HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Psychology

have examined a dissertation entitled

"Measuring, monitoring, and maintaining memories in a partially observable mind"

presented by Jordan William Suchow

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. George Alvarez

Signature _____

Typed name: Prof. Patrick Cavanagh

Signature _____

Typed name: Prof. Martin Nowak

Signature _____

Typed name: Prof. Daniel Schacter

Date: April 11, 2014

Measuring, monitoring, and maintaining memories in a partially observable mind

A DISSERTATION PRESENTED
BY
JORDAN WILLIAM SUCHOW
TO
THE DEPARTMENT OF PSYCHOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
PSYCHOLOGY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS

APRIL 2014

Dissertation advisor: Professor George A. Alvarez Jordan William Suchow

Measuring, monitoring, and maintaining memories in a partially observable mind

Abstract

Visual memory holds in mind details of objects, textures, faces, and scenes. After initial exposure to an image, however, visual memories rapidly degrade because they are transferred from iconic memory, a high-capacity sensory buffer, to working memory, a low-capacity maintenance system. How does visual memory maintenance work? This dissertation builds the argument that the maintenance of short-term visual memories is analogous to the act of breathing: it is a dynamic process with a default behavior that explains much of its usual workings, but which can be observed, overridden, and controlled. Chapter 1 shows how the act of trying to remember more information causes people to forget faster and to remember less ("load-dependent forgetting" and "overreaching"). It then shows how the paradigm of evolution can be applied to the problem of maintenance, with memories competing over a limited memory-supporting commodity, explaining these effects. Chapter 2 presents experiments on metamemory, the ability of people to observe and make decisions about their own memories. The experiments isolate a component of metamemory that monitors a memory's quality as it degrades over time. Chapter 3 connects memory to metamemory, drawing on work from reinforcement learning and decision theory to liken the problem of memory maintenance to that of an agent who sequentially decides what to prioritize in a partially observable mind.

Contents

In memory of Sol Graff, my grandfather.

Acknowledgments

Thank you to George Alvarez, my advisor.

Thank you to the other members of my dissertation committee: Patrick Cavanagh, Martin Nowak, and Dan Schacter. It's a pleasure to have learned from you.

Thanks also to Ken Nakayama and Yaoda Xu for contributing to the great environment in the visionlab.

Thank you to fellow members of the visionlab and psychology department, past and present:

Arash Afraz, Jorge Almeida, Sam Anthony, Eve Ayeroff, Julie Belkova, Katie Bettencourt, Tim Brady, Donal Cahill, Sasen Cain, Jon Cant, Ramakrishna Chakravarthi, Garga Chatterjee, Sarah Cohan, Michael Cohen, Joe DeGutis, Judy Fan, Daryl Fougnie, Lúcia Garrido, Laura Germine, Jon Gill, Jason Haberman, Fred Halper, Morgan Henry, Laura Herman, Su Keun Jeong, Justin Jungé, Talia Konkle, Andy Leber, Bria Long, Camille Morvan, Marnix Naber, Maryam Vaziri Pashkam, Irene Pepperberg, Julie Rhee, Adena Schachner, Anna Shafer-Skelton, Mirta Stantic, Viola Störmer, Charles Stromeyer, Roger Strong, Arin Tuerk, Ruosi Wang, Jeremy Wilmer, Daw-An Wu, Jiedong Zhang, and Xiaoyu Zhang.

Thank you to Celia Raia and other members of the psychology department staff for help in navigating the process.

Thank you to Trinidad Zuluaga and Sarah Cormiea for their assistance with data collection.

Chapter 1 is in collaboration with Ben Allen, Martin Nowak, and George Alvarez. Chapter 2 is in collaboration with Daryl Fougnie and George Alvarez. Appendix A is in collaboration with Ben Allen. Appendix B is in collaboration with Tim Brady, Daryl Fougnie, and George Alvarez.

Thank you to Ben Allen for many helpful conversations.

Thank you to Ned Block, Josh Greene, Susan Carey, Jeremy Wolfe, Guven Guzeldere, and Stefano Anzellotti for their questions and thoughts at CBB lunch. Thanks to Justin Halberda, Steve Luck, Wei Ji Ma, and Christopher Tyler for comments at VSS. Thanks to David Rand for the introductions and encouraging words.

Thank you to Douglas Hofstadter for playing along, much to my delight.

Some of the computations in this thesis were run on the Odyssey cluster, supported by the FAS *Science Division Research Computing Group*.

This work was in part supported by a National Science Foundation CAREER Grant (BCS-0953730) to George Alvarez; NIH grant 1F32EY020706 to Daryl Fougnie; and a grant to the Center of Excellence for Learning in Education, Science, and Technology (CELEST), a Science of Learning Centers program of the National Science Foundation (NSF SBE-0354378).

Thank you to Josh Tenenbaum, Noah Goodman, Alan Yuille, and the *Institute for Pure and Applied Mathematics* at UCLA for including me in the 2011 graduate summer school in probabilistic models of cognition.

Thank you to all the professors, including Larry Maloney and Paul Bloom, who, a decade ago, took the time to respond to a letter from a high school sophomore eager to dip his toes in the waters of research.

Thank you to Denis Pelli, the one whose response changed my career. Denis invited me to a lab meeting in 2003 and over the years has taught me more than anyone.

Thank you to Miriam Graff, my grandmother, for her commitment to my education.

Thank you to my parents, Eva and Steven Suchow, who from the very beginning have been supremely supportive of my research. As I've grown older, the form of the support has changed, progressing from science museum trips, Lego Mindstorms, and balsa wood airplanes propelled by rubber bands, to rides to lab meetings, patient listening to the bumblings of someone learning a new field, and careful readthroughs of draft upon draft of everything I write, including this dissertation.

And thank you to Ariella, for standing beside me.

0

Introduction

WHEN IN 2013 GEOFFREY MUTAI crossed the finish line of the New York City marathon, he was breathing hard, having run 26.2 miles in just two hours and eight minutes, winning the race (Belson, 2013). As many elite runners seem to do, he soon slowed to a walk, placed his hands on his hips, took a few deep breaths, and continued on with an almost impossible calmness that masked much of the complexity going on inside.

Strenuous exercise, such as running a marathon or biking up a hill, changes the body in many ways:

internal temperature rises, lactic acid builds up in the muscles, and metabolic production of carbon dioxide increases (Katz & Sahlin, 1988; Cheuvront & Haymes, 2001). In response, the brain makes adjustments — for example, increasing the rate of breathing to maintain steady levels of oxygen and carbon dioxide in the blood (Feldman & McCrimmon, 2008). Patterns of breathing change in response to many environmental factors, such as the presence of toxins or the unavailability of oxygen, most often with the goal of maintaining homeostasis and, ultimately, surviving (Feldman & McCrimmon, 2008).

Breathing, though usually automatic, can be brought under conscious control. For example, *Lamaze International* teaches breathing techniques to expectant mothers to distract them from the pain of childbirth (Lothian, 2011). And the *Association Internationale pour le Développement de l'Apnée* oversees the sport of deep freediving, in which competitors plunge as far as possible below sea while holding their breath. Sighing, singing, speaking and shushing all involve the deliberate sucking of air inward and out. Lastly, becoming aware of one's own breathing — and the temporary horror of wondering what would happen if the thought were never to leave your mind — is a vivid demonstration of the interplay between effortful and automatic respiration (Kahneman, 2011; Stanovich & West, 2000; Wegner, 2009).

