, in which they made speeded responses to a central target. Immediately after each response, a green check mark indicating a correct response or a red X indicating an incorrect response was presented. In Phase 4, participants completed a final recognition memory test in which all the words from Phases 1 and 2 were presented along with novel lures. Participants made old or new judgements using a confidence rating scale from 1 to 4, with 1 being definitely new and 4 being definitely old. **(B)** We analyzed two regions of interest (ROIs), a left central ROI (FC5, FC1, C3, CP5, CP1, FC3, C1, C5, CP3) and a right central ROI (CP6, CP2, C4, FC6, FC2, CP4, C6, C2, FC4).

Recognition Task Experimental Design

We conducted two recognition memory experiments (E1, E2) and manipulated test phase instructions between subjects. Participants' goal was to successfully retrieve study items (E1) or to detect new items (E2). Stimuli consisted of 1602 words, drawn from the Toronto Noun Pool (Friendly et al., 1982). From this set, 640 words were randomly selected for each participant. Participants completed four phases

(Figure 4.1A). Phase 3 was divided into two subsets; the practice subset preceded Phase 1 and the main subset preceded Phase 3.

Phase 1. In each of 10 runs, participants viewed a list containing 16 words, yielding a total of 160 trials. On each trial, participants saw a single word presented for 2000 ms followed by a 1000 ms inter-stimulus interval (ISI). In E1, participants were instructed to study the presented word in anticipation for a later memory test and did not make any behavioral responses. In E2, participants were instructed to read the words silently and did not make any behavioral responses.

Phase 2. Participants completed a recognition memory test with different memory goals. On each trial, participants viewed a word which had either been presented during Phase 1 (target) or had not been presented (lure). In E1, participants' task was to make an old or new judgement for each word using a confidence rating scale from 1 to 4, with 1 being definitely new and 4 being definitely old. In E2, the task was framed as a detection phase in which participants' task was to detect new words that were not presented in Phase 1. Participants made an old or new judgment without the use of a confidence rating scale for each word by pressing one of two buttons ("d" or "k"). Response mappings were counterbalanced across participants. Phase 2 trials were self-paced and separated by a 1000 ms ISI. There were a total of 320 test trials with all 160 Phase 1 words presented along with 160 novel lures.

Phase 3. Prior to beginning Phase 1, participants completed three practice runs of a flanker task in which they made speeded responses to a central target in a string of congruent (e.g. >>>>>>>) or incongruent (e.g. <<<><<<) arrows. Feedback was presented immediately after each response as either a green check mark indicating

a correct response or a red X indicating an incorrect response. Response duration, the interval in which a response was accepted, was initially set to 375 ms based on pilot data. To maintain difficulty and ensure an approximately balanced number of correct and incorrect responses during the main subset of Phase 3, response duration was individually adjusted based on participants' accuracy following each practice run. If accuracy was below 50%, response duration increased by 25 ms, if accuracy was above 50%, response duration decreased by 25 ms. Thus, after completing the three practice runs, the final response duration could be a minimum of 300 ms and a maximum of 450 ms. After completing Phase 2, participants completed the main subset of Phase 3 which consisted of six runs of the flanker task. Throughout the main subset of Phase 3, the response duration was fixed to that obtained from the final practice run.

Phase 4. Participants completed a final recognition memory test in which all the words from Phase 1 and 2 were presented along with novel lures. Trials were self-paced and participants made old or new judgements for each word using a confidence rating scale from 1 to 4, with 1 being definitely new and 4 being definitely old. Trials were separated by a 1000 ms ISI. There were a total of 640 test trials with all 320 Phase 2 words presented along with 320 novel lures. As our analyses focus on Phases 2 and 3, we do not consider the final test data further.

EEG Data Acquisition and Preprocessing

EEG recordings were collected using a BrainVision system and an ActiCap equipped with 64 Ag/AgCl active electrodes positioned according to the extended 10-20 system. All electrodes were digitized at a sampling rate of 1000 Hz and were

referenced to electrode FCz. Offline, electrodes were later converted to an average reference. Impedances of all electrodes were kept below 50kΩ. Electrodes that demonstrated high impedance or poor contact with the scalp were excluded from the average reference. Bad electrodes were determined by voltage thresholding (see below).

Custom python codes were used to process the EEG data. We applied a high pass filter at 0.1 Hz, followed by a notch filter at 60 Hz and harmonics of 60 Hz to each participant's raw EEG data. We then performed three preprocessing steps (Nolan et al., 2010) to identify electrodes with severe artifacts. First, we calculated the mean correlation between each electrode and all other electrodes as electrodes should be moderately correlated with other electrodes due to volume conduction. We z-scored these means across electrodes and rejected electrodes with z-scores less than -3. Second, we calculated the variance for each electrode, as electrodes with very high or low variance across a session are likely dominated by noise or have poor contact with the scalp. We then z-scored variance across electrodes and rejected electrodes with a $|z| >= 3$. Finally, we expect many electrical signals to be autocorrelated, but signals generated by the brain versus noise are likely to have different forms of autocorrelation. Therefore, we calculated the Hurst exponent, a measure of long-range autocorrelation, for each electrode and rejected electrodes with a $|z| >= 3$. Electrodes marked as bad by this procedure were excluded from the average re-reference. We then calculated the average voltage across all remaining electrodes at each time sample and re-referenced the data by subtracting the average voltage from the filtered EEG data. We used wavelet-enhanced independent component analysis (Castellanos & Makarov, 2006) to remove artifacts from eyeblinks and saccades.

EEG Data Analysis

For E1 and E2, we applied the Morlet wavelet transform (wave number 6) to the entire EEG time series across electrodes, for each of 46 logarithmically spaced frequencies (2-100 Hz; Long & Kahana, 2015). After log-transforming the power, we downsampled the data by taking a moving average across 100 ms time intervals from -1000 to 3000 ms relative to the response and sliding the window every 25 ms, resulting in 157 time intervals (40 non-overlapping). Mean and standard deviation power were calculated across all trials and across time points for each frequency. Power values were then z-transformed by subtracting the mean and dividing by the standard deviation power. We followed the same procedure for the flanker task, with 117 overlapping (30 non-overlapping) time windows from 1000 ms preceding to 2000 ms following the response. We focus exclusively on the theta band (4-8 Hz) for all analyses.

Regions of Interest

We examined theta power across two regions of interest (ROIs; Figure 4.1B), left central (FC5, FC1, C3, CP5, CP1, FC3, C1, C5, CP3) and right central (CP6, CP2, C4, FC6, FC2, CP4, C6, C2, FC4). We specifically focus on the frontocentral region as prior work has demonstrated that theta power in these regions dissociates both hits and correct rejections (Burgess & Gruzelier, 1997; Klimesch et al., 1997; Gruber et al., 2008; Nyhus & Curran, 2010) and positive and negative feedback (Cohen et al., 2007; Marco-Pallarés et al., 2008; Mas-Herrero et al., 2015).

Univariate Analyses

We performed two univariate contrasts. First, to test the effect of instructions on theta power, our two conditions of interest were hits (correctly recognized targets) and correct rejections (CRs, correctly rejected lures) in Phase 2 in E1 and E2. Second, to test whether theta power is greater following negative relative to positive outcomes, our two conditions of interest were correct and incorrect responses in the Phase 3 flanker task. For each contrast and participant, we calculated z-transformed theta power across the ROI in each of the two conditions, across 100 ms intervals from 500 ms pre-response to 500 ms post-response, as well as averaged over the 500 ms pre-response interval and the 500 ms post-response interval.

Pattern Classification Analyses

Pattern classification analyses were performed using penalized (L2) logistic regression implemented via the sklearn module (0.24.2) in Python and custom Python code. We used our pilot data to determine classifier features. We compared performance of two pattern classifiers trained to dissociate positive vs. negative feedback trials in the flanker task. The classifier trained on spectral signals averaged across 63 electrodes and 46 frequencies yielded higher cross subject classification accuracy (M = 58.8%, SD = 5.04%) than the classifier trained on spectral signals averaged across the theta frequency band and 20 central electrodes (M = 56.4%, SD = 4.81%). Therefore, for all classification analyses, classifier features were comprised of spectral power across 63 electrodes and 46 frequencies. Before pattern classification analyses were performed, an additional round of z-scoring was performed across features (electrodes and frequencies) to eliminate trial-level differences in spectral

power (Kuhl & Chun, 2014; Long & Kuhl, 2018; D. E. Smith et al., 2022). Thus, mean univariate activity was matched precisely across all conditions and trial types. Classifier performance was assessed in two ways. "Classification accuracy" represented a binary coding of whether the classifier successfully guessed the type of flanker feedback, positive or negative. We used classification accuracy for general assessment of classifier performance (i.e., whether feedback could be decoded). "Classifier evidence" was a continuous value reflecting the logit-transformed probability that the classifier assigned the correct feedback label (positive, negative) for each trial. Classifier evidence was used as a trial-specific, continuous measure of feedback information, which was used to assess the degree of positive feedback evidence present following hit and CRs during Phase 2.

Cross Study Feedback Classification

To measure feedback evidence in E1 and E2, we conducted three stages of classification using similar methods as in our prior work (Long, 2023). First, we conducted within participant leave-one-run-out cross-validated classification (penalty parameter = 1) on all participants who completed the flanker task (N = 72). A classifier trained on the 500 ms post-response interval can reliably decode positive versus negative feedback (M = 63.2%, SD = 6.97%), however, during this interval, feedback is visually presented. Therefore, it is possible that the decoder is leveraging properties of the visual stimuli instead of feedback specific responses. To avoid this potential confound, we trained the classifier on the 100 ms time interval preceding the response as this interval is less contaminated by visual inputs and is likely to capture internal feedback signals. For each participant, we generated true and null classification accuracy values. We permuted condition labels (positive

feedback, negative feedback) for 1000 iterations to generate a null distribution for each participant. Second, we conducted leave-one-participant-out cross-validated classification (penalty parameter = 0.0001) on the selected participants to validate the feedback classifier and obtain classification accuracy of 62.7% which is significantly above chance ($t_{71} = 13.73$, $p < 0.0001$, $d = 2.304$), indicating that the cross subject feedback classifier is able to distinguish positive and negative feedback. Finally, we applied the cross subject feedback classifier to the Phase 2 trials of E1 and E2, specifically in 100 ms intervals from 500 ms pre-response to 500 ms post response. We extracted classifier evidence, the logit-transformed probability that the classifier assigned a given Phase 2 trial a label of positive feedback or negative feedback. This approach provides a trial-level estimate of positive feedback evidence during hits and CRs.

Statistical Analyses

We used mixed effects ANOVAs and *t*-tests to assess the effect of experiment (E1, E2) and response (hit, CR) on reaction times. We used mixed effects ANOVAs and *t*-tests to assess the effect of experiment (E1, E2), response (hit, CR) and time interval on central theta power and feedback evidence.

We used paired-sample *t*-tests to compare classification accuracy across subjects to chance decoding accuracy, as determined by permutation procedures. Namely, for each subject, we shuffled the condition labels of interest (e.g., "positive" and "negative" for the feedback classifier) and then calculated classification accuracy. We repeated this procedure 1000 times for each subject and then averaged the 1000 shuffled accuracy values for each subject. These mean values were used as

subject-specific empirically derived measures of chance accuracy.

4.4 Results

Faster reaction times when probe type and goal match.

We previously reported the impact of task goals on correct rejection (CR) rates and did not find a significant difference in CR rates across E1 and E2 (D. E. Smith & Long, 2024). Therefore, we conducted an exploratory analysis to investigate reaction times (RTs) for hits and CRs across both experiments. In a typical recognition memory experiment, participants are asked to recognize old items. The implicit assumption is that 'old' is equal to 'not-new', but behavioral evidence (Brainerd et al., 2021) suggests that recognizing old items is not equivalent to detecting new items. To the extent that encoding specificity drives memory judgments, we would expect to find facilitated responses (faster RTs) when the goal and probe type match (e.g. goal is to detect new and probe type is lure) compared to when the goal and probe type do not match (e.g. goal is to detect new and probe type is target). Alternatively, to the extent that participants use general or gist level information to reject items, we would expect to find faster RTs when the goal and probe type do not match compared to when the goal and probe type match.

As participants in E1 made confidence judgments and participants in E2 did not, we divided response types into six conditions, CRs (E2), high confidence CRs (E1, response of "1") and low confidence CRs (E1, responses of "2"), hits (E2), high confidence hits (E1, response of "4") and low confidence hits (E1, responses of

"3"; Figure 4.2). For a direct comparison between experiments, we excluded low confidence responses and conducted a 2×2 mixed effects ANOVA with factors of experiment (E1, E2) and response type (hit, CR). We do not find a main effect of experiment ($F_{1,70} = 0.137$, $p = 0.712$, $\eta_p^2 = 0.002$). We find a significant main effect of response type ($F_{1,70} = 42.91$, $p < 0.0001$, $\eta_p^2 = 0.38$) and a significant interaction between response type and experiment ($F_{1,70} = 22.05$, $p < 0.0001$, $\eta_p^2 = 0.24$). This interaction was driven by significantly faster CRs in E2 (M = 1302 ms, SD = 328.1 ms) than high confidence CRs in E1 (M = 1479 ms, SD = 379.6 ms; $t_{70} = 2.080$, $p = 0.0411$, $d = 0.4973$) and numerically faster high confidence hits in E1 (M = 1123 ms, SD = 200.3 ms) than hits in E2 (M = 1244 ms, SD = 421.1 ms; $t_{70} = 1.529$, $p = 0.1308$, $d = 0.3655$). These findings are consistent with the encoding specificity account. When the participant's goal (e.g. detect new) matches the probe type (e.g. lure) responses are faster. These results demonstrate that the manipulated task goals impact recognition memory responses.