This dissertation is not about breathing. Rather, it is about remembering and forgetting our visual experiences. But just as computing machines have provided a useful analogy for thinking about cognition, so too can breathing provide a useful analogy for thinking about the dynamics and controllability of memory. This dissertation builds the argument that the maintenance of short-term visual memories is analogous to the act of breathing: it is a dynamic process with a default behavior that explains much of its usual workings, but which can be observed, overridden, and controlled. The default behavior of maintenance and degradation (i.e., its dynamics in the absence of top-down control) is

well described by an evolutionary process operating over the units of a memory-supporting cognitive commodity, like attention. Memories can be observed through real-time metamemory, an inward-looking process that tracks the status of a memory as it degrades over time. And the default behavior can be overridden and controlled by the processes of directed forgetting and self-directed remembering, which redirect maintenance in accordance with the goals.

The plan of the dissertation is as follows:
This introduction lays the groundwork by reviewing the relevant literature on visual working memory, metamemory, and directed forgetting.

Chapter 1, a manuscript on the time course of visual memory, describes a set of high-quality forgetting functions measured through participants' performance on a visual memory task. These forgetting functions reveal the inadequacies of current time-based models. The manuscript then uses *evolutionary dynamics* — a mathematical framework for describing how information is reproduced in an environment that is subject to mutation, selection, and random drift — to construct a new model of the time course of maintenance and degradation in visual working memory.

Chapter 2, a manuscript on metamemory, provides an experimental existence proof of a certain kind of real-time metamemory, whereby, in real time, the memorizer monitors the current quality of visual memories. This process is useful both because it provides a mechanism for judging confidence in our memories, and because it can provide the foundation of an adaptive maintenance strategy that selectively targets memories depending on their strength.

Chapter 3, a manuscript on the control of working memory, begins with an experiment on directed forgetting, in which a memorizer controls the contents of memory according to the demands of the task. It continues with an experiment on self-directed remembering, in which people adaptively adjust the target of maintenance according to metamemory beliefs (e.g., by preferentially maintaining

the best-remembered item when asked to do so). The chapter goes on to present a formal framework for thinking about memory maintenance, drawing on work from the fields of reinforcement learning and decision theory to cast the problem of memory maintenance as a partially-observable Markov decision process. In doing so, it extends the model proposed in Chapter 1 to include self-directed remembering as well as the real-time metamemory revealed in Chapter 2.

The final chapter reviews the contributions of the dissertation.

Working memory is a storage system that actively holds information in mind and allows for its manipulation, providing a workspace for thought (Baddeley, 1992; Cowan, 2005). Its capacity is strikingly limited, perhaps to only a few sights or sounds (Miller, 1956). Using working memory is effortful: pupils dilate, skin conductance rises, and secondary tasks become impossible to perform well (Kahneman, 1973). Much of the research on working memory has focused on characterizing its limits and determining what gives rise to them. For example, working memory capacity is known to be lower in young children and the elderly (Dobbs & Rule, 1989; Salthouse & Babcock, 1991; Zacks, 1989; Gathercole et al., 2004), correlates strongly with a person's fluid intelligence (Conway et al., 2003; Cowan, 2005), is affected by sleep schedule (Chee & Choo, 2004; Steenari et al., 2003), and can be impaired in people with mental disorders such as schizophrenia (Goldman-Rakic, 1994; Gold et al., 1997; Manoach, 2003; Joormann & Gotlib, 2008; Martinussen et al., 2005). From this work, we have learned a considerable amount about how much can be remembered and who is best at remembering it.

Over the past 50 years, there have been two dominant approaches to studying the processes that underlie working memory: the first is cognitive psychology, and the second is cognitive neuroscience. Cognitive psychology is the study of how mental processes affect behavior and thought (Neisser,

1967). Cognitive neuroscience is the study of the neural substrates of cognition — e.g., the brain regions and networks that are active when a person thinks or observes — and how those substrates together produce experience and behavior (Gazzaniga, 2004).

The cognitive psychology approach to studying memory begins with designing a behavioral task and then proceeds by adjusting the task demands, altering the information that is presented (e.g., the number of stimuli, the format of presentation, or the category of object or sound) or the duration of the *retention interval*, which is the time between when the information is initially presented and when it is later accessed to perform a task. This approach has a long tradition in experimental psychology that goes back to Ebbinghaus' 1885 book *Über das Gedächtnis*, in which he measured forgetting curves that track the amount remembered as it falls over time.

The cognitive neuroscience approach, in contrast, catalogs the neural substrates of encoding and maintaining visual information in working memory. Working memory is supported by a control network that includes the prefrontal cortex, basal ganglia, and parietal cortex (Baddeley, 1992; Voytek & Knight, 2010; Xu & Chun, 2005; Todd & Marois, 2004), acting in conjuction with posterior regions that support storage (Postle et al., 1999; D'Esposito & Postle, 1999; Harrison & Tong, 2009; Emrich et al., 2013). One of the most relevant findings is the existence of *contralateral delay activity* (CDA), which is event-related potential (ERP) activity sustained during the maintenance interval of a memory task. The amplitude of CDA tracks how much is remembered, and individual differences in CDA amplitude are correlated with individual differences in memory capacity (Vogel & Machizawa, 2004).

Cognitive psychology and cognitive neuroscience are not the only approaches to understanding working memory. A third approach, the computational approach, delineates three levels of analysis that are needed to explain any cognitive phenomenon: the computational, the algorithmic, and the physical (Marr, 1982; Chater & Oaksford, 1998). The computational level considers the structure of

the problem that the person (or system) faces, the information that is available, and the logic of the possible solutions to that problem. The algorithmic level considers the rules by which a solution is carried out. The physical level considers how those rules translate to a working biological system.

Though this flavor of thinking is largely absent from theories of visual working memory, there is a strong precedent for it in the context of long term memory. John Anderson, a cognitive scientist at Carnegie Mellon, provided (alongside colleagues) the first rational analysis of human memory (Anderson, 1989; Anderson & Milson, 1989), introducing the approach to the study of higher-level cognition (Chater & Oaksford, 1998). Anderson proposed that the problem of long term memory is that of retrieving relevant information in a setting where retrieval is costly and relevance is uncertain (Anderson & Milson, 1989). In later work with Lael Schooler, the two found that the availability of a memory reflects the probability that the information it contains will be needed again in the future, as determined by the statistics of reoccurrence as experienced in the environment (Anderson & Schooler, 1991). The classic power-law forgetting functions of long-term memory (Ebbinghaus, 1913; Wixted & Carpenter, 2007; Wickelgren, 1974) were shown to naturally reflect the power-law reoccurence of experiences in the environment — e.g., the names found in headlines of *The New York Times* or words encountered in the verbal interactions of children (Anderson & Schooler, 1991).

One example of a rational analysis of visual working memory is from Sims et al. (2012), which proposes that the problem of visual working memory is that of transmitting visual information from one point in time to another. In this formulation, solutions to the problem of visual memory take the form of encoders and decoders that maximize performance on a task. Sims et al. (2012) found that formulating the problem of working memory in this way does an excellent job of predicting behavioral performance and its dependence on the statistics of the visual input.

This dissertation poses the problem of visual working memory differently because it takes seri-

ously the idea that maintenance is an active and controllable process (Jonides et al., 2008). In this light, the problem of visual working memory is seen as that of encoding visual information, maintaining it over a short duration, and then later accessing it to perform a task well. This dissertation combines the cognitive psychology and computational approaches to explore the rules by which memory maintenance operates. This entails considering the space of possible maintenance strategies and how that space is constrained by experiments on the default behavior, structure, and observability of memory. One such constraint is metamemory.