Post-response theta power decreases for hits vs CRs regardless of goals.

Our first goal was to replicate our previous finding that theta power is modulated by successful retrieval leading up to and following memory responses (D. E. Smith et al., 2024). To specifically test for a pre- vs. post-response dissociation in theta power for hits and CRs across E1 and E2, we averaged signals within the 500 ms pre- and post-response time intervals (Figure 4.3, insets). To the extent that theta power is modulated by successful retrieval, the same post-response signals should dissociate hits vs. CRs regardless of memory goals. However, to the extent

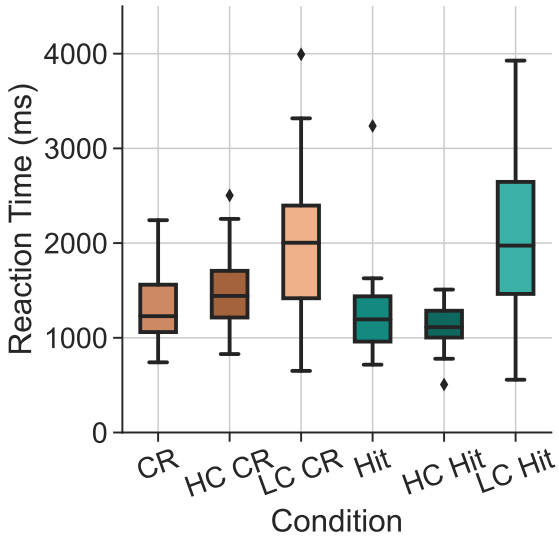

Figure 4.2: Test phase instructions modulate reaction times. Reaction times for hits and correct rejections (CRs). We divided hits and CRs in E1 into high confident (HC Hit, dark teal; HC CR, dark orange) and low confident (LC Hit, light teal; LC CR, light orange) responses. Box-and-whisker plots show median (center line), upper and lower quartiles (box limits), 1.5x interquartile range (whiskers) and outliers (diamonds).

that theta power dissociations are driven by goal attainment, we would expect decreased post-response theta for hits compared to CRs in E1 and decreased post-response theta for CRs compared to hits in E2. Following our pre-registration and our prior work, we conducted two $2 \times 2 \times 2$ mixed effects ANOVA, one for each ROI (left central; LC, right central; RC) with factors of experiment (E1, E2), average time interval (pre-response, post-response) and response type (hit, CR). We report the results of this ANOVA in Table 4.1 and highlight the key findings here. If successful retrieval drives the post-response theta decrease for hits, we would expect to find an interaction between response type and time interval and no experiment driven interactions. Alternatively, if goal attainment drives the post-response theta decrease, we would expect to find an interaction between experiment, time, and response type. Across both ROIs, we find a significant main effect of time interval (LC: $F_{1,70} = 24.79$, $p < 0.0001$, $\eta_p^2 = 0.2615$; RC: $F_{1,70} = 15.46$,

$p = 0.0002$, $\eta_p^2 = 0.1809$). We find a significant interaction between response type and time interval (LC: $F_{1,70} = 52.015$, $p < 0.0001$, $\eta_p^2 = 0.4263$; RC: $F_{1,70} = 44.00$, $p < 0.0001$, $\eta_p^2 = 0.386$). The three-way interaction between response type, time interval, and experiment was not significant (LC: $F_{1,70} = 0.154$, $p = 0.6959$, $\eta_p^2 = 0.0022$; RC: $F_{1,70} = 0.2046$, $p = 0.6525$, $\eta_p^2 = 0.0029$). Bayes factor analysis revealed that a model without the experiment term is preferred to a model with the three-way interaction by a factor of 754.6 in LC and 83.6 in RC. Together, these results indicate that theta power is modulated by successful retrieval leading up to and following memory responses regardless of task goals.

Table 4.1: Theta power for left and right central ROIs as a function of experiment, response type, and average time interval, mixed effects ANOVAs

		Left Central			Right Central		
Effect	df	F	p	η_p^2	F	p	η_p^2
Main effect of experiment	(1,70)	0.8738	0.3531	0.0123	0.9006	0.3459	0.0127
Main effect of response type	(1,70)	1.112	0.2953	0.0156	11.51	**0.0011**	0.1412
Main effect of time interval	(1,70)	24.79	< **0.0001**	0.2615	15.46	**0.0002**	0.1809
Interaction of experiment × response type	(1,70)	0.2755	0.6013	0.0039	4.545	**0.0365**	0.061
Interaction of experiment × time interval	(1,70)	1.778	0.1867	0.0248	1.672	0.2002	0.0233
Interaction of response type × time interval	(1,70)	52.01	< **0.0001**	0.4263	44.00	< **0.0001**	0.3860
Interaction of experiment × response type × time interval	(1,70)	0.1540	0.6959	0.0022	0.2046	0.6525	0.0029

Note: bold values indicate p < 0.05.

Robust feedback decoding in the flanker task.

Before testing our central hypothesis, we validated that feedback in the flanker task could be reliably decoded. We conducted a multivariate pattern classification analysis in which we trained a classifier to discriminate positive versus negative feedback trials based on a feature space comprised of all 63 electrodes and 46 frequencies ranging from 2-100 Hz. For this analysis, we averaged z-power over the 100 ms preceding the response. Using leave-one-participant-out cross-validated classification (penalty parameter = 0.0001), mean classification accuracy was 62.7%

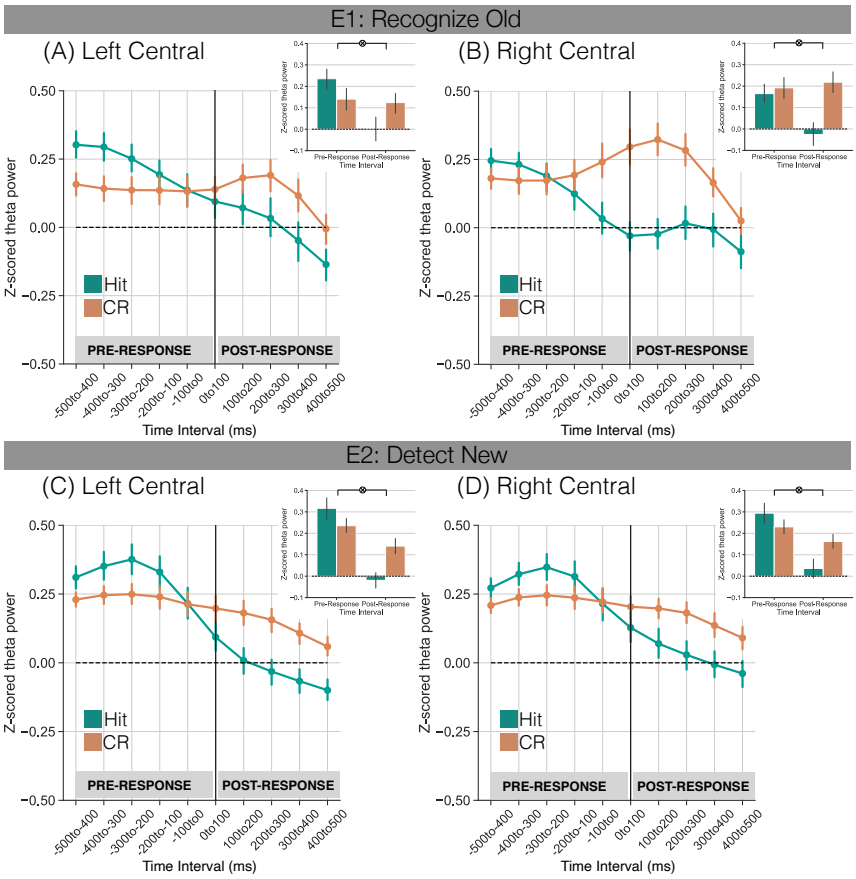

Figure 4.3: Theta power dissociations across hits and correct rejections preceding and following memory responses in E1 and E2. Response-locked z-transformed theta power (4-8 Hz) for the left and right central ROIs. The solid vertical black line indicates when the response was made. Hits are shown in teal, correct rejections (CRs) are shown in orange. Error bars reflect standard error of the mean. **(A-B)** E1, participants recognize old items. **(A)** Over left central ROI, we find a significant interaction between response type and time interval ($p < 0.001$) driven by numerically greater pre-response theta power for hits than CRs and greater post-response theta power for CRs than hits. **(B)** Over right central ROI, we find a significant interaction between response type and time interval ($p < 0.001$) driven by greater post-response theta power for CRs than hits. **(C-D)** E2, participants detect new items. **(C)** Over left central ROI, we find a significant interaction between response type and time interval ($p < 0.001$) driven by greater pre-response theta power for hits than CRs and greater post-response theta power for CRs than hits. **(D)** Over right central ROI, we find a significant interaction between response type and time interval ($p < 0.001$) driven by numerically greater pre-response theta power for hits than CRs and greater post-response theta power for CRs than hits.

(SD = 7.79%), which was significantly greater than chance, as determined by permutation tests ($t_{71} = 13.73$, $p < 0.0001$, $d = 2.304$; Figure 4.4A).

Greater positive feedback evidence for hits vs CRs regardless of goals.

Our central goal was to test the hypothesis that successful retrieval is intrinsically rewarding. Our interpretation is that post-response theta power dissociations reflect a feedback signal, in which there is greater positive feedback following hits. However, participants did not receive any explicit rewards or feedback during E1 and E2, therefore we cannot definitively say whether these post-response signals reflect positive feedback. In order to test our hypothesis, we used the flanker task to develop an independent measure of feedback. We trained a classifier to discriminate positive versus negative feedback trials in flanker and then tested the classifier on Phase 2 hits and CRs in E1 and E2. To the extent that successful retrieval is intrinsically rewarding, we expect to find greater positive feedback evidence following hits compared to CRs in both experiments. Alternatively, to the extent that feedback signals reflect goal attainment, we expect to find greater positive feedback evidence following CRs compared to hits in E2 relative to E1.

To investigate the effect of instructions and response type on feedback evidence over time, we trained a classifier to discriminate positive versus negative feedback trials using the average z-power from the 100 ms preceding the response and then tested the classifier on ten 100 ms time intervals from 500 ms preceding and following the response in Phase 2 of E1 and E2. To specifically test for a pre- vs. post-response dissociation in positive feedback evidence following hits and CRs across E1 and E2, we averaged signals within the 500 ms pre- and post-response time intervals (Figure 4.4B,C). Following our pre-registration, we conducted a 2 $\times$ 2 $\times$ 2 mixed effects ANOVA with factors of experiment (E1, E2), average time

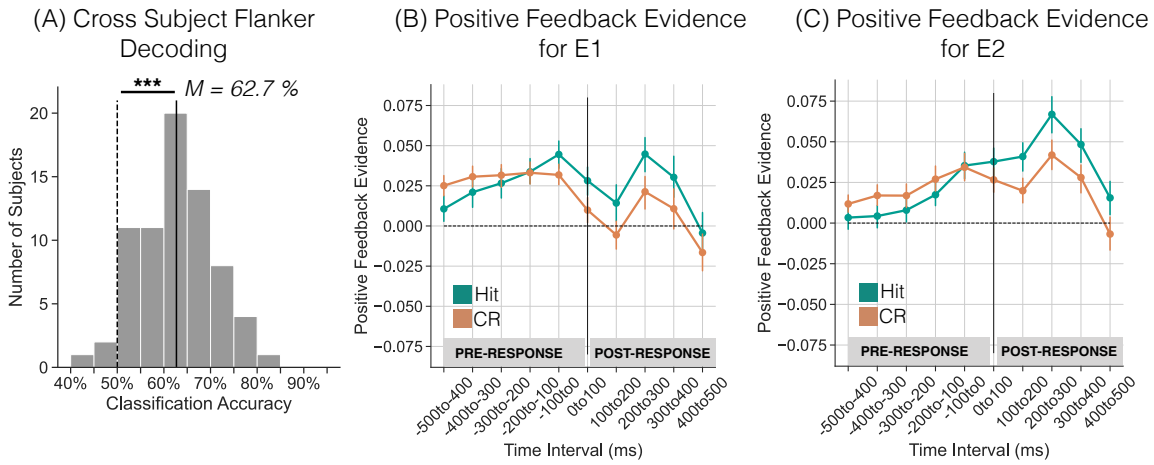

Figure 4.4: Positive feedback evidence over time in E1 and E2. (A) Mean classification accuracy across all participants (solid vertical line) is shown along with a histogram of classification accuracies for individual participants (gray bars) and mean classification accuracy for permuted data across all participants (dashed vertical line). Mean classification accuracy was 62.7%, which differed significantly from chance (two-tailed, paired t-test, $p < 0.0001$) **(B-C)** Positive y-axis values indicate greater positive feedback evidence. The solid vertical line at time 0-100 ms indicates the response. Each panel shows positive feedback evidence separated by response type (teal: hit; orange: CR) across the 500 ms pre-response and post-response time intervals. There is greater post-response positive feedback evidence for hits compared to CRs ($p = 0.0001$), regardless of experimental goals. Error bars represent standard error of the mean. $^{***}p < 0.001$

interval (pre-response, post-response) and response type (hit, CR). We do not find a significant main effect of experiment ($F_{1,70} = 0.2409$, $p = 0.6251$, $\eta_p^2 = 0.0034$), response type ($F_{1,70} = 3.6323$, $p = 0.0608$, $\eta_p^2 = 0.0493$), or time interval ($F_{1,70} = 0.0091$, $p = 0.9241$, $\eta_p^2 = 0.0001$). The interaction between response type and experiment was not significant ($F_{1,70} = 0.0493$, $p = 0.8249$, $\eta_p^2 = 0.0007$). We find a significant interaction between time interval and experiment ($F_{1,70} = 5.496$, $p = 0.0219$, $\eta_p^2 = 0.0728$) and between time interval and response type ($F_{1,70} = 22.01$, $p < 0.0001$, $\eta_p^2 = 0.2392$). The interaction between time interval and response type was driven by significantly greater post-response positive feedback evidence for hits (M = 0.03, SD = 0.05) compared to CRs (M = 0.01, SD = 0.05; $t_{71} = 4.105$, $p = 0.0001$, $d = 0.3824$). The three-way interaction between response type, time interval, and experiment was not significant ($F_{1,70} = 0.3104$, $p = 0.5792$, $\eta_p^2 = 0.0044$). Bayes factor analysis

revealed that a model without the experiment term is preferred to a model with the three-way interaction by a factor of 4.56. These results suggest that successful retrieval may be intrinsically rewarding.

4.5 Discussion

The aim of this study was to identify whether test phase theta signals reflect intrinsic reward or goal attainment. We conducted two independent recognition memory experiments in which we manipulated participants' test phase goals to either recognize old items (E1) or detect new items (E2). We recorded scalp EEG and used a cross-study decoding approach (Long, 2023) to measure response-locked positive feedback evidence during the test phase of each experiment. We find post-response theta power decreases for hits compared to CRs regardless of task goals and greater positive feedback evidence for hits compared to CRs across both experiments. Together, these findings suggest that successful retrieval is intrinsically rewarding.

We find faster reaction times when the task goal (recognize old, detect new) and probe type (target, lure) match. This finding is in line with the encoding specificity hypothesis whereby performance is improved when the study and test phase contexts are more similar (Brainerd et al., 2021). The theory of transfer-appropriate processing may also account for this finding, which emphasizes that the same processes engaged during study will be engaged again during the test (Morris, Bransford, & Franks, 1977; Roediger, 1990). The current findings are counter to previous behavioral work that finds better memory accuracy when the stimulus type and instruction don't match, selectively for new items or lures (Brainerd et

al., 2021), consistent with the Fuzzy Trace Theory (FTT) noncompensatory gist principle. One possibility to account for the difference in findings may be due the high level of CR rates in the current study which on average were above 80% across both experiments. The overall higher performance may indicate differences in the processes in which participants' engaged across the current compared to prior work. Our results add to existing work demonstrating that task goals can impact the memory and decision processes participants engage when responding to stimuli.

Consistent with prior EEG work (D. E. Smith et al., 2024), we find greater post-response theta power for CRs compared to hits, regardless of the task goals. Finding the same result regardless of the two task goals suggests that there might be something inherent about recognizing a stimulus as 'old', even when that is counter to the task goal. Our interpretation is that the post-response theta dissociation that we have identified in our work represents an intrinsic feedback signal in response to successful retrieval. Along with neuro-imaging work that finds greater reward system activity during hits compared to CRs (Spaniol et al., 2009), this interpretation is consistent with work from reward and feedback based learning literature that finds theta power increases following incorrect relative to correct responses and negative relative to positive outcomes (Mazaheri et al., 2009; Cavanagh & Frank, 2014). As both hits and CRs constitute accurate trials, the dissociation in activity for hits compared to CRs suggests that the signal change is not driven by dissociations between correct and incorrect trials, but rather is a response to successful retrieval. Thus, the post-response dissociation in theta power for hits and CRs may represent an intrinsic feedback signal in response to successful retrieval, rather than a general positive feedback signal following accurate responses.

Independent of the task goal, we find greater positive feedback evidence for hits than CRs, suggesting that successful retrieval is intrinsically rewarding. Our interpretation is that the reward system – in particular, the striatum – generates a positive feedback signal in response to successful retrieval. Previous work has shown that when a study item is associated with a potential reward – e.g. a reward that will be received if the item is remembered at test – leads to the item being better remembered later (Loftus & Wickens, 1970; Marini et al., 2011). This memory enhancement for study items assigned a potential reward is driven by correlated activity between the ventral tegmental area, striatum and hippocampus, in which reward signals up-regulate memory encoding mechanisms in the hippocampus (Adcock et al., 2006; Wolosin, Zeithamova, & Preston, 2012). Potentially, in the same way that study phase anticipation of a future extrinsic reward can enhance memory formation via striatal-hippocampal connectivity (Adcock et al., 2006; Wolosin et al., 2012), the same mechanisms may be engaged by intrinsic reward/feedback. Future work will be needed to directly test this possibility.

An intrinsic feedback signal following successful retrieval has the potential to influence our ability to remember information. During the study phase of a memory experiment, an item and the strategy used to encode an item are associated together and then reinstated during memory retrieval (Polyn, Kragel, Morton, McCluey, & Cohen, 2012). Accordingly, an intrinsic reward signal during the test phase should influence the item as well as the reinstated strategy. Therefore, test phase reward signals in response to successful retrieval may influence the use of memory strategies. The direct investigation of test phase intrinsic reward on reinforcing memory strategies to improve memory performance presents an exciting avenue for future work.

Together, our findings suggest that there is a positive feedback signal in response to successful retrieval. A direction for future research will be to directly investigate the extent to which the positive feedback signal reinforces the information retrieved and impacts subsequent memory performance. These findings are highly relevant to a growing body of literature characterizing the relationship between memory and reinforcement learning. Together, these findings demonstrate that successful retrieval is intrinsically rewarding, which has implications across many cognitive contexts.

Chapter 5

The impact of post-retrieval test-phase extrinsic reward on subsequent memory

Devyn E. Smith & Nicole M. Long

5.1 Abstract

The anticipation of extrinsic reward facilitates memory formation. However, it is unclear how reward following memory retrieval influences the information that is retrieved and later memory. Here, we conducted four behavioral experiments in which we manipulated test phase reward delivery. Across all experiments, participants studied word-image pairs and then completed two rounds of retrieval

practice, followed by a final recognition test. During retrieval practice, participants were given a word cue and instructed to bring to mind the associated image. Participants rated the vividness of their memory for the image and every response had a 50% chance of receiving reward feedback. Although we find no impact of either retrieval practice or reward on subsequent memory performance we find that retrieval practice improves recognition for highly vivid items, but impairs recognition for items low in vividness. We find some evidence that repeated rewards improve recognition for high vivid items. Together, these results suggest that the benefit of both retrieval practice and reward may be dependent on the strength of the memory that is retrieved.

5.2 Introduction

A viable approach for improving memory may be through the use of extrinsic reward (e.g. monetary compensation) as information that is valuable or rewarding is prioritized over information that is less rewarding (Loftus & Wickens, 1970). However, the extent to which reward *following* memory retrieval impacts subsequent behavior is unknown. Memory reinforcement may be better accomplished through direct reward of what is retrieved, rather than through study phase manipulation of potential future reward. Prior behavioral work that has investigated test phase extrinsic reward has found mixed results (Shigemune et al., 2017; Castanheira et al., 2022). However, these studies used anticipatory methods such that participants were aware prior to retrieval that there was a potential to receive a reward for remembering. Additionally, these studies investigated the immediate influence of reward rather than potential long term impacts of reward on later memory.

Thus, it is an open question how receiving a reward immediately following memory retrieval impacts neural processing and subsequent behavior. The aim of this study was to investigate how extrinsic reward following memory retrieval impacts subsequent memory.

Practicing retrieval without any rewards is well known to improve later memory (Roediger & Karpicke, 2006; Karpicke & Roediger, 2008; Karpicke, 2012). According to transfer appropriate processing, retrieval practice benefits memory because the same processes engaged during retrieval practice will be engaged during the final memory test (Morris et al., 1977). Another explanation for the benefits of retrieval practice is that of desirable difficulties whereby a desirable amount of challenge or difficulty improves long-term retention (R. A. Bjork, 1994; E. L. Bjork & Bjork, 2011). Retrieval practice can facilitate memory but also increase errors to similar novel stimuli if general information common to both study items and lures is strengthened (McDermott, 2006) thus, presenting reward during practice following retrieval may or may not serve to facilitate later memory.

It is well established that associating a study item with a potential reward – e.g. reward that will be received if the item is remembered at test – impacts the likelihood that the item is later remembered (Loftus & Wickens, 1970; Marini et al., 2011; Elliott, Blais, McClure, & Brewer, 2020), with higher potential rewards leading to better subsequent memory. Enhanced subsequent memory for study items assigned a potential reward is driven by correlated activity between reward regions (e.g ventral tegmental area, striatum) and the hippocampus, in which reward signals up-regulate memory encoding mechanisms in the hippocampus (Adcock et al., 2006; Wolosin et al., 2012). Given that the hippocampus supports both memory encoding and memory retrieval (Eichenbaum, 2004; Diana, Yonelinas,

& Ranganath, 2007; Long et al., 2017), a similar interaction between the reward system and hippocampus during a memory test may also serve to enhance memory performance.

Our current understanding of the role of test phase reward is unclear as extant studies suggest both that test phase rewards improve memory and have no effect on memory. Shigemune and colleagues (Shigemune et al., 2017) had participants study words in high or low difficulty tasks and then complete a memory test. Prior to each test trial, participants were given a cue to inform them that correctly recognizing a study item would result in a high or low reward. Hit rates (rates of correctly recognizing study items) were higher in the high compared to low reward condition in the high difficulty task. However, another study with a similar design (Castanheira et al., 2022) found no effect of potential reward during test. Thus, test phase reward yields conflicting behavioral outcomes, leaving our understanding of the role of test phase reward limited. Furthermore, as both of these studies measured the influence of reward on immediate memory performance, the extent to which these test phase rewards impact later memory is unknown.

A limitation of all existing studies – regardless of whether potential reward is manipulated during the study or test phase – is that reward delivery is always anticipatory. In these motivated memory studies, individuals are aware prior to encountering a study or test stimulus that there is the potential to receive a reward for remembering that stimulus. Such a design will impact how upcoming stimuli are processed. However, rewards following retrieval should modify neural processing after a stimulus is retrieved. It may be through post-retrieval mechanisms, rather than pre-retrieval anticipatory mechanisms, that reward reinforces the contents of retrieval. A post-retrieval reward may thus lead to alterations in memory

representations and future behavior, effects which cannot be investigated in anticipatory designs. Thus, how extrinsic reward immediately following memory retrieval impacts the information that is retrieved remains an open question.

Our hypothesis is that extrinsic reward following retrieval will reinforce the information that is retrieved and modulate subsequent memory. To test our hypothesis, we conducted four behavioral experiments (E1, E2, E3, E4) in which we manipulated test phase reward delivery. In each experiment, participants completed three phases. Across all experiments, participants studied word-image pairs and then completed two rounds of retrieval practice, followed by a final recognition test. During retrieval practice, participants were given a word cue and instructed to bring to mind the associated image. They rated the vividness of their memory for the image on a scale from one (least vivid) to four (most vivid). Every response had a 50% chance of receiving reward feedback. To the extent that reward reinforces the information that is retrieved, we should find increased memory performance for reward items compared to no-reward items.

5.3 Methods

Participants

168 native English speakers from the University of Virginia community participated, with forty two participants enrolled in each experiment (E1: 22 female; age range = 18-22, mean age = 19.2 years; E2: 27 female; age range = 18-21, mean age = 18.8 years; E3: 26 female; age range = 18-21, mean age = 19 years; E4: 25 female; age

range = 18-21, mean age = 19.1 years). All participants had normal or corrected-to-normal vision. Informed consent was obtained in accordance with University of Virginia Institutional Review Board for Social and Behavioral Research and participants received class credit for their participation. Our sample size was determined *a priori* based on pilot data (E2, N = 14) described in the pre-registration report of this study (https://osf.io/gebm4). A total of two participants (one each from E2 and E4) were excluded from the final dataset due to a d' > 2.5 SDs of the mean across the four experiments. Thus data are reported for the remaining 166 participants. All raw, de-identified data and the associated experimental and analysis codes used in this study will be made available via the Long Term Memory Lab Website upon publication.

Recognition Task Experimental Design

We conducted four recognition memory experiments (E1, E2, E3, E4) each with three phases (Figure 5.1) and manipulated test phase reward delivery between subjects. Stimuli consisted of 1602 words, drawn from the Toronto Noun Pool (Friendly et al., 1982) and three categories of images: 490 common objects (e.g., banjo), drawn from an image database with multiple exemplars per object category (Konkle et al., 2010), 96 famous faces (e.g., Paul Rudd) and 96 famous scenes (e.g., Taj Mahal; Lee, Samide, Richter, & Kuhl, 2018). From this set, 192 words and 288 images were selected for each participant. The images consisted of an equal number (96) of objects, faces, and scenes. Of the 288 images, a subset of 192 were presented in Phase 1, with 64 images drawn from each visual category. Only one exemplar per object category appeared during Phase 1 (e.g. one banjo). Word-image associations were randomly generated for

each participant and randomly assigned to condition (e.g. target or lure, see below).