In 2000 and 2004, ken jennings and brad rutter were contestants on the American game show *Jeopardy!*, in which players display their knowledge of trivia by quickly responding with a question whose answer matches the provided clue. Jennings and Rutter were two of the most successful contestants in the show's history, winning just shy of 6 million USD between them (Jeopardy Productions, 2014). Strong players often know the answer immediately and have it ready before buzzing in, but metamemory provides a useful tool when they don't. Metamemory is an awareness of one's memories and the systems that store them. It is particularly useful in *Jeopardy!* because of two properties of the game's design: the first player to "buzz in" is the one given the chance to respond, and incorrect responses are penalized. Metamemory can thus help in two distinct ways. First, a player can press the buzzer even before having recalled the correct response, using the following second or two to retrieve the relevant information. Second, a player can avoid buzzing in when they are uncertain or unlikely to respond correctly.

Metamemory is useful beyond *Jeopardy!*, extending to our everyday lives. It helps us to determine that we are uncertain, to know when to ask for a reminder and who to ask for it, and to form beliefs about our ability to remember certain kinds of experiences and events. Metamemory is often studied

in the context of long term memory, where it is invoked to explain phenomena such as tip-of-the-tongue states and the feeling of knowing (Flavell & Wellman, 1977; Wellman, 1977; Brown, 1991). Healthy individuals have a rich set of metamemory skills that guide learning, decision making, and action (Metcalfe & Shimamura, 1994). Neurological diseases, such as Alzheimer's and Korsakoff's syndrome, adversely affect metamemory judgments, causing a mismatch between what is remembered and what is believed to be remembered (Pannu & Kaszniak, 2005).

Memories can be forgotten intentionally. Experiments on this process of "directed forgetting" often ask participants to study some information and then later direct them to remember or forget specific elements of what was studied (Muther, 1965; Bjork et al., 1968). Memory tends to be better for the to-be-remembered information than for the to-be-forgotten information. For example, in Woodward & Bjork (1971), participants studied a list of words and later were asked to recall as many of them as possible. This is the popular *free recall* paradigm used extensively in studies of long-term memory. Following each word's presentation, a cue appeared instructing the participant either to remember or to forget the word. Later, participants were asked to recall all the words from the studied list, regardless of how those words initially had been marked. The recall task was challenging. Critically, its difficulty depended on how the word had been marked: those marked as to-be-remembered were recalled 23.3% of the time, whereas those marked as to-be-forgotten were recalled only 4.7% of the time. This is the hallmark of directed forgetting, which has been demonstrated in both long- and short-term memory and was a popular topic of memory research in the 1970's (Woodward & Bjork, 1971; Muther, 1965; Bjork et al., 1968; Block, 1971; Bjork, 1972; Epstein, 1972; Burwitz, 1974; MacLeod, 1975).

Directed forgetting is intimately related to cognitive control and to the processes that determine

our conscious thoughts from moment to moment (Macrae et al., 1997). For example, experimentally adding cognitive load decreases people's ability to suppress unwanted thoughts (Wegner & Erber, 1992), and young children and the elderly have deficits in attentional processing, which makes it more difficult for them to abandon memories and thoughts that are no longer relevant (Harnishfeger & Bjorklund, 1993; Bjorklund & Harnishfeger, 1990; Passolunghi & Siegel, 2001; Hartman & Hasher, 1991).

Memory, metamemory, and directed forgetting are the building blocks of this dissertation, which begins by describing the default behavior of memory maintenance, goes on to show that the memory state produced by that undirected behavior can be observed through metamemory, elaborates and constrains the ways that maintenance can be controlled, and then proposes a formal framework for the problem of memory maintenance in a partially observable mind.

We begin with the default behavior.

1

Evolutionary dynamics of visual memory

1.0 Abstract

Visual memory enables a viewer to hold in mind details of objects, textures, faces, and scenes. After initial exposure to an image, however, visual memories rapidly degrade because they are transferred from iconic memory, a high-capacity sensory buffer, to working memory, a low-capacity maintenance system. Here, we extend the classic depiction of the dynamics of visual memory maintenance to include competitive interactions between memories, fluid reallocation of a memory-supporting

commodity, and a stability threshold that determines the weakest memory that can still be maintained. The proposed model, based on these principles, can be understood as an evolutionary process with memories competing over a limited memory-supporting commodity. The model reproduces the time course of visual working memory observed in a series of experiments. Notable features of this time course include load-dependent stability and overreaching, in which the act of trying to remember more information causes people to forget faster, and to remember less, respectively. Our results demonstrate that evolutionary models provide quantitative insights into the mechanisms of memory maintenance.

1.1 Introduction

Memories typically degrade until, at last, they are forgotten. From its inception, research on memory degradation has characterized *forgetting functions* that track the downfall of how much is remembered over time (Wixted & Ebbesen, 1991; Ebbinghaus, 1913). A forgetting function is shaped by the processes that degrade and maintain memories, and its functional form often provides a signature of the underlying mechanisms (Wixted, 1990). Examining forgetting functions can thus reveal important insights, having previously provided some of the primary evidence for iconic memory (Sperling, 1960) and for rational theories of adaptive forgetting (Anderson, 1989; Anderson & Schooler, 1991).

In the case of visual memory, which holds in mind the details of what was seen, at least three subsystems contribute to storage and maintenance. Each subsystem has a characteristic timescale, format, and neural substrate. Iconic memory, a high-capacity sensory buffer, operates over short time scales (0.05–1 s) and is thought to result from persistence of activation in mid- to high-level visual areas such as the lateral occipital complex and temporal cortex (Sperling, 1960; Keysers et al., 2005; Ferber et al., 2005). Visual working memory, an active maintenance system, operates over moderate

time scales (0.5–20 s) and is supported by a network that includes prefrontal cortex, basal ganglia, and parietal cortex (Baddeley, 1992; Voytek & Knight, 2010; Xu & Chun, 2005; Todd & Marois, 2004). Visual long-term memory, a high-capacity passive store, operates over lengthy time scales (minutes to decades) and recruits much of the same machinery that supports more general forms of long-term memory, such as the hippocampus (Brady et al., 2011b). Other subsystems have been proposed, each with its own particular properties and substrates (Sligte et al., 2008; Magnussen, 2000; Wood, 2009). Together, these systems maintain visual memories, allowing us to remember what we see.

The classic forgetting function of visual memory, which is applicable to short and moderate time scales, has a brief period of rapid decline followed by a long plateau, a form that is attributed to the quick fading of iconic memory and the stability of working memory (Sperling, 1960). This model has survived for over 50 years with only slight modification (Box 1) (Zhang & Luck, 2009). Here, we extend the classic model to account for new data, leveraging the tools of evolutionary biology to model memories as entities that compete for a limited mental commodity that is shared among them. In a series of experiments, we asked participants to remember a set of objects, and then after a short delay, to report the color of a randomly selected object using a graded continuous-report procedure (Fig. 1 and Methods). We tested a five-hundredfold range of durations (0.03–16 s) and a twelve-fold range of loads (1–12 objects), randomly interleaving them all. We used the resulting data to derive a set of high-quality forgetting curves (Methods). These revealed effects unnoticed by previous studies, which have typically kept load or duration fixed or varied each in a blocked design, rarely manipulating both. Complete coverage of duration and load enables us to compare changes over time with minimal contamination by differences in initial encoding; and by using graded rather than discrete responses, we gain higher efficiency in detecting the presence of weak memories and the ability to

distinguish weak memories from those that are completely forgotten.

Box 1

The classic model of the time course of visual memory

Experiments in the 1960s revealed the existence of iconic memory, a high-capacity sensory buffer whose contents are short-lived, fading within a second (Sperling, 1960). At the time, working memory was assumed to be stable, with stored objects remaining there indefinitely. The forgetting function of the classic model, defined by the number of objects remembered at time t, is given by

$$y = \beta + \exp(-t/\tau),$$

where τ is the mean lifetime of an iconic memory and β is the capacity of working memory (Appendix A).

The pure death model

The pure death model extends the classic model to accommodate degradation in working memory. We formalize this degradation as a concurrent process of exponential decay that happens more slowly than iconic decay (Appendix A).

The sudden death model

The sudden death model extends the pure death model by proposing a 4-second window of time in which working memory is immune to degradation (Zhang & Luck, 2009) (Appendix A).

Figure 1: The memory task. Participants stare at a small cross at the center of the screen. A set of colorful dots briefly appears. After a delay, the location of one of the dots (selected at random) is marked with a cue. The participant is asked to report the color of the dot that appeared at the marked location. Performance is assessed by comparing the reported color to the true color. Here, the participant makes a large error, reporting the green object as orange.

1.2 Results

1.2.0 Memory stability depends on load

Participants' errors on the memory task were used to derive forgetting curves that track the number of remembered objects as it falls over time (Fig. 2a, Methods). Fitting an exponential function separately to the data from each memory load, we found that the rate of forgetting depends on the total amount of information held in mind, with lone memories lasting longest (estimated mean lifetime of 157 s) and higher loads leading to progressively shorter lifetimes (Fig. 2b–d, Methods). The relationship between memory load and mean lifetime is well described by a power law with exponent -1.7 ($r^2 = 0.98$, $p = 6.5 \times 10^{-8}$), such that halving the load leads to roughly a tripling in mean lifetime (Fig. 2b). This relationship was also found when limiting analysis to durations greater than 1 s, where iconic memory plays no role (Sperling, 1960; Yang, 1999) (Fig. 2c). In the initial analysis, we assumed that the forgetting function is exponential-like (Box 1, Eq. 1). To test whether load-dependent stability is robust to this assumption, we also considered another functional form—a power law. Power law forgetting has been observed in long-term memory (Wixted & Ebbesen, 1991) and is common because it can arise both normatively (i.e., as the optimal solution to a task) (Anderson, 1989) and as an artifact of averaging exponential-like forgetting functions that differ in timescale (Anderson & Tweney, 1997; Anderson, 2001). We found a comparable effect of load-dependent stability under power law forgetting (Fig. 2d).

1.2.1 Crossovers and overreaching

The lines of the forgetting functions for each load cross (Fig 3a). At short durations, presenting a greater number of objects causes more to be remembered. At long durations, however, the oppo-

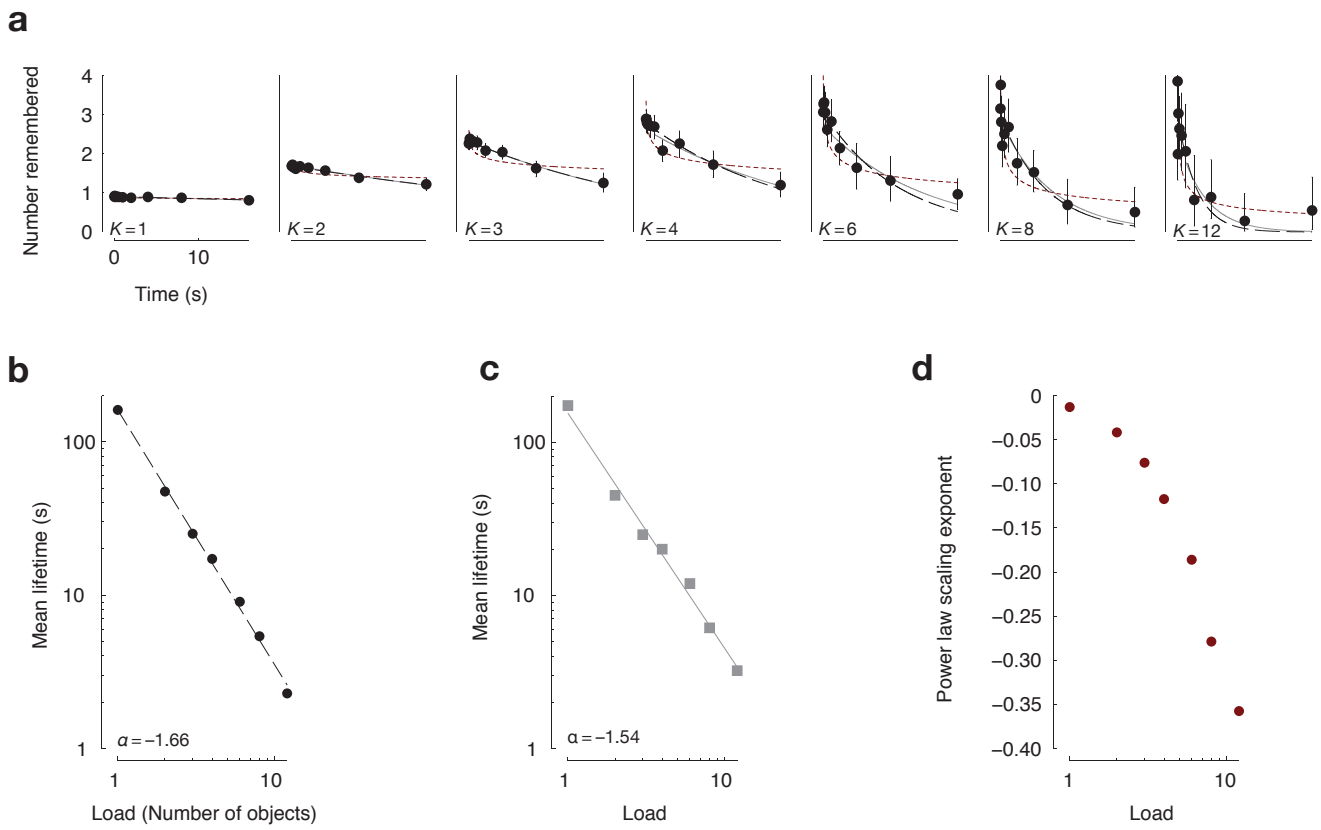


Figure 2: Load-dependent forgetting: the more you try to remember, the faster you forget. (a) Subplots show the empirical forgetting function for each load (K = 1, 2, 3, 4, 6, 8, or 12 objects), tracking the number of remembered objects as it falls over time. The dashed black and solid grey curves assume an exponential form to the forgetting function; the former is fit to all the data and the latter considers only durations of at least 1 s, where iconic memory plays no role. The dotted red curve assumes that forgetting follows a power law. Curves were fit to the data from each load separately (Methods). Error bars here and in other figures are 95% credible intervals. (b) Comparing lifetimes across loads, the relationship is well described by a power law with exponent -1.66. (c) This relationship also holds for the estimates derived from durations of at least 1 s, where iconic memory plays no role. (d) A similar relationship is found for degradation under power law forgetting, quantified by the scaling exponent, which falls precipitously with load.

site is often true: presenting a greater number of objects causes fewer to be remembered (Fig 3b). Crossovers in the forgetting function imply that the relationship between the number of objects presented and the number remembered changes with time. (Fig 3c).