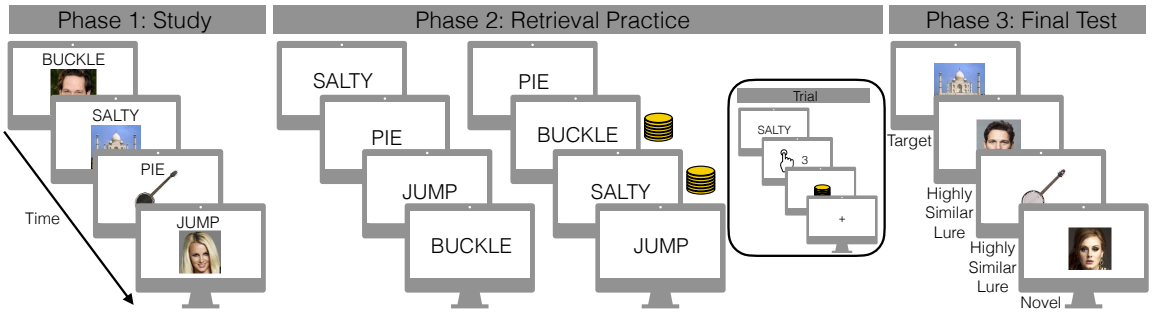

Figure 5.1: Task design. During Phase 1, participants studied word-image pairs; images were from one of three categories: famous faces (e.g. Paul Rudd), famous scenes (e.g. Taj Mahal), and common objects (e.g. banjo). In Phase 2, participants completed two rounds of retrieval practice. Participants saw an individual word and were instructed to bring to mind the image associated with each word and make vividness ratings on a scale from 1 to 4, with 1 being least vivid and 4 being most vivid. In E1, during the first round of retrieval practice, every response had a 50% chance of receiving reward feedback displayed as coins. In E2, during the second round of retrieval practice, every response had a 50% chance of receiving reward. In E3, participants received rewards during both rounds of retrieval practice. In the first round, as in E1, every response had a 50% chance of receiving reward. If reward followed an item in the first round of retrieval practice, reward followed the same item in the second round. In E4, participants did not receive any reward during either round of retrieval practice. The temporal dynamics of a trial during retrieval practice are as follows: participants see a word, rate the vividness of their memory for the image, then a reward could immediately follow. All participants then completed Phase 3, a final recognition memory test that included images only. Test probes included previously studied images (targets), highly similar lures (non-identical images depicting the same person, place, or object as the targets), and novel images. Participants made old or new judgements using a confidence rating scale from 1 to 4, with 1 being definitely new and 4 being definitely old.

Phase 1: Study. In each of four runs, participants studied 48 word-image pairs, yielding a total of 192 trials. On each trial, participants saw a word-image pair presented for 2000 ms followed by a 3500 ms distractor interval. The distractor interval was comprised of alternating fixation and digit presentation (fixation, digit, fixation, digit, fixation). During digit presentation, participants saw a single digit, 1 through 10, and were instructed to press one of two buttons ("1" or "2") to indicate if the number was odd or even. Each fixation was 500 ms and each digit presentation was 1000 ms.

Phase 2: Retrieval Practice. Participants completed two rounds of retrieval practice of 96 trials each. A total of 96 words from Phase 1 were presented. The same words were presented in both rounds of retrieval practice in random order. Each round was further sub-divided into two runs with 48 trials. In each run, an equal number of words (16) associated with an image from each visual category were presented. On each trial, participants were presented with a word cue for 4000 ms and instructed to bring to mind the associated image from the study phase. Participants rated the vividness of their memory of the retrieved image on a scale from 1 to 4, with 1 being least vivid and 4 being most vivid. To motivate participants to use the full response scale, if the same vividness response was made on five or more consecutive trials, participants received a message instructing them to use the full scale. Following the word cue, participants saw either a square of scrambled images or feedback displayed as coins overlaid on the square of scrambled images for 1000 ms, followed by a 500 ms inter-stimulus interval (ISI).

We manipulated reward delivery across experiments. In E1, during the first round of retrieval practice, every response had a 50% chance of receiving reward feedback. In E2, during the second round of retrieval practice, every response had a 50% chance of receiving reward feedback. In E3, participants received rewards during both rounds of retrieval practice. In the first round, as in E1, every response had a 50% chance of receiving reward feedback. If reward followed an item in the first round of retrieval practice, reward followed the same item in the second round. In E1, E2, and E3, 48 total words were rewarded (16 words associated with images from each visual category). In E4, participants did not receive any reward during either round of retrieval practice.

Phase 3: Recognition Test. Participants completed a final recognition memory

test for images only. Trials were self-paced and participants made old or new judgements for each image using a confidence rating scale from 1 to 4, with 1 being definitely new and 4 being definitely old. Trials were separated by a 500 ms ISI. There were a total of 288 test trials. To reduce test phase interference, participants were only tested on either a target (e.g. original image of Paul Rudd) or the similar lure (e.g. new image of Paul Rudd), as in prior work (Lee et al., 2018). Test probes included 96 previously studied images (targets), 96 highly similar lures (non-identical images depicting the same face, scene, or object as the Phase 1 image), and 96 novel face, scene, or object images. There we an equal number of images from each visual category for each test probe condition (i.e. 32 novel scenes). For E1, E2, and E3, half of the targets were rewarded during retrieval practice and half were not rewarded. Similarly, half of the lures were associated with an image that was rewarded during retrieval practice and half were associated with an image that was not rewarded during retrieval practice. We refer to these as "rewarded targets" and "rewarded lures" although the rewards are always presented during the Phase 2 retrieval practice and not during the final Phase 3 recognition test.

Statistical Analyses

We used mixed effects ANOVAs to assess the effect of retrieval practice, reward structure, and vividness on hit rate and false alarm rate. We used an independent samples *t*-test to compare E1-E3 reward and no reward hit rate and false alarm rate to E4 hit rate and false alarm rate.

5.4 Results

No evidence that retrieval practice impacts subsequent memory.

Repeated testing or retrieval practice improves long term retention relative to restudying (Roediger & Karpicke, 2006), therefore across experiments, we expected to find greater hit rates for targets that were presented during the retrieval practice phase compared to those that were not practiced. To the extent that retrieval practice increases errors to related lures (McDermott, 2006), we also expected to find greater false alarm (FA) rates for similar lures. We conducted a 4×2 mixed effects ANOVA with experiment (E1, E2, E3, E4) and practice condition (practice, no practice) as factors and hit rate as the dependent variable (Figure 5.2A). We do not find a main effect of experiment ($F_{3,162} = 0.218$, $p = 0.884$, $\eta_p^2 = 0.004$) or a main effect of practice condition ($F_{1,162} = 0.665$, $p = 0.416$, $\eta_p^2 = 0.0041$). Bayes factor analysis revealed that a model without practice condition is preferred to a model with the practice condition by a factor of 7.97. The interaction between experiment and practice condition was also not significant ($F_{3,162} = 0.238$, $p = 0.869$, $\eta_p^2 = 0.0044$). We then conducted a second 4×2 mixed effects ANOVA with experiment (E1, E2, E3, E4) and practice condition (practice, no practice) as factors and FA rate as the dependent variable (Figure 5.2B). We do not find a main effect of experiment ($F_{3,162} = 0.633$, $p = 0.594$, $\eta_p^2 = 0.01$) or a main effect of practice condition ($F_{1,162} = 2.025$, $p = 0.157$, $\eta_p^2 = 0.01$). Bayes factor analysis revealed that a model without practice condition is preferred to a model with the practice condition by a factor of 5.93. The interaction between experiment and practice condition was also not significant ($F_{3,162} = 0.838$, $p = 0.475$, $\eta_p^2 = 0.02$). Together, these findings provide no evidence that repeated practice impacts subsequent memory.

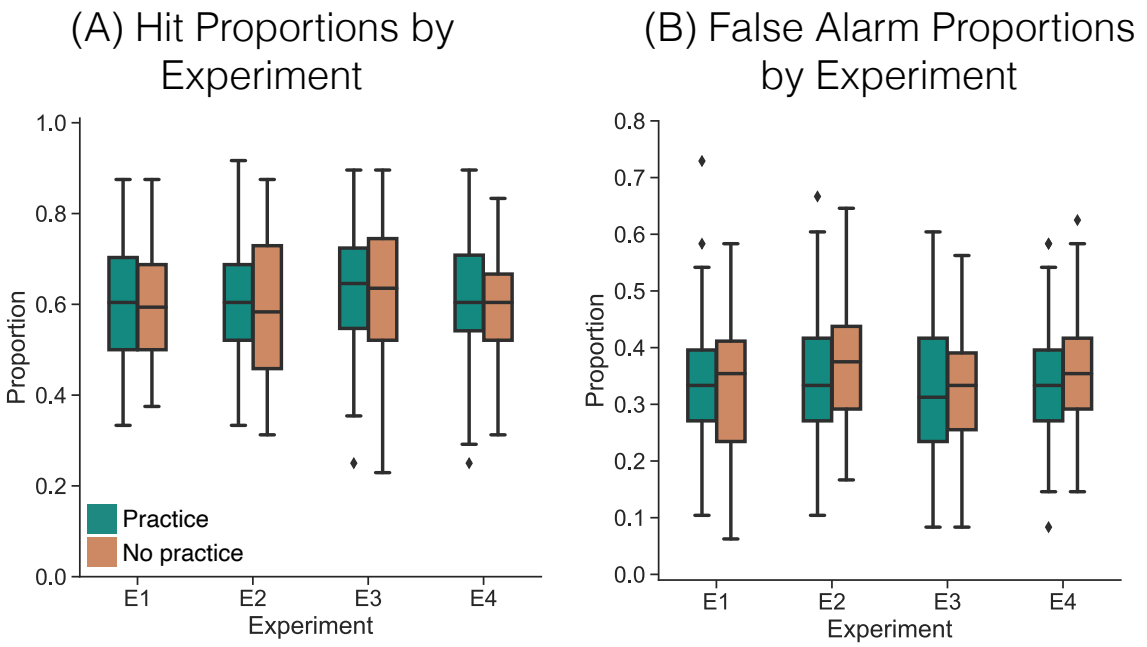

Figure 5.2: Influence of retrieval practice on hit and false alarm rates. **A** Hit rate proportions for practiced and not practiced targets for each experiment. We do not find a significant interaction between practice condition and experiment ($p = 0.869$). **B** False alarm (FA) rate proportions for practiced lures and not practiced lures for each experiment. We do not find a significant interaction between practice condition and experiment ($p = 0.475$). Practiced is shown in teal and not practiced is shown in orange. Box-and-whisker plots show median (center line), upper and lower quartiles (box limits), 1.5x interquartile range (whiskers) and outliers (diamonds).

No evidence that rewarded retrieval practice impacts subsequent memory.

Our central goal was to test the hypothesis that extrinsic reward following retrieval reinforces the information that is retrieved and modulates subsequent memory. Specifically, we expected to find greater hit rates for reward E1-E3 targets compared to E4 targets. Insofar as the reward/no-reward structure in E1-E3 constitutes a "mixed list" (as opposed to "pure list") condition (Ratcliff, Clark, & Shiffrin, 1990), we might expect to find lower hit rates for no-reward E1-E3 targets compared to E4 targets. That is, not receiving a reward in a context of reward may be worse

for memory than not receiving a reward in a context of no-reward. Following our preregistration, we treated the data from E1-E3 as one experiment and used independent samples *t*-tests to compare reward and no-reward hit rates to E4 hit rates. We did not find a significant difference in hit rates (t_{164} = 0.0757, p = 0.9397, d = 0.0141, BF = 0.19) between reward E1-E3 trials (M = 0.6103, SD = 0.1578) and E4 (M = 0.6082, SD = 0.1387; Figure 5.3A) or between no-reward E1-E3 trials (M = 0.602, SD = 0.1576) and E4 (t_{164} = -0.2247, p = 0.8225, d = 0.042, BF = 0.20). Next, we compared reward and no-reward FA rates to E4 FA rates and based on our pilot data, expected to find lower FA rates for reward E1-E3 lures compared to E4 lures. However, we did not find a significant difference in FA rates (t_{164} = -0.1082, p = 0.914, d = 0.0204, BF = 0.19) between reward E1-E3 trials (M = 0.3363, SD = 0.1374) and E4 (M = 0.3389, SD = 0.1149; Figure 5.3B) or between no reward E1-E3 trials (M = 0.3383, SD = 0.1364) and E4 (t_{164} = -0.0248, p = 0.9803, d = 0.0047, BF = 0.19).

Although we found no difference in hit or FA rates between E1-E3 and E4, it is possible that reward structure has an impact on subsequent memory. Both hit rates and FA rates might be greater following two rounds of rewarded retrieval practice (E3) compared to only one round of rewarded retrieval practice (E1, E2). Additionally, hit rates may differ as a function of when during practice rewards are delivered (E1 vs. E2). To the extent that first round retrieval rewards strengthen image representations and thereby facilitate second round retrieval, E1 hit rates should be greater than E2 hit rates. To the extent that the lack of rewards during second round retrieval weakens representations, E1 hit rates should be lower than E2 hit rates. We conducted a 2 × 3 mixed effects ANOVA with reward (reward, no-reward) and experiment (E1, E2, E3) as factors and hit rate as the dependent variable (Figure 5.4A). We do not find a main effect of experiment ($F_{2,122}$ = 0.43, p

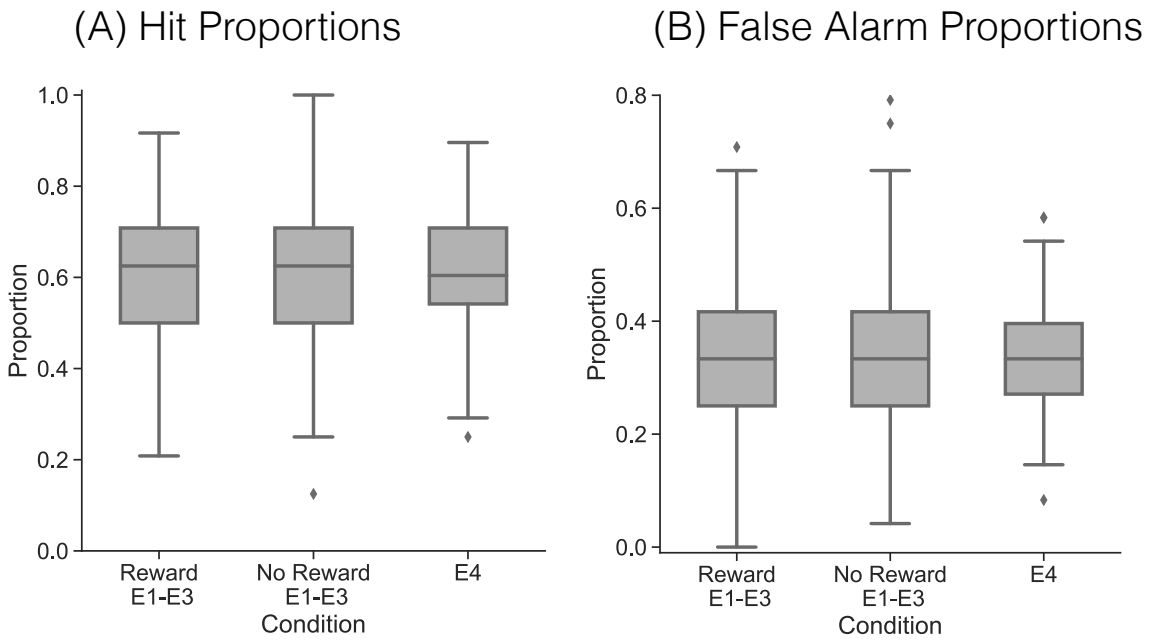

Figure 5.3: Influence of test phase reward on hit and false alarm rates. **A** Hit rate proportions for reward and no reward E1-E3 targets and E4 targets. We do not find a significant difference in hit rates for reward or no reward E1-E3 targets compared to E4 targets. **B** False alarm (FA) rate proportions for reward and no reward E1-E3 lures and E4 lures. We do not find a significant difference in FA rates for reward or no reward E1-E3 lures compared to E4 lures. Box-and-whisker plots show median (center line), upper and lower quartiles (box limits), 1.5x interquartile range (whiskers) and outliers (diamonds).