The presence of crossovers suggests a flawed strategy of the participants, who presumably control how many objects they encode and maintain. Like a bodybuilder who herniates a disk by straining to lift too heavy a weight, our participants performed worse because they tried to encode and maintain more than they could handle—they overreached. A comparable effect has been reported for tracking many moving objects at once, which is a task that is demanding of attention (Holcombe & Chen, 2013). Alternatively, it is possible that participants chose appropriately when deciding how many objects to encode or maintain, but that the presence of distracting objects led to flawed execution of the chosen strategy (Vogel et al., 2005). Crossovers are inconsistent with the classic model and its variants, whose lines occasionally meet, but never cross (Box 1; for details, see Appendix A).

1.2.2 Evolutionary model

To explain these results, we propose an account of visual memory rooted in evolutionary dynamics, a mathematical framework for describing how information is reproduced in a setting that is subject to mutation, selection, and random drift (Nowak, 2006). Specifically, we describe an evolutionary process operating over a commodity that supports memory. This commodity may take any one of a number of forms, including (for example) cycles of a time-based refreshing process (Vergauwe et al., 2009) or populations of neurons in prefrontal cortex representing "token" encodings of visual events (Bowman & Wyble, 2007). No matter its particular form, what defines a commodity is being a limited asset, at least partially shared across memories, whose availability affects performance. A shared commodity stands in contrast to a purely local substrate that represents specific stimulus attributes in

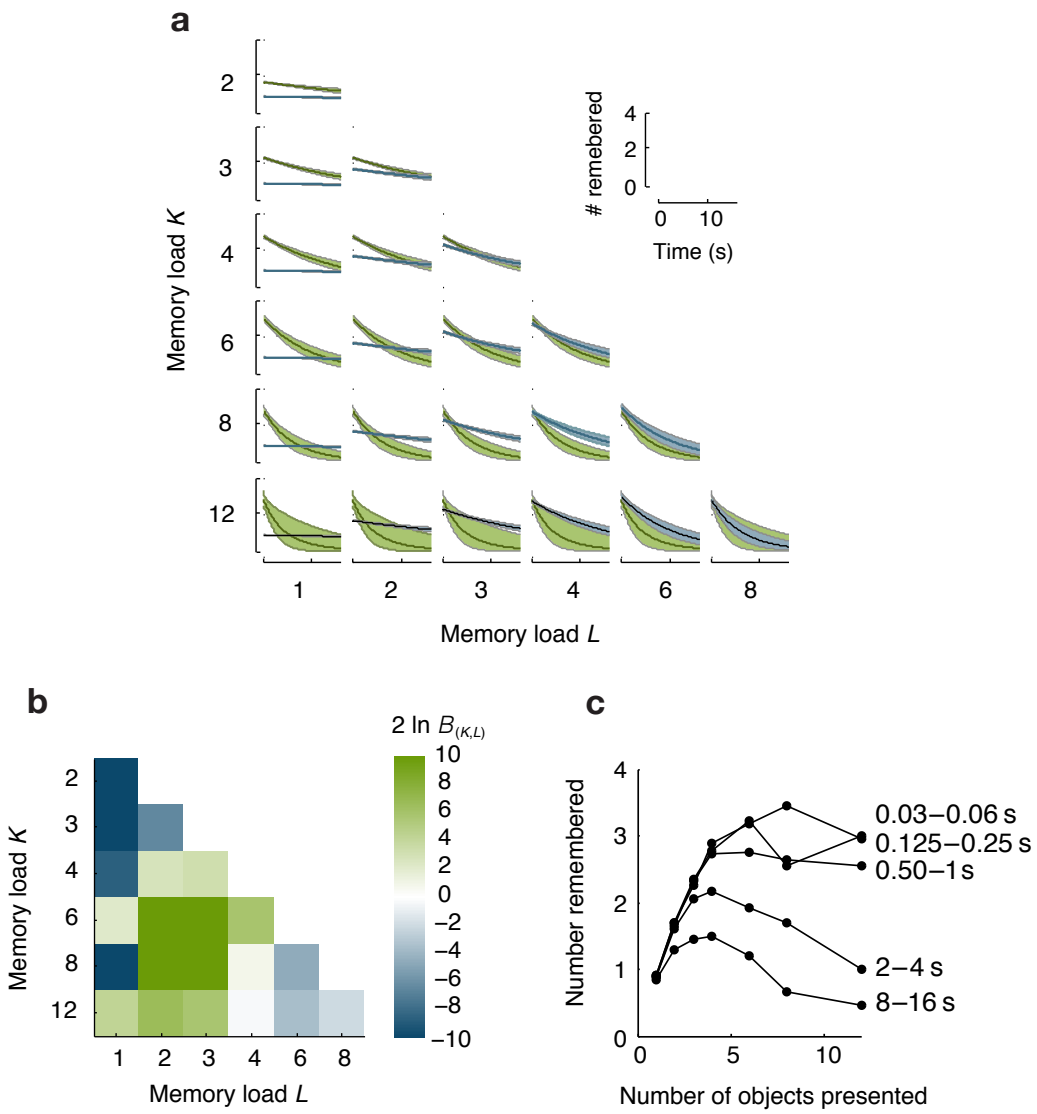


Figure 3: Crossovers in the forgetting function and overreaching. (a) Each subplot is a pairwise comparison of the forgetting functions for two memory loads, with the greater load K plotted in green and the lesser load L in blue. Shaded error bars show 95% credible intervals at each time point. (b) The strength of evidence for a crossover between the forgetting functions for a pair of loads K and L is expressed as twice the natural logarithm of the Bayes factor $B(K,L)$ in favor of a model with crossover over one without it (Kass & Raftery, 1995) (Methods). In the heat map, $B(K,L)$ is coded as positive (green) when the evidence favors the crossover model and as negative (blue) when it favors a model without crossover. (c) The crossover effect implies that the relationship between the number of objects presented and the number remembered will change over time. The plateau at ≈ 3 objects for short durations is considered to be the signature of visual working memory's meager capacity. However, the non-monotonic curves seen for durations greater than 1 s are new and suggest a failure on the part of participants, who would have performed better by trying to encode less of the display.

particular locations of the visual field (Franconeri et al., 2013). Though such location- and content-based substrates are essential for encoding information into working memory, they are perhaps less relevant to memory maintenance, which may operate over a more pluripotent medium (Bowman & Wyble, 2007).

Recent work has sought to determine both the quantization (Zhang & Luck, 2008; Bays & Husain, 2008) of the commodity and the structure of the memories that it forms [e.g., whether they form bound objects (Luck & Vogel, 1997), bags of unbound features (Fougnie & Alvarez, 2011), or hierarchical bundles of features (Brady et al., 2011b)]. In the general case, the commodity is divided into N quanta, each of which is dedicated to some information about a memory structure. Discrete "slot"-based models set $N \approx 4$, whereas "continuous resource" models consider the limit as N tends to infinity (Zhang & Luck, 2008; Bays & Husain, 2008; Cowan, 2001; Wilken & Ma, 2004).

We model the evolution of the quantal population using a generalization of the Moran process. The Moran process is a model of evolution in finite populations that was originally used to describe the dynamics of allele frequencies (Moran, 1958), and which has recently been leveraged to describe evolutionary processes in diverse settings, including frequency-dependent selection, emergence of cooperative behavior, and cultural evolution of language (Nowak et al., 2004; Fudenberg et al., 2006; Komarova & Nowak, 2003). The Moran process begins with a population of quanta (the units of the commodity) that have been assigned to structures (which may be objects, features, bundles, etc.). At each time step, a quantum becomes degraded, losing the information that it stores. In the same step, the lost information is replaced by the contents of another quantum, randomly selected from them all (Fig. 4). Our generalization further introduces a stability threshold: if at any point a structure has fewer than s quanta assigned to it, it becomes inaccessible to the maintenance process and the associated quanta lose their assignment, floating freely until they are reassigned (Fig. 4, grey dots). This

threshold is comparable to a recently-proposed lower bound on the fidelity of an accessible memory (Brady et al., 2013) and has the effect of limiting the number of structures that can be stored to approximately N/s. When the stability threshold is a single quantum, we can derive the forgetting function analytically (Appendix A); for greater values of the stability threshold, the forgetting function is obtained numerically. Over time, the number of quanta assigned to a structure drifts. Eventually, either a single structure reaches fixation, with all the quanta assigned to it, or corruption prevails and all the quanta are left free-floating and unassigned.

Various cognitive processes could give rise to these dynamics. First consider a process of active maintenance that recycles the memory commodity, repurposing quanta dedicated to lost memories in order to provide redundancy to those that remain. Alternatively, consider a process of interference where at each time step a quantum becomes corrupted, taking on the value and assignment of an intruding quantum. In these ways, the evolutionary process can be seen as a formal model of memory maintenance in the face of degradation due to interference or decay.

Each component of the evolutionary model — the commodity, the degradation process, and the stability threshold — contributes to the resulting dynamics. When a memory structure loses a quantum and hits the stability threshold, that structure is lost. This happens quickly at first, but more slowly over time, because the loss of one memory lends stability to those that remain. When there are many objects to remember, the memory commodity is spread thinly, with fewer quanta per memory structure, and so each one stands closer to the stability threshold. In contrast, when there are fewer objects to remember, the representation of each one is more stable. This discrepancy accounts for the relationship between lifetime and load and may also explain the remarkable stability of lone memories, which need not compete at all for the memory commodity.

The proposed evolutionary process reproduces the observed forgetting functions of visual mem-

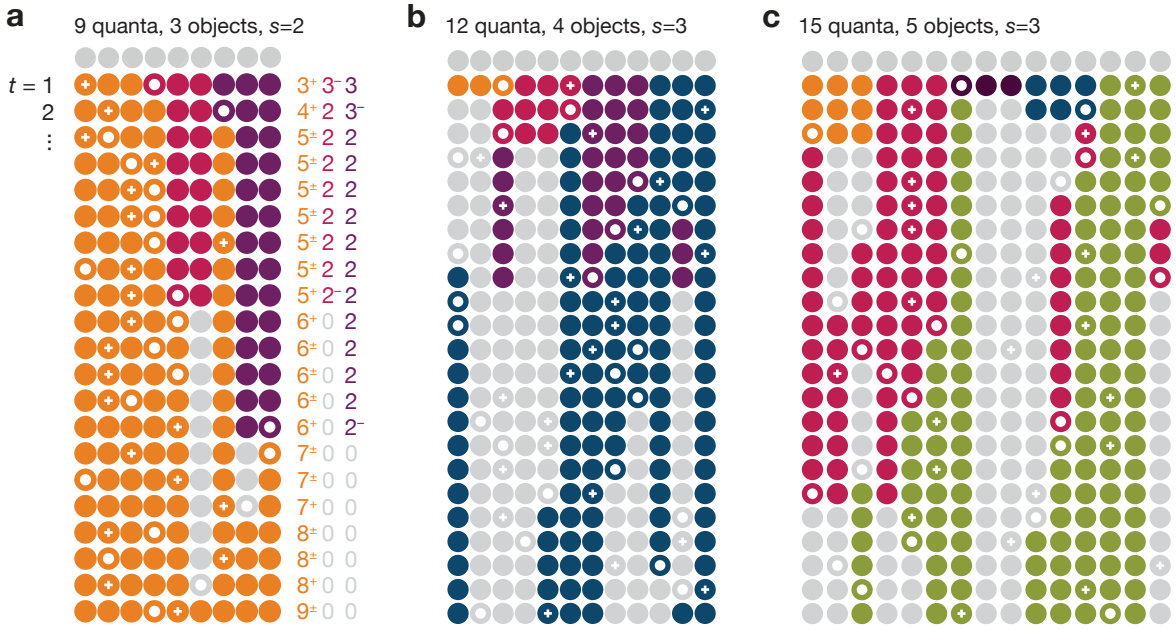


Figure 4: Modeling the evolution of a memory-supporting commodity. (a) In the top row, a pool of 9 unallocated quanta (grey dots) that wait to be assigned. In the second row, each quantum is assigned to one of three structures, labeled in orange, red, and purple. Subsequent rows show the process as it plays out over time, one time step per row. An empty circle denotes the quantum that died and a circle with a plus mark denotes the quantum that was selected to replace it. When the number of quanta assigned to a structure drops below the threshold (s = 2 in panel **a** and s = 3 in panels **b** and **c**), the remaining quanta become inaccessible to the maintenance process and lose their assignment (greyed-out dots). Numbers to the right of the panel count the number of quanta dedicated to each structure at that time step, with superscripts showing which structure gains, +, loses, –, or both, $\pm$. The number of stored structures corresponds to the number of unique colors in a row. In this run, the orange structure reaches fixation. (b) A second iteration of the process, with 12 quanta and 4 objects. At the last time step that is displayed, the blue structure is the only one left, but it has not yet reached fixation, with much of the commodity left unassigned. (c) A third iteration, with 15 quanta and 5 objects. Red takes an early lead, but is eventually overcome by green.

ory, showing effects of load-dependent forgetting and overreaching, effects that are inconsistent with the classic, pure death, and sudden death accounts, which show neither effect (Fig. 5, Methods, and Appendix A). In the classic account (Fig. 5, grey dashed lines), only iconic memory degrades; the stability of working memory produces flat forgetting functions with no slope and which do not cross. In the pure death account (Fig. 5, blue dashed lines), working memory decays at a fixed rate that is independent of load; this produces sloped lines that share a common decay rate (mean lifetime) and never cross. The same is also true of the sudden death account (Fig. 5, yellow dashed lines), which extends the pure death account by proposing a 4-second window of time in which working memory is immune to degradation (Zhang & Luck, 2009). Only the proposed evolutionary process produces both effects (Fig. 5, green solid lines).

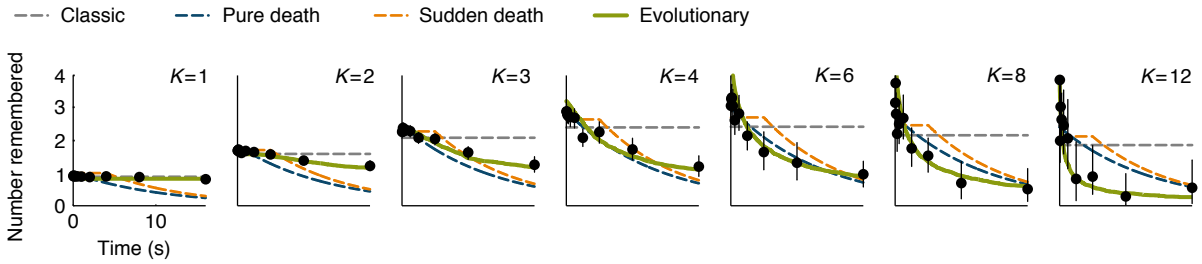


Figure 5: Comparing the forgetting functions of the classic, pure death, sudden death, and evolutionary models. Subplots show the forgetting function for a particular load (1, 2, 3, 4, 6, 8, or 12 objects). Competing models fail to capture important aspects of the data. The classic model (grey dashed line) does not change over time. The pure death model (blue dashed line) has a fixed rate of forgetting, one that is too quick for low loads (K = 1, 2, and 3) and too slow for high loads (K = 8 and 12). The sudden death model behaves similarly to the pure death model. The evolutionary model succeeds, with slow forgetting at low loads and quick forgetting at high loads (best-fit parameters N = 58, p = 0.82, t_{step} = 0.01, and s = 7). The form of each forgetting function is derived in Appendix A. We also considered the effect of individual differences on the predictions of each model (see Figs. 7–12).