= 0.652, $\eta_p^2 = 0.007$) or a main effect of reward ($F_{1,122} = 0.394$, $p = 0.531$, $\eta_p^2 = 0.0032$). The interaction between experiment and reward was also not significant ($F_{2,122} = 2.126$, $p = 0.124$, $\eta_p^2 = 0.03$). We then conducted a second 2×3 mixed effects ANOVA with reward and experiment (E1, E2, E3) as factors and FA rate as the dependent variable (Figure 5.4B). We do not find a main effect of experiment ($F_{2,122} = 0.467$, $p = 0.628$, $\eta_p^2 = 0.0076$) or a main effect of reward ($F_{1,122} = 0.034$, $p = 0.853$, $\eta_p^2 = 0.0003$). The interaction between experiment and reward was also not significant ($F_{2,122} = 0.020$, $p = 0.980$, $\eta_p^2 = 0.0003$). Together, these findings suggest that test phase extrinsic reward following memory retrieval does not impact subsequent memory.

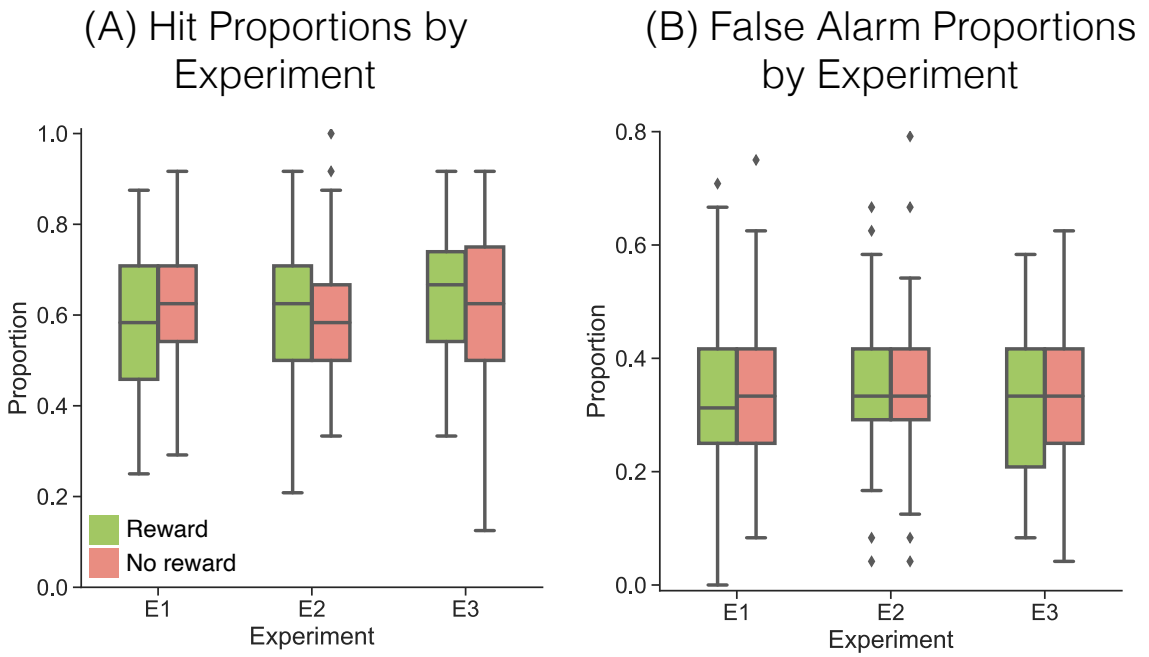

Figure 5.4: Influence of reward structure on hit and false alarm rates. **A** Hit rate proportions for reward and no reward targets by experiment (E1, E2, E3). We do not find a significant interaction between reward and experiment ($p = 0.124$). **B** FA rate proportions for reward and no reward lures by experiment (E1, E2, E3). We do not find a significant interaction between reward and experiment ($p = 0.980$). Reward is shown in green and no reward is shown in coral. Box-and-whisker plots show median (center line), upper and lower quartiles (box limits), 1.5x interquartile range (whiskers) and outliers (diamonds).

Retrieval practice vividness modulates subsequent memory.

Prior work has shown that vividness during retrieval practice impacts subsequent memory (Lee et al., 2018). Our inability to detect an impact of both retrieval practice and rewarded retrieval practice on later memory may be accounted for by variations in vividness. We did not provide rewards based on vividness ratings, meaning that strongly remembered (high vivid) and weakly remembered (low vivid) images were equally likely to have been rewarded. Reward may have differential effects across these two cases, whereby reward for a high vivid item may strengthen the stored representation and improve subsequent memory, but

reward for a low vivid item may not be able to strengthen the representation if insufficient information is retrieved. Alternatively, high vivid items may be so well remembered during practice that reward does not provide any additional reinforcement and instead, reward may serve to strengthen the representation of the low vivid items only.

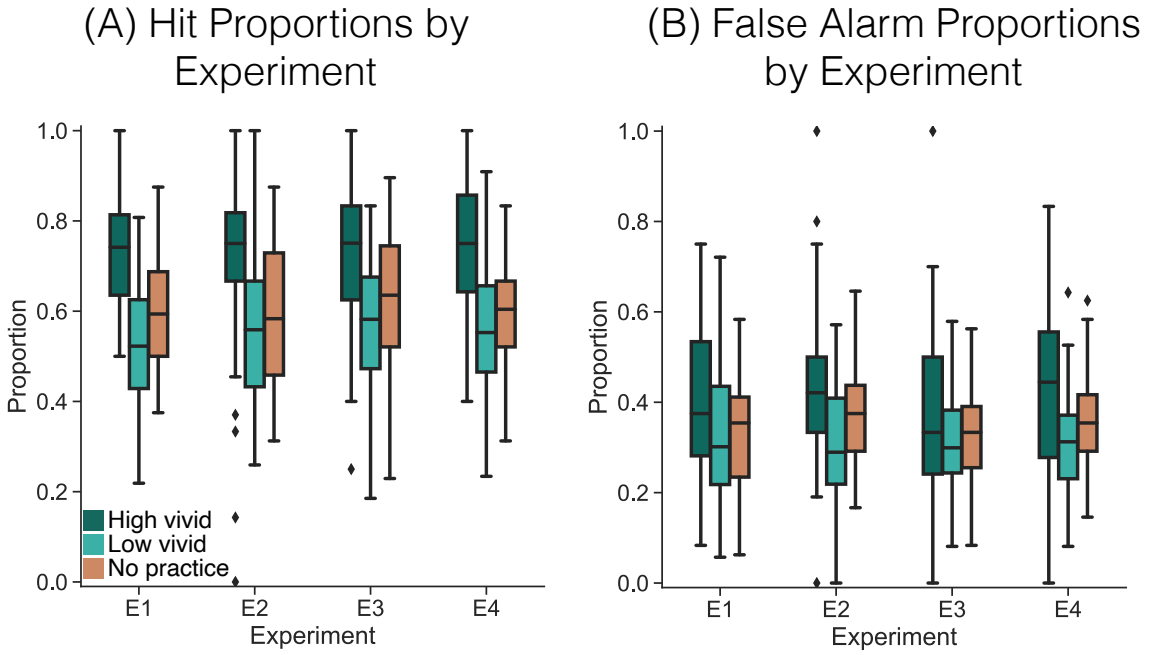

Figure 5.5: Influence of vividness on hit and false alarm rates. **A** Hit rate proportions for high (3 or 4) and low vividness (1 or 2) ratings during retrieval practice and not practiced targets for each experiment. We find a significant main effect of vividness driven by greater hit rates for the high vivid condition ($p < 0.0001$). **B** False alarm (FA) rate proportions for high and low vividness ratings during retrieval practice and not practiced lures for each experiment. We find a significant main effect of vividness driven by greater FA rates for the high vivid condition ($p < 0.0001$). High vividness ratings are shown in dark teal, low vividness ratings are shown in light teal, and not practiced is shown in orange. Box-and-whisker plots show median (center line), upper and lower quartiles (box limits), 1.5x interquartile range (whiskers) and outliers (diamonds).

Our first goal was to assess the impact of vividness during retrieval practice on subsequent memory. We specifically focused on vividness ratings during the first round of retrieval practice as the ratings during this round provide an estimate of initial memory strength that is not (yet) modulated by practice and/or reward. We

divided vividness ratings into high and low vividness groups (high = 3 and 4, low = 1 and 2). Our expectation is that greater retrieval practice vividness should be associated with both higher hit rates and higher FA rates. That is, we expect that images strongly remembered (high vivid) during the retrieval practice phase will be correctly endorsed as old when presented as targets during the final recognition test. However, to the extent that strong memories contain gist level information (Brainerd & Reyna, 2002; Lee et al., 2018), highly vivid remembering of gist-level information is likely to lead participants to erroneously endorse lures associated with highly vivid images as old. We conducted a 4×3 mixed effects ANOVA with experiment (E1, E2, E3, E4) and condition (high vivid, low vivid, no practice) as factors and hit rate as the dependent variable (Figure 5.5A). We do not find a main effect of experiment ($F_{3,162} = 0.17$, $p = 0.916$, $\eta_p^2 = 0.003$). We find a main effect of condition ($F_{2,324} = 131.8$, $p < 0.0001$, $\eta_p^2 = 0.45$) driven by greater hit rates for the high vivid condition (M = 0.74, SD = 0.17) compared to the no practice condition (M = 0.60, SD = 0.15; $t_{330} = 7.793$, $p < 0.0001$, $d = 0.858$) and greater hit rates for the no practice condition compared to the low vivid condition (M = 0.55, SD = 0.15; $t_{330} = 2.817$, $p = 0.0051$, $d = 0.3101$). The interaction between experiment and condition was not significant ($F_{6,324} = 0.648$, $p = 0.691$, $\eta_p^2 = 0.01$). We then conducted a second 4×3 mixed effects ANOVA with experiment (E1, E2, E3, E4) and condition (high vivid, low vivid, no practice) as factors and FA rate as the dependent variable (Figure 5.5B). We do not find a main effect of experiment ($F_{3,162} = 0.909$, $p = 0.438$, $\eta_p^2 = 0.02$). We find a main effect of condition ($F_{2,324} = 25.95$, $p < 0.0001$, $\eta_p^2 = 0.14$) driven by greater FA rates for the high vivid condition (M = 0.40, SD = 0.19) compared to the no practice condition (M = 0.35, SD = 0.12; $t_{330} = 3.120$, $p = 0.002$, $d = 0.3435$) and greater FA rates for the no practice condition compared to the low vivid condition (M = 0.31, SD = 0.14; $t_{330} = 2.504$, $p = 0.0128$, $d = 0.2756$). The interaction between

experiment and practice condition was also not significant ($F_{6,324} = 0.964$, $p = 0.449$, $\eta_p^2 = 0.02$). Together, these findings suggest that retrieval practice vividness ratings modulate subsequent memory.

5.5 Discussion

The aim of this study was to investigate how extrinsic reward following memory retrieval impacts subsequent memory. We conducted four independent recognition memory experiments in which we manipulated test phase reward delivery. We measured the impact of retrieval practice, reward, and vividness on a final recognition test. We find no effect of retrieval practice or reward on subsequent memory performance. However, we find that retrieval practice high vividness ratings were associated with greater hit rates and greater false alarm (FA) rates. Together, these findings suggest that retrieval practice and reward following memory retrieval may only be beneficial for strongly retrieved memories.

It is well established that repeated testing or retrieval practice facilitates memory (Karpicke & Roediger, 2008), however, we find no effect of retrieval practice on subsequent memory performance. This may be due to the format of the final test used in the current studies. Prior work has found inconclusive evidence regarding the impact of retrieval practice when the final test format is recognition with some studies reporting a benefit for repeated testing (Roediger & McDermott, 1995) and others not (Jones & Roediger, 1995). One well known theory that accounts for the benefit of retrieval practice is transfer-appropriate processing, which emphasizes that the same processes engaged during repeated testing or retrieval practice will be engaged during the final memory test (Morris et al., 1977; Roediger, 1990).

Potentially, as the retrieval practice phase in the current studies are not matched – practice was cued recall whereas the final test was recognition – this may also contribute to the lack of retrieval practice benefit in the present work.

We replicate prior work and find higher vividness ratings during retrieval practice are associated with greater hit rates and false alarm rates (Lee et al., 2018). We also find lower hit rates for items low in vividness. Our interpretation is that the consequences of retrieval vary depending on what information is reactivated. Higher hit rates for items rated high in vividness and lower hit rates for items rated low in vividness, may reflect the "zone of destruction". The zone of destruction posits that items that are moderately reactivated are weakened whereas items that are strongly reactivated are strengthened (Detre, Natarajan, Gershman, & Norman, 2013). Possibly, low vivid items in the present study were partially reactivated and therefore weakened by practice. These findings are comparable to the retrieval induced forgetting literature whereby retrieving a non-target item is thought to lead to suppression or inhibition of that item, impairing later memory (Anderson et al., 1994). Furthermore, impaired subsequent memory for low vivid items is consistent with prior work demonstrating that incomplete reinstatement via violation of context expectation weakens the memory representation and leads to forgetting (G. Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014). Taken together, vividness ratings during the retrieval practice phase may potentially be explained by whether the correct or incorrect image was reactivated.

We do not find evidence to suggest that rewards following a retrieval test impact subsequent memory. Our finding is consistent with prior behavioral work that also found no effect of potential reward during test (Castanheira et al., 2022). One possibility for this lack of effect may be due to insufficient rounds of retrieval prac-

tice as participants completed only two rounds of retrieval practice. Increasing the number of rounds or only rewarding trials in which participants are able to vividly recall the associated stimuli could modulate subsequent memory. Additionally, in the current study, every response had a 50% chance of receiving a reward in order to prevent participants from anticipating that a specific response (e.g. high vividness) or category (e.g. faces) would be rewarded because reward anticipation can influence participant decisions (Bowen, Marchesi, & Kensinger, 2020). However, reward prediction errors (RPEs) may drive reinforcement, independent of reward delivery, and modulate subsequent memory. RPEs occur when the predicted outcome deviates from what is received, a process that can drive learning (Schultz, Dayan, & Montague, 1997). As both positive and negative RPEs – receiving unexpected rewards or unexpected punishments – modulate neural signals and drive behavior (Zaghloul et al., 2009; Scimeca, Katzman, & Badre, 2016; Jang, Nassar, Dillon, & Frank, 2019; Ergo, De Loof, & Verguts, 2020; Rouhani & Niv, 2021), RPEs may be a potential mechanism for reinforcement. Future work will be needed to directly test this possibility.

Taken together, our findings indicate that the consequences of retrieval vary depending on what information is reactivated. A direction for future research will be to directly investigate the impact on memory of selectively rewarding items based on how those items are remembered. More broadly, we contribute to a growing body of literature characterizing the role of test phase reward on subsequent memory performance.

Chapter 6

General Discussion

In the previous chapters, I examined the neural mechanisms of episodic memory retrieval using recognition paradigms coupled with scalp electroencephalographic (EEG) recordings. My results suggest that temporal overlap can induce a retrieval state, an intrinsic reward signal occurs in response to successful retrieval and the strength of a retrieved memory modulates subsequent memory.

In Chapter 2, I used scalp EEG to identify a memory retrieval state. Participants completed a mnemonic state task and were instructed to either encode the present event or retrieve a past, overlapping event. The critical manipulation was the temporal distance between the first and second object, whereby the shorter the temporal distance between two objects, the greater their temporal contextual overlap. Interference can occur between experiences that overlap due to a trade-off between encoding the present event and retrieving the past event, which can lead to forgetting (Anderson, 2003; Kuhl et al., 2010). Using pattern classification analyses, I found that when two events overlap and are experienced nearby in

time, the memory system is biased toward a retrieval state. This induction of the retrieval state occurs independent from top-down demands to encode or retrieve and impairs subsequent memory for the past event. Critically, our neural results suggest that the retrieval state we observe is likely the result of a general retrieval mode (Rugg & Wilding, 2000), rather than a reflection of retrieval success or effort.

The results of this chapter demonstrate that a retrieval state can be engaged independent from the top-down demands to encode or retrieve, suggesting that bottom-up stimulus features may induce memory states. Prior work has shown that memory states predict subsequent memory (Long & Kuhl, 2019) and can influence behavior and decision-making (Duncan et al., 2012; Duncan & Shohamy, 2016). Specifically, behavioral evidence has suggested that brain states linger following a memory judgement and influence ongoing memory judgments (Patil & Duncan, 2018). To the extent that the retrieval state can be induced automatically and influence on-going processing, this state has the potential to widely influence cognition. It is necessary to be able to control both initiation and inhibition of the retrieval state to leverage this state when task-relevant and to switch out of this state when task-irrelevant. Future work is necessary to determine how the retrieval state may be initiated and controlled to better understand the consequences of the retrieval state on behavior.

In Chapter 3, I investigated the processes occurring prior to and following a memory response using response-locked theta power. Theta has been shown to support episodic memory and decision-making (Nyhus & Curran, 2010; Cavanagh & Frank, 2014). Specifically, theta power is greater for hits compared to correct rejections (CRs) after stimulus onset (Burgess & Gruzelier, 1997; Klimesch et al., 2000; Düzel et al., 2003) and greater following incorrect compared to correct responses

and following negative relative to positive outcomes (Mazaheri et al., 2009; Cohen, 2014b). Together, these findings suggest that theta power prior to a response may reflect memory processes and post-response theta power may reflect a feedback signal. I replicated previous work and found greater theta power for hits than CRs. Following the memory response, I found that the theta power dissociation 'flips' such that theta power is greater for CRs than hits, potentially reflecting a positive feedback signal selectively for hits.

In Chapter 4, I measured post-retrieval positive feedback evidence in response to successful retrieval. I conducted two recognition memory experiments and manipulated the test phase goals. In a typical recognition experiment, the participants' goal to identify old items is confounded with successful retrieval. Therefore, test phase reward signals could reflect the attainment of a task goal (Han et al., 2010). However, the abundance of evidence that finds reward system activity during hits compared to CRs, in the absence of extrinsic reward (Spaniol et al., 2009; Clos et al., 2015), suggests that successful retrieval may be intrinsically rewarding (Speer et al., 2014). I replicate the findings in Chapter 3, and find theta power decreases selectively for hits relative to CRs, regardless of task goals. I used an independently validated feedback classifier to measure feedback evidence in each experiment. I find that regardless of task goals, following a response, there is greater positive feedback evidence for hits than CRs, suggesting that successful retrieval is intrinsically rewarding.

The results of Chapters 3 and 4 suggest that there is a positive feedback signal in response to successful retrieval. How might intrinsic reward following retrieval reinforce mnemonic strategies? Mnemonic strategies are techniques used in an attempt to promote memory, such as grouping items based on a shared meaning.

Whereas mnemonic strategies can influence the ability to remember, the neural mechanisms underlying the reinforcement of these strategies are unknown. It might be through an intrinsic feedback signal that strategies are reinforced and promote sustained improvements in memory performance. According to the binding of items and context (BIC) model, during study the hippocampus binds an item to its spatiotemporal context, where context includes the mnemonic strategy used to encode the item (Diana et al., 2007). Furthermore, according to retrieved context theory (Polyn & Kahana, 2007), retrieval of a study item leads to retrieval of its associated context. Therefore, the mnemonic strategy encoded with an item should be retrieved when the item is later remembered (Polyn et al., 2012). Accordingly, if reward accompanies successful retrieval, retrieval of the item should reinforce the mnemonic strategy that was initially used to encode the item. However, not all mnemonic strategies are created equal. One consequence of intrinsic reward during retrieval is that any mnemonic strategy associated with that item will be reinforced, even if that strategy is suboptimal. Therefore, it is insufficient to learn effective mnemonic strategies, individuals also need to avoid ineffective mnemonic strategies. A critical next step is to establish the extent to which intrinsic reward reinforces memory processes and impacts subsequent memory.

In Chapter 5, I used extrinsic reward following memory retrieval to investigate the extent to which reward reinforces the contents of retrieval and impacts subsequent memory. I conducted four behavioral experiments and manipulated test phase reward delivery. Across all four experiments, participants studied word-image pairs and then completed two rounds of retrieval practice, followed by a final recognition test. The structure of reward delivery varied during the retrieval practice phase in which participants made vividness judgments. It is well es-

tablished that study phase reward anticipation facilitates memory formation and subsequent memory (Loftus & Wickens, 1970; Marini et al., 2011). However, memory reinforcement may be better accomplished through direct reward of what is retrieved, rather than through study phase manipulation of potential future reward. I find no impact of reward on subsequent memory performance, but I did find an impact of vividness on subsequent memory, suggesting that the benefit of reward may be dependent on the strength of the memory that is retrieved.

The results of this chapter demonstrate that retrieval practice improves recognition for highly vivid items, but impairs recognition for items low in vividness. These findings may reflect the "zone of destruction" whereby items that are moderately reactivated are weakened whereas items that are strongly reactivated are strengthened (Detre et al., 2013). Potentially, as a result of partial reactivation, low vivid items in the present study may be weakened by practice. These findings are in line with the retrieval induced forgetting literature whereby retrieving a non-target item is thought to lead to suppression or inhibition of that item, impairing later memory (Anderson et al., 1994). Furthermore, my findings of impaired later memory for items remembered with low vividness is consistent with prior work demonstrating that incomplete reinstatement via violation of context expectation weakens the memory representation and leads to forgetting (G. Kim et al., 2014). Expectation violation, or reward prediction errors (RPEs), occur when the predicted outcome deviates from what is received. RPEs may be a potential mechanism for reinforcement as both positive and negative RPEs – receiving unexpected rewards or unexpected punishments – modulate neural signals and drive behavior (Zaghloul et al., 2009; Scimeca et al., 2016; Jang et al., 2019; Ergo et al., 2020; Rouhani & Niv, 2021). Therefore, test phase RPEs may have a differential impact

on memory depending on what information is reactivated. Future work should work to identify whether RPEs or extrinsic reward – regardless of expectations – promote subsequent memory.

In conclusion, the work in my dissertation provides a substantial contribution to our understanding of episodic retrieval mechanisms. Formation and subsequent retrieval of memories are essential for daily decision making. Therefore, elucidating the neural mechanisms that support retrieval are critical across cognition. I believe that the insights my work has provided will lead to a better understanding of how people successfully retrieve memories and how to improve memory performance.

References

Achim, A. M., & Lepage, M. (2005). Dorsolateral prefrontal cortex involvement in memory post-retrieval monitoring revealed in both item and associative recognition tests. *NeuroImage*, *24*, 1113-1121.

Adcock, A. R., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. (2006). Reward-motivated learning: Mesolimbic activation precedes memory formation. *Neuron*, *50*, 507-517.

Addante, R. J., Ranganath, C., & Yonelinas, A. P. (2012). Examining ERP correlates of recognition memory: Evidence of accurate source recognition without recollection. *NeuroImage*, *61*(1), 439-450.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*, 415-445.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063-1087.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review*, *102*(1), 68-100.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & S. Patrick (Eds.),

Contemporary developments in mathematical psychology: I. learning, memory and thinking. (p. 243-293). W.H. Freeman.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300.

Besson, G., Ceccaldi, M., Didic, M., & Barbeau, E. J. (2012). The speed of visual recognition memory. *Visual Cognition*, *20*(10), 1131-1152.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, H. L. M, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (p. 55-64). New York: Worth Publishers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (p. 185-205). Cambridge, MA: MIT Press.

Bowen, H. J., Marchesi, M. L., & Kensinger, E. A. (2020). Reward motivation influences response bias on a recognition memory task. *Cognition*, *203*, 1-13.

Brainerd, C., Bialer, D., Chang, M., & Upadhyay, P. (2021). A fundamental asymmetry in human memory: Old not equal not-new and new not equal not-old. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(12), 1850-1867.

Brainerd, C., & Reyna, V. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, *11*(5), 164-169.

Brunec, I. K., Robin, J., Olsen, R. K., Moscovitch, M., & Barense, M. D. (2020). Integration and differentiation of hippocampal memory traces. *Neuroscience and Biobehavioral Reviews*, *118*, 196-208.

Burgess, A. P., & Gruzelier, J. H. (1997). Short duration synchronization of human theta rhythm during recognition memory. *NeuroReport*, *8*(4), 1039-1042.

Castanheira, K. d. S., Lalla, A., Ocampo, K., Otto, A. R., & Sheldon, S. (2022). Reward at encoding but not retrieval modulates memory for detailed events. *Cognition*, *219*(104957), 1-10.

Castellanos, N. P., & Makarov, V. A. (2006). Recovering EEG brain signals: Artifact suppression with wavelet enhanced independent component analysis. *Journal of Neuroscience Methods*, *158*(2), 300-312.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414-421.

Cavanagh, J. F., Frank, M. J., Klein, T. J., & Allen, J. J. (2010). Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *NeuroImage*, *49*(4), 3198-3209.

Cavanagh, J. F., Zambrano-Vazquez, L., & Allen, J. J. (2012). Theta lingua franca: A common mid-frontal substrate for action monitoring processes. *Psychophysiology*, *49*(2), 220-238.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual review of psychology*, *62*, 73-101.

Clos, M., Schwarze, U., Gluth, S., Bunzeck, N., & Sommer, T. (2015). Goal- and retrieval-dependent activity in the striatum during memory recognition. *Neuropsychologia*, *72*, 1-11.

Cohen, M. X. (2014a). *Analyzing neural time series data: theory and practice*. MIT press.

Cohen, M. X. (2014b). A neural microcircuit for cognitive conflict detection and signaling. *Trends in Neurosciences*, *37*(9), 480-490.

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates

feedback-related negativity and EEG spectra. *NeuroImage*, *35*, 968-978.