1.3 Discussion

We modeled memory degradation as an evolutionary process operating over the quantal units of a memory-supporting commodity. The proposed model, a generalization of the Moran process with a stability threshold, captures important features of the forgetting functions of visual working memory. First, the stability of a memory depends on the load, with lesser loads leading to progressively longer lifetimes. Lone memories are remarkably stable, with lifetimes on the order of minutes. In contrast, in the context of high loads, memories degrade quickly, in just a handful of seconds. The relationship between load and mean lifetime—load-dependent forgetting—is well characterized by a power law. Second, the lines of the forgetting functions at different loads cross, implying a failure of the participants to encode or maintain an amount of information that was commensurate with their abilities to maintain it. This is overreaching. These features follow naturally from the proposed evolutionary process, but are absent from previous models.

In the context of visual working memory, encoding and maintenance are often viewed as a process in which a limited store fills up during encoding and then remains mostly stable, perhaps with whole object representations being lost one by one over time. Importantly, in this view, encoding and maintenance happen independently over stored objects, resulting in exponential decay functions with a rate shared across different loads. Load-dependent forgetting suggests an alternate view: visual memory representations compete for a commodity that is at least partially shared among them, such that the success of maintenance for one structure is affected by that for the others, thereby introducing a dependency of forgetting on load. Our proposed evolutionary model is the simplest instantiation of this principle, with a mental commodity fully shared across representations.

The capacity of working memory is often described as a small and stable "magical" number, one

that began life as 7±2 chunks, later slipped to 4, and now hovers restlessly at 3, 2, or even just one (Cowan, 2001; Miller, 1956; Rensink, 1999; McElree, 2001). The proposed evolutionary model and its load-dependent stability may help to explain why these magical numbers are so pervasive. In our model, when content is lost from memory, due to either interference or decay, the commodity that had previously been assigned to that content is co-opted for use elsewhere. This interference (or re-allocation) brings stability to the structures that remain, such that it becomes increasingly difficult to lose the next one. No matter how many structures the quantal resource is initially spread among, within a short time most of the structures are lost. The forgetting functions therefore spend the bulk of their time hovering at only a small handful of objects.

It is conceivable that the proposed process could be used to describe both iconic and working memory, together, as a single process. Iconic memory was initially considered to be a unitary system, but was later fractionated into two distinct subcomponents, one providing visible persistence (i.e., the experience of seeing a stimulus after its removal), the other providing informational persistence (i.e., remembering something about a stimulus after its removal) (Coltheart, 1980). Visible persistence is distinct in its phenomenology from working memory, as memories are rarely experienced as being seen, but informational persistence and working memory have long been conflated. For example, studies of visual working memory often test at durations of 500–1000 ms, a point in time at which there is a non-negligible contribution of iconic memory to task performance (Yang, 1999). We find that the evolutionary model provides excellent fits to the forgetting functions of iconic memory that have been measured in previous experiments (Fig. 6). Experimental evidence of a distinct iconic storage system underlying informational persistence comes from a variety of experiments, not all of which rely on its timing (e.g., see Becker et al., 2000). However, the closeness of fit between model and data suggests that informational persistence in iconic memory may be the initial moments of

maintenance in a lengthier short-term storage system.

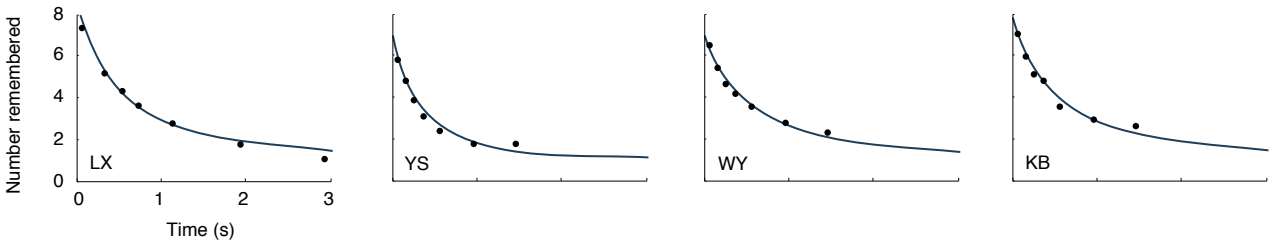


Figure 6: The evolutionary model can be used fit the full time course of visual memory as a single process. Data are replotted from Yang (1999). Each subplot is data from the participant whose initials appear in the bottom left corner. The stability threshold was fixed at $s = 1$ (see Methods).

Evolutionary dynamics provides a rich framework in which to extend our account of visual memory. For example, it is likely that the neural substrate over which visual memory maintenance operates is in some way structured—perhaps as a gridded visuotopic maps like those found in visual areas in the brain, or as a scale-free network, like so many other biological systems (Franconeri et al., 2013; Gardner et al., 2008; Schira et al., 2010). Evolutionary graph theory, which extends evolutionary dynamics to structured populations, is a natural tool for specifying the interaction network of the memory commodity and exploring how such structure impacts the stability of memories (Lieberman et al., 2005). Similarly, frequency-dependent fitness, where the success of an individual depends on the abundance of that individual's type, is analogous to a memory maintenance policy that selectively maintains memories according to their stability (e.g., by purifying, selectively maintaining the strongest memories, or balancing, selectively maintaining those memories on the brink) (Nowak et al., 2004).

By constructing an evolutionary model of memory degradation that operates over the natural units of visual memory allocation and maintenance—those of a memory-supporting commodity, rather than whole objects—we are able to build better process models of memory maintenance and its dy-

namics. Here, we focused on short-term visual memories. But just as the framework of evolutionary theory has been applied across many domains and scales, from alleles to words and from cells to societies, so too might our approach, when appropriately extended, be applied to memory maintenance in more complex systems, such as the transactional and collective memories of groups.

1.4 Methods

Participants

We recruited 1000 participants using Amazon Mechanical Turk, an online labor market where people perform short computer-based tasks for pay (Berinsky et al., 2012; Buhrmester et al., 2011; Ipeirotis, 2010; Mason & Suri, 2012). Each participant was paid $0.50 for a few minutes of work. Recruitment and testing was executed in accordance with Harvard University regulations and approved by the Committee on the Use of Human Subjects in Research under the Institutional Review Board for the Faculty of Arts and Sciences. Recruitment was limited to participants based in the US with approval rates of at least 90%.

Stimuli

The stimulus consisted of a set of 1–12 colorful dots. The dots were arranged in a ring around a small central fixation point. Each dot appeared in one of twelve locations spaced equally around the ring, with the constraint that each dot had its own location. Dots were randomly assigned one of 180 equally spaced colors drawn from a circle (radius 59°, center L=54, a=18 and b=−8) cut out from the CIE $L^*a^*b^*$ color space. Stimuli were rendered in a browser. The viewing distance was approximately 50 cm.