Crone, N. E., Miglioretti, D. L., Gordon, B., Sieracki, J. M., Wilson, M. T., Uematsu, S., & Lesser, R. P. (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. i. alpha and beta event-related desynchronization. *Brain*, *121*(12), 2271-2299.

Crowder, R. (1976). *Principles of learning and memory*. Lawrence Erlbaum.

Cruse, D., & Wilding, E. L. (2009). Prefrontal cortex contributions to episodic retrieval monitoring and evaluation. *Neuropsychologia*, *47*, 2779-2789.

Cruse, D., & Wilding, E. L. (2011). Temporally and functionally dissociable retrieval processing operations revealed by event-related potentials. *Neuropsychologia*, *49*, 1751-1760.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, *28*(6), 923-938.

Curran, T. (2004). Effects of attention and confidence on the hypothesized ERP correlates of recollection and familiarity. *Neuropsychologia*, *42*, 1088-1106.

Curran, T., & Cleary, A. M. (2003). Using ERPs to dissociate recollection from familiarity in picture recognition. *Cognitive Brain Research*, *15*, 191-205.

Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 2-17.

Curran, T., & Hancock, J. (2007). The FN400 indexes familiarity-based recognition of faces. *NeuroImage*, *36*(2), 464-471.

Danker, J. F., & Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, *136*(1), 87-102.

Detre, G. J., Natarajan, A., Gershman, S. J., & Norman, K. A. (2013). Moderate

levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia*, *51*(12), 2371-2388.

de Zubicaray, G. I., McMahon, K. L., Eastburn, M. M., Finnigan, S., & Humphreys, M. S. (2005). fMRI evidence of word frequency and strength effects in recognition memory. *Cognitive Brain Research*, *24*, 587-598.

de Zubicaray, G. I., Miozzo, M., Johnson, K., Schiller, N. O., & McMahon, K. L. (2011). Independent distractor frequency and age-of-acquisition effects in picture–word interference: fMRI evidence for post-lexical and lexical accounts according to distractor type. *Journal of Cognitive Neuroscience*, *24*(2), 482-495.

Diana, R. A., Vilberg, K. L., & Reder, L. M. (2005). Identifying the ERP correlate of a recognition memory search attempt. *Brain Research. Cognitive Brain Research*, *24*(3), 674-684.

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *TRENDS in Cognitive Sciences*, *11*(9), 379-386.

Dickerson, K. C., & Adcock, A. R. (2018). Stevens' handbook of experimental psychology and cognitive neuroscience. In (Fourth ed., chap. Motivation and Memory). John Wiley & Sons, Inc.

Duncan, K. D., Sadanand, A., & Davachi, L. (2012). Memory's penumbra: Episodic memory decisions induce lingering mnemonic biases. *Science*, *337*(6093), 485-487.

Duncan, K. D., & Shohamy, D. (2016). Memory states influence value-based decisions. *Journal of Experimental Psychology: General*, *145*(11), 1420-1426.

Düzel, E., Habib, R., Schott, B., Schoenfeld, A., Lobaugh, N., McIntosh, A., ... Heinze, H. (2003). A multivariate, spatiotemporal analysis of electromagnetic

time-frequency data of recognition memory. *NeuroImage*, *18*(2), 185-197.

Duzel, E., Neufang, M., & Heinze, H.-J. (2005). The oscillatory dynamics of recognition memory and its relationship to event-related responses. *Cerebral Cortex*, *15*, 1992-2002.

Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron*, *44*, 109-120.

El-Kalliny, M. M., Wittig, J. H. J., Sheehan, T. C., Sreekumar, V., Inati, S. K., & Zaghloul, K. A. (2019). Changing temporal context in human temporal lobe promotes memory of distinct episodes. *Nature Communications*, *10*(203).

Elliott, B. L., Blais, C., McClure, S. M., & Brewer, G. A. (2020). Neural correlates underlying the effect of reward value on recognition memory. *NeuroImage*, *206*, 1-9.

Ergo, K., De Loof, E., & Verguts, T. (2020). Reward prediction error and declarative memory. *Trends in Cognitive Sciences*, *24*(5), 388-397.

Evans, L. H., Williams, A. N., & Wilding, E. L. (2015). Electrophysiological evidence for retrieval mode immediately after a task switch. *NeuroImage*, *108*, 435-440.

Fliessbach, K., Weis, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *NeuroImage*, *32*, 1413-1421.

Folkerts, S., Rutishauser, U., & Howard, M. W. (2018). Human episodic memory retrieval is accompanied by a neural contiguity effect. *The Journal of Neuroscience*, *38*(17), 4200-4211.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *The Journal of Neuroscience*, *35*(2), 485-494.

Frank, M. J., Woroch, B. S., & Curran, T. (2005). Error-related negativity predicts

reinforcement learning and conflict biases. *Neuron*, *47*, 495-501.

Friedman, D., & Johnson Jr., R. (2000). Event-related potential (ERP) studies of memory encoding and retrieval: A selective review. *Microscopy Research and Technique*, *51*, 6-28.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*, 375-399.

Fries, P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, *9*(10), 474-480.

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385-390.

Gimbel, S. I., & Brewer, J. B. (2011). Reaction time, memory strength, and fMRI activity during memory retrieval: Hippocampus and default network are differentially responsive during recollection and familiarity judgments. *Cognitive Neuroscience*, *2*(1), 19-23.

Glazer, J. E., Kelley, N. J., Pornpattananangkul, N., Mittal, V. A., & Nusslock, R. (2018). Beyond the FRN: Broadening the time-course of EEG and ERP components implicated in reward processing. *International Journal of Psychophysiology*, *132*, 184-202.

Gordon, A. M., Rissman, J., Kiani, R., & Wagner, A. D. (2014). Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. *Cerebral Cortex*, *24*, 3350-3364.

Gruber, T., Tsivilis, D., Giabbiconi, C.-M., & Müller, M. M. (2008). Induced electroencephalogram oscillations during source memory: familiarity is reflected

in the gamma band, recollection in the theta band. *Journal of Cognitive Neuroscience*, *20*(6), 1043-1053.

Guan, Q., Ma, L., Chen, Y., Luo, Y., & He, H. (2023). Midfrontal theta phase underlies evidence accumulation and response thresholding in cognitive control. *Cerebral Cortex*, *33*, 8967-8979.

Han, S., Huettel, S. A., Raposo, A., Adcock, A. R., & Dobbins, I. G. (2010). Functional significance of striatal responses during episodic decisions: Recovery or goal attainment? *The Journal of Neuroscience*, *30*(13), 4767-4775.

Hanczakowski, M., Zawadzka, K., & Macken, B. (2015). Continued effects of context reinstatement in recognition. *Memory & Cognition*, *43*(5), 788-797.

Hasselmo, M. E. (2005). What is the function of hippocampal theta rhythm?—- linking behavioral data to phasic properties of field potential and unit recording data. *Hippocampus*, *15*(7), 936-949.

Hasselmo, M. E., Bodelon, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, *14*(4), 793-817.

Hasselmo, M. E., & McGaughy, J. (2004). High acetylcholine levels set circuit dynamics for attention and encoding and low acetylcholine levels set dynamics for consolidation. In (Vol. 145, p. 207-231). Progress in Brain Research.

Hasselmo, M. E., & Stern, C. E. (2014). Theta rhythm and the encoding and retrieval of space and time. *NeuroImage*, *85*, 656-666.

Hayama, H. R., Johnson, J. D., & Rugg, M. D. (2008). The relationship between the right frontal old/new ERP effect and post-retrieval monitoring: Specific or non-specific? *Neuropsychologia*, *46*, 1211-1223.

Healey, M. K., Crutchley, P., & Kahana, M. J. (2014). Individual differences in memory search and their relation to intelligence. *Journal of Experimental Psychology:*

General, *143*(4), 1553-1569.

Henson, R. N., Hornberger, M., & Rugg, M. D. (2005). Further dissociating the processes involved in recognition memory: An fMRI study. *Journal of Cognitive Neuroscience*, *17*(7), 1058-1073.

Herron, J., & Wilding, E. (2004). An electrophysiological dissociation of retrieval mode and retrieval orientation. *NeuroImage*, *22*, 1554-1562.

Herweg, N. A., Apitz, T., Leicht, G., Mulert, C., Fuentemilla, L., & Bunzeck, N. (2016). Theta-alpha oscillations bind the hippocampus, prefrontal cortex, and striatum during recollection: Evidence from simultaneous EEG–fMRI. *The Journal of Neuroscience*, *36*(12), 3579-3587.

Herweg, N. A., Solomon, E. A., & Kahana, M. J. (2020). Theta oscillations in human memory. *Trends in Cognitive Sciences*, *24*(3), 208-227.

Hidi, S. (2016). Revisiting the role of rewards in motivation and learning: implications of neuroscientific research. *Educational Psychology Review*, *28*, 61-93.

Hill, P. F., Horne, E. D., Koen, J. D., & Rugg, M. D. (2021). Transcranial magnetic stimulation of right dorsolateral prefrontal cortex does not affect associative retrieval in healthy young or older adults. *Neuroimage: Reports*, *1*, 1-9.

Hornberger, M., Rugg, M. D., & Henson, R. N. (2006a). ERP correlates of retrieval orientation: Direct versus indirect memory tasks. *Brain Research*, *1071*, 124-136.

Hornberger, M., Rugg, M. D., & Henson, R. N. (2006b). fMRI correlates of retrieval orientation. *Neuropsychologia*, *44*, 1425-1436.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269-299.

Howard, M. W., Viskontas, I. V., Shankar, K. H., & Fried, I. (2012). Ensembles of human mtl neurons "jump back in time" in response to a repeated stimulus.

Hippocampus, *22*, 1833-1847.

Jacobs, J., Hwang, G., Curran, T., & Kahana, M. J. (2006). EEG oscillations and recognition memory: Theta correlates of memory retrieval and decision making. *NeuroImage*, *32*(2), 978-987.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513-541.

Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction errors during decision making strengthen memory encoding. *Nature human behavior*, *3*(7), 719-732.

Johansson, M., & Mecklinger, A. (2003). The late posterior negativity in ERP studies of episodic memory: Action monitoring and retrieval of attribute conjunctions. *Biological Psychology*, *64*, 91-117.

Johnson, J. D., Price, M. H., & Leiker, E. K. (2015). Episodic retrieval involves early and sustained effects of reactivating information from encoding. *NeuroImage*, *106*, 300-310.

Jones, T. C., & Roediger, H. L. (1995). The experiential basis of serial position effects. *European Journal of Cognitive Psychology*, *7*(1), 65-80.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103-109.

Kahana, M. J. (2012). *Foundations in human memory*. Oxford University Press.

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157-163.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966-968.

Kay, K., & Frank, L. M. (2019). Three brain states in the hippocampus and cortex. *Hippocampus*, *29*, 184-238.

Kelley, W. M., Miezin, F. M., McDermott, K. B., Buckner, R. L., Raichle, M. E., Cohen, N. J., ... Petersen, S. E. (1998). Hemispheric specialization in human dorsal frontal cortex and medial temporal lobe for verbal and nonverbal memory encoding. *Neuron*, *20*(5), 927-936.

Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences*, *111*(24), 8997-9002.

Kim, H. (2011). Neural activity that predicts subsequent memory and forgetting: A meta-analysis of 74 fMRI studies. *NeuroImage*, *54*(3), 2446-2461.

Kim, H. (2013). Differential neural activity in the recognition of old versus new events: An activation likelihood estimation meta-analysis. *Human Brain Mapping*, *34*, 814-836.

Klimesch, W., Doppelmayr, M., Schimke, H., & Ripper, B. (1997). Theta synchronization and alpha desynchronization in a memory task. *Psychophysiology*, *34*, 169-176.

Klimesch, W., Doppelmayr, M., Schwaiger, J., Winkler, T., & Gruber, W. (2000). Theta oscillations and the ERP old/new effect: Independent phenomena? *Clinical Neurophysiology*, *111*(5), 781-793.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558-578.

Kota, S., Rugg, M. D., & Lega, B. C. (2020). Hippocampal theta oscillations support successful associative memory formation. *The Journal of Neuroscience*, *40*(49), 9507-9518.

Kuhl, B. A., & Chun, M. M. (2014). Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *The Journal of Neuroscience*, *34*(23),

8051-8060.

Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences*, *108*(14), 5903-5908.

Kuhl, B. A., Shah, A. T., DuBrow, S., & Wagner, A. D. (2010). Resistance to forgetting associated with hippocampus- mediated reactivation during new learning. *Nature Neuroscience*, *13*(4), 501-508.

Lee, H., Samide, R., Richter, F. R., & Kuhl, B. A. (2018). Decomposing parietal memory reactivation to predict consequences of remembering. *Cerebral Cortex*, *29*, 1-14.

Loftus, G. R., & Wickens, T. D. (1970). Effect of incentive on storage and retrieval processes. *Journal of Experimental Psychology*, *85*(1), 141-147.

Lohnas, L. J., & Kahana, M. J. (2014). Compound cuing in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 12-24.

Long, N. M. (2023). The intersection of the retrieval state and internal attention. *Nature Communications*, *14*(3861), 1-13.

Long, N. M., & Kahana, M. J. (2015). Successful memory formation is driven by contextual encoding in the core memory network. *NeuroImage*, *119*, 332-337.

Long, N. M., & Kahana, M. J. (2017). Modulation of task demands suggests that semantic processing interferes with the formation of episodic associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(2), 167-176.

Long, N. M., & Kahana, M. J. (2019). Hippocampal contributions to serial-order memory. *Hippocampus*, 1-8.

Long, N. M., & Kuhl, B. A. (2018). Bottom-up and top-down factors differentially influence stimulus representations across large-scale attentional networks.

The Journal of Neuroscience, *38*(10), 2495-2504.

Long, N. M., & Kuhl, B. A. (2019). Decoding the tradeoff between encoding and retrieval to predict memory for overlapping events. *NeuroImage*, *201*.

Long, N. M., & Kuhl, B. A. (2021). Cortical representations of visual stimuli shift locations with changes in memory states. *Current Biology*, *31*, 1119-1126.

Long, N. M., Öztekin, I., & Badre, D. (2010). Separable prefrontal cortex contributions to free recall. *The Journal of Neuroscience*, *30*(33), 10967-10976.

Long, N. M., Sperling, M. R., Worrell, G. A., Davis, K. A., Gross, R. E., Lega, B. C., ... Kahana, M. J. (2017). Contextually mediated spontaneous retrieval is specific to the hippocampus. *Current Biology*, *27*, 1-6.

Luft, C. D. B. (2014). Learning from feedback: The neural mechanisms of feedback processing facilitating better performance. *Behavioural Brain Research*, *261*, 356-368.

Luu, P., Tucker, D. M., & Makeig, S. (2004). Frontal midline theta and the error-related negativity: Neurophysiological mechanisms of action regulation. *Clinical Neurophysiology*, *115*, 1821-1835.

Manning, J. R., Polyn, S. M., Baltuch, G. H., Litt, B., & Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences*, *108*(31), 12893-12897.

Manns, J. R., Howard, M. W., & Eichenbaum, H. (2007). Gradual changes in hippocampal activity support remembering the order of events. *Neuron*, *56*(3), 530-540.

Marco-Pallarés, J., Cucurell, D., Cunillera, T., García, R., Andrés-Pueyo, A., Münte, T. F., & Rodríguez-Fornells, A. (2008). Human oscillatory activity associated to reward processing in a gambling task. *Neuropsychologia*, *46*(1), 241-248.

Marini, F., Marzi, T., & Viggiano, M. P. (2011). "wanted!" the effects of reward on face recognition: electrophysiological correlates. *Cogn Affect Behav Neurosci*, *11*, 627-643.

Mas-Herrero, E., Ripollés, P., HajiHosseini, A., Rodríguez-Fornells, A., & Marco-Pallarés, J. (2015). Beta oscillations and reward processing: Coupling oscillatory activity and hemodynamic responses. *NeuroImage*, *119*, 13-19.

Mazaheri, A., Nieuwenhuis, I. L., Dijk, H. v., & Jensen, O. (2009). Prestimulus alpha and mu activity predicts failure to inhibit motor responses. *Human Brain Mapping*, *30*, 1791-1800.

McDermott, K. B. (2006). Paradoxical effects of testing: repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, *34*(2), 261-267.

McGeoch, J. A. (1942). *The psychology of human learning: an introduction*. Longmans.

Mecklinger, A. (2000). Interfacing mind and brain: A neurocognitive model of recognition memory. *Psychophysiology*, *37*, 565-582.

Meeter, M., Murre, J., & Talamini, L. (2004). Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus*, *14*, 722-741.

Moore, I. L., & Long, N. M. (2024). Semantic associations restore neural encoding mechanisms. *Learning & Memory*, *31*(3), 1-13.

Morris, D. C., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519-533.

Moscovitch, M. (1994). Cognitive resources and dual-task interference effects at retrieval in normal people: The role of the frontal lobes and medial temporal cortex. *Neuropsychology*, *8*(4), 524-534.

Mulder, M. J., & van Maanen, L. (2013). Are accuracy and reaction time affected via different processes? *PLoS One*, *8*(11), 1-8.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving bold activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636-2643.

Nelson, D. L., Zhang, N., & McKinney, V. M. (2001). The ties that bind what is known to the recognition of what is new. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1147-1159.

Nolan, H., Whelan, R., & Reilly, R. (2010). Faster: Fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods*, *192*(1), 152-162.

Nyhus, E., & Curran, T. (2010). Functional role of gamma and theta oscillations in episodic memory. *Neuroscience and Biobehavioral Reviews*, *34*, 1023-1035.

Ostlund, B., Donoghue, T., Anaya, B., Gunther, K. E., Karalunas, S. L., Voytek, B., & Pérez-Edgar, K. E. (2022). Spectral parameterization for studying neurodevelopment: How and why. *Developmental Cognitive Neuroscience*, *54*(101073), 1-14.

Patil, A., & Duncan, K. D. (2018). Lingering cognitive states shape fundamental mnemonic abilities. *Psychological Science*, *29*(1), 45-55.

Perfect, T. J., Stark, L.-J., Tree, J. J., Moulin, C. J., Ahmed, L., & Hutter, R. (2004). Transfer appropriate forgetting: The cue-dependent nature of retrieval-induced forgetting. *Journal of Memory and Language*, *51*, 399-417.

Pinner, J. F., & Cavanagh, J. F. (2017). Frontal theta accounts for individual differences in the cost of conflict on decision making. *Brain Research*, *1672*, 73-80.

Polyn, S. M., & Kahana, M. J. (2007). Memory search and the neural representation of context. *TRENDS in Cognitive Sciences*, *12*(1), 24-30.

Polyn, S. M., Kragel, J. E., Morton, N. W., McCluey, J. D., & Cohen, Z. D. (2012). The neural dynamics of task context in free recall. *Neuropsychologia*, *50*, 447-457.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129-156.

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*, 816-847.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 163-178.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873-922.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260-281.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59-83.

Richter, F. R., Chanales, A. J. H., & Kuhl, B. A. (2016). Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. *NeuroImage*, *124*(Pt A), 323-335.

Ries, S. K., Dronkers, N. F., & Knight, R. T. (2016). Choosing words: Left hemisphere, right hemisphere, or both? Perspective on the lateralization of word retrieval. *Annals of the New York Academy of Sciences*, *1369*(1), 111-131.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*(9), 1043-1056.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803-814.

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(2634), 1-11.

Rouhani, N., & Niv, Y. (2021). Signed and unsigned reward prediction errors dynamically enhance learning and memory. *eLife*, *10*(e61077), 1-28.

Rugg, M. D., Henson, R. N., & Robb, W. G. (2003). Neural correlates of retrieval processing in the prefrontal cortex during recognition and exclusion tasks. *Neuropsychologia*, *41*, 40-52.

Rugg, M. D., Ruth, M. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., & Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature*, *392*, 595-598.

Rugg, M. D., & Wilding, E. L. (2000). Retrieval processing and episodic memory. *Trends in Cognitive Sciences*, *4*(3), 108-115.

Satterthwaite, T. D., Ruparel, K., Elliott, M. A., Gerraty, R. T., Calkins, M. E., Hakonarson, H., ... Wolf, D. H. (2012). Being right is its own reward: Load and performance related ventral striatum activation to correct responses during a working memory task in youth. *NeuroImage*, *61*, 723-729.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical*

transactions of the Royal Society of London. Series B, Biological sciences, *362*(1481), 773-786.

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, *1*, 1-8.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schwarze, U., Bingel, U., Badre, D., & Sommer, T. (2013). Ventral striatal activity correlates with memory confidence for old- and new-responses in a difficult recognition test. *PLoS One*, *8*(3), 1-7.

Scimeca, J. M., Katzman, P. L., & Badre, D. (2016). Striatal prediction errors support dynamic control of declarative memory decisions. *Nature Communications*, *7*(13061), 1-15.

Sederberg, P. B., Gauthier, L. V., Terushkin, V., Miller, J. F., Barnathan, J. A., & Kahana, M. J. (2006). Oscillatory correlates of the primacy effect in episodic memory. *NeuroImage*, *32*, 1422-1431.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(4), 893-912.

Sederberg, P. B., Miller, J. F., Howard, M. W., & Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition*, *38*(6), 689-699.

Senftleben, U., & Scherbaum, S. (2021). Mid-frontal theta during conflict in a value-based decision task. *Journal of Cognitive Neuroscience*, *33*(10), 2109-2131.

Shenhav, A., Straccia, M. A., Musslick, S., Cohen, J. D., & Botvinick, M. M. (2018). Dissociable neural mechanisms track evidence accumulation for selection of attention versus action. *Nature Communications*, *9*(2485), 1-10.

Shigemune, Y., Tsukiura, T., Nouchi, R., Kambara, T., & Kawashima, R. (2017).

Neural mechanisms underlying the reward-related enhancement of motivation when remembering episodic memories with high difficulty. *Human Brain Mapping*, *38*, 3428-3443.

Smith, D. E., & Long, N. M. (2024). *Top-down task goals induce the retrieval state.* (BioRxiv. https://www.biorxiv.org/content/10.1101/2024.03.04.583353v1)

Smith, D. E., Moore, I. L., & Long, N. M. (2022). Temporal context modulates encoding and retrieval of overlapping events. *The Journal of Neuroscience*, *42*(14), 3000-3010.

Smith, D. E., Wheelock, J. R., & Long, N. M. (2024). *Response-locked theta dissociations reveal potential feedback signal following successful retrieval.* (BioRxiv. https://www.biorxiv.org/content/10.1101/2024.01.11.575166v1)

Smith, S. M., Handy, J. D., Hernandez, A., & Jacoby, L. L. (2018). Context specificity of automatic influences of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(10), 1501-1513.

Spaniol, J., Davidson, P. S., Kim, A. S., Han, H., Moscovitch, M., & Grady, C. L. (2009). Event-related fMRI studies of episodic encoding and retrieval: Meta-analyses using activation likelihood estimation. *Neuropsychologia*, *47*, 1765-1779.

Speer, M. E., Bhanji, J. P., & Delgado, M. R. (2014). Savoring the past: Positive memories evoke value representations in the striatum. *Neuron*, *84*, 1-10.

Spitzer, B., Gloel, M., Schmidt, T. T., & Blankenburg, F. (2014). Working memory coding of analog stimulus properties in the human prefrontal cortex. *Cerebral Cortex*, *24*, 2229-2236.

Tang, Y.-Y., Rothbart, M. K., & Posner, M. I. (2012). Neural correlates of establishing, maintaining, and switching brain states. *Trends in Cognitive Sciences*, *16*(6), 330-337.

Trujillo, L. T., & Allen, J. J. (2007). Theta EEG dynamics of the error-related negativity. *Clinical Neurophysiology*, *118*, 645-668.

Tuladhar, A. M., Huurne, N. t., Schoffelen, J.-M., Maris, E., Oostenveld, R., & Jensen, O. (2007). Parieto-occipital sources account for the increase in alpha activity with working memory load. *Human Brain Mapping*, *28*, 785-792.

Tulving, E. (1972). *Episodic and semantic memory* (E. Tulving & W. Donaldson, Eds.). Academic Press, Inc.

Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press.

Tulving, E. (1985). How many memory systems are there? *American Psychologist*, *40*(4), 385-398.

Tulving, E. (1993). What is episodic memory? *Current Directions in Psychological Science*, *2*(3), 67-70.

Uncapher, M. R., Boyd-Meredith, J. T., Chow, T. E., Rissman, J., & Wagner, A. D. (2015). Goal-directed modulation of neural memory patterns: Implications for fMRI-based memory detection. *The Journal of Neuroscience*, *35*(22), 8531-8545.

Underwood, B. (1948). Retroactive and proactive inhibition after five and forty-eight hours. *Journal of Experimental Psychology*, *38*(1), 29-38.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, *35*(2), 254-262.

Vigneau, M., Beaucousin, V., Hervé, P.-Y., Jobard, G., Petit, L., Crivello, F., ... Tzourio-Mazoyer, N. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. *NeuroImage*, *54*, 577-593.

Voss, J., & Paller, K. (2008). Neural substrates of remembering: Electroencephalographic studies: Learning and memory: A comprehensive reference. In

J. H. Byrne (Ed.), *Memory systems.* Elsevier.

Wang, D., Buckner, R. L., & Liu, H. (2014). Functional specialization in the human brain estimated by intrinsic hemispheric interaction. *The Journal of Neuroscience*, *34*(37), 12341-12352.

Wang, F., & Diana, R. A. (2017). Temporal context in human fMRI. *Current Opinion in Behavioral Sciences*, *17*, 57-64.

Wang, L., Gu, Y., Zhao, G., & Chen, A. (2020). Error-related negativity and error awareness in a go/no-go task. *Scientific Reports*, *10*(4026), 1-12.

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society Open Science*, *3*(150670), 1-17.

Wilding, E. L., & Rugg, M. D. (1996). An event-related potential study of recognition memory with and without retrieval of source. *Brain*, *119*, 889-905.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152-176.

Wolosin, S. M., Zeithamova, D., & Preston, A. R. (2012). Reward modulation of hippocampal subfield activation during successful associative encoding and retrieval. *Journal of Cognitive Neuroscience*, *24*(7), 1532-1547.

Woodruff, C. C., Uncapher, M. R., & Rugg, M. D. (2006). Neural correlates of differential retrieval orientation: Sustained and item-related components. *Neuropsychologia*, *44*, 3000-3010.

Yonelinas, A., Ranganath, C., Ekstrom, A., & Wiltgen, B. (2019). A contextual binding theory of episodic memory: Systems consolidation reconsidered. *Nature Reviews Neuroscience*, *20*(6), 364-375.

Yonelinas, A. P. (2001a). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of*

London. Series B, Biological Sciences, *356*(1413), 1363-1374.

Yonelinas, A. P. (2001b). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, *130*(3), 361-379.

Zaghloul, K. A., Blanco, J. A., Weidemann, C. T., McGill, K., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2009). Human substantia nigra neurons encode unexpected financial rewards. *Science*, *323*(5920), 1496-1499.

Zeithamova, D., & Preston, A. R. (2017). Temporal proximity promotes integration of overlapping events. *Journal of Cognitive Neuroscience*, *29*(8), 1311-1323.