Procedure

The participant pressed the space bar to begin the trial. The stimulus immediately appeared for 250 ms and then disappeared. Participants were asked to remember the colors of the presented dots. The screen remained blank for the retention interval. Once the waiting period was over, a small cue appeared in the location of one of the dots, selected at random. The participant used the mouse to select the remembered color of the cued dot. Colors were selected by moving the mouse in a circle around the center of the display. A dot appeared at the center of the display and was continuously updated with the currently selected color. Participants registered their selection by clicking. No feedback was provided. There were ten possible retention intervals (1/32, 1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, or 16 s) and seven possible memory loads (1, 2, 3, 4, 6, 8, and 12 objects), for a total of $7 \times 10 = 70$ test trials. The order of the trials was randomized so that the participant would not know at the time of encoding how long they would need to remember the objects. There were 6 practice trials, 1–6 objects in ascending order, all with a retention interval of 1 s. There were negligible practice effects during the test trials, suggesting that our training procedure was sufficient for participants to perform the task well (Fig. 13).

Extracting the empirical forgetting functions

First we excluded participants who showed weak evidence of having faithfully completed the task. To do this, for each participant, we compared two models of their performance using the Akaike information criterion. The first model was a two-parameter model (Zhang & Luck, 2008) where with probability $1 - g$ the participant remembers the stimulus with fixed fidelity σ, the dispersion parameter of a von Mises distribution (a circular analogue to the normal distribution), and with probability g guesses blindly. The other model was a zero-parameter model where the participant always

guessed blindly. Since our null model — complete guessing for all 76 trials — is so weak, our criterion for inclusion was strict, $\text{AIC}_C \geq 10$, which constitutes strong evidence of the presence of memory (Akaike, 1974). This strict criterion may inadvertently exclude participants with poor working memory, though the results we find are comparable when relaxing the inclusion criterion to $\text{AIC}_C \geq 3$, which constitutes moderate evidence of memory.

Next, we combined participants' data into a super-subject. The main manipulations of time and load were performed within each subject — one trial per condition per participant — but the analysis combined the data together. Though this is necessary to achieve sufficiently precise measurements, it leaves open the possibility that variability among people in the form of individual differences will affect the shape of the measured curves (see later sections). We fit a four-parameter variable-precision model (Fougnie et al., 2012; van den Berg et al., 2012) to arrive at an estimate of the guess rate g separately for each duration and load K. The product $(1 - g)K$, the average number of remembered objects, is plotted in Figures 2, 3, and 5. Analysis was performed using MemToolbox 1.0.0. (Suchow et al., 2013).

Estimating mean lifetimes

Mean lifetimes were estimated by fitting an exponential decay model to the raw error data. The exponential decay model is a time-based generalization of the two-component model described in the previous section. In the exponential decay model, the number of remembered objects Y falls exponentially with time t, such that $Y(t) = \beta \exp(t/\tau)$, where τ is the mean lifetime and β is the number of encoded objects at $t = 0$. Memory quality at each duration, as quantified by the dispersion parameter of the corresponding von Mises distribution, was allowed to vary freely. A loose prior was placed over each parameter for the purposes of estimation. The prior on β was uniform over the full range, 0 to

the number of presented objects. The prior on bias was uniform over the full range, $-\pi$ to π radians. The prior on τ was log-normal with a mean of 20 s and a standard deviation of 2 ln units. The prior on the dispersion parameter of each von Mises distribution was log-normal with a mean of 7.4 and a standard deviation of 1 ln unit. The model was fit with MCMC using PyMC2 version 2.2 (Patil et al., 2010).

Strength of evidence for crossover

For each possible pairing of tested set sizes, we measured the strength of evidence in favor of a model where the forgetting function for the greater set size crosses over that for lesser set size (i.e., where it starts higher and ends lower) to one where it does not cross over. Strength of evidence was measured using the Bayes factor, the ratio of the posterior odds to the prior odds. The prior odds were $1:1$. The prior probabilities on model parameters were the same as in the previous section.

Fitting the evolutionary model to the Yang (1999) data. In Fig. 6, we fit the evolutionary model to data from Yang (1999)'s experiments on iconic memory (Yang, 1999). The model was fit by minimizing the squared error between the data and the model's predictions using Nelder–Mead simplex search over the model's parameters (Lagarias et al., 1998) (Appendix A).

1.5 Additional results

In the following subsections, we consider the effects of individual differences, practice, and differences between in-lab and online testing.

1.5.0 Individual differences

People vary considerably in the capacity of their working memory systems, and these individual differences are correlated with intelligence, reasoning abilities, and reading comprehension (Unsworth & Engle, 2005; Daneman & Carpenter, 1980; Fukuda et al., 2010; Kyllonen & Christal, 1990). Our analysis procedure, which combines data from multiple participants into a single super-subject, masks such variability, and it is therefore important to consider the ways in which the presence of individual differences might impact our results.

First, variability might alter the predictions of the classic, sudden death, or pure death models, undermining our claim that they fail to capture features of the empirical forgetting curves. Second, variability might alter the predictions of the proposed evolutionary model, undermining the logic whereby a tight fit between model and data lends support to the model. We examine each of these possibilities below.

1.5.1 Individual differences in the classic model

In the classic model, variability can arise through individual differences in the initial capacity K, which is the number of structures encoded in working memory. Through simulation, we inject individual differences by drawing $1 - g$, the probability of encoding each object, from a Beta distribution with parameters chosen to cover a reasonable range of variability. Figure 7 shows that individual differences of this sort have no impact on the resulting curves.

1.5.2 Variability in the pure death model

In the pure death model, variability can arise in two ways: through individual differences in the initial capacity β, and through individual differences in the mean lifetime τ. Variability in β is modeled

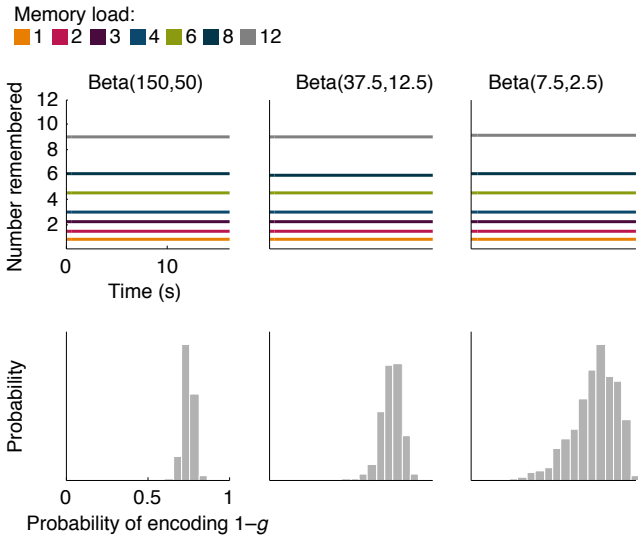


Figure 7: Individual differences in the classic model. The bottom row shows histograms of 1-*g*, the probability of successfully encoding an object in working memory. The top row shows the resulting forgetting functions, averaged over participants. Moving rightward, columns have greater individual differences.

in the same way as it is in the classic model. Through simulation, we inject variability into the mean lifetime by drawing t from a log normal distribution. Figure 8 shows that variability in β has no impact on the resulting curves and that variability in τ bends each curve, but does not change the relationship between them, which would be needed to reproduce the effects of load-dependent stability or crossover.

1.5.3 Variability in the sudden death model

In the sudden death model, variability can arise in three ways: through individual differences in (1) the initial capacity β, (2) the mean lifetime τ, and (3) the length of the window of initial stability t_d. Variability in β and τ are modeled in the same way as in the classic and pure death models. Through simulation, we inject variability into t_d by drawing it from a log normal distribution. As before, variability in β has no impact on the resulting curves. Figure 9 shows that variability in τ has the same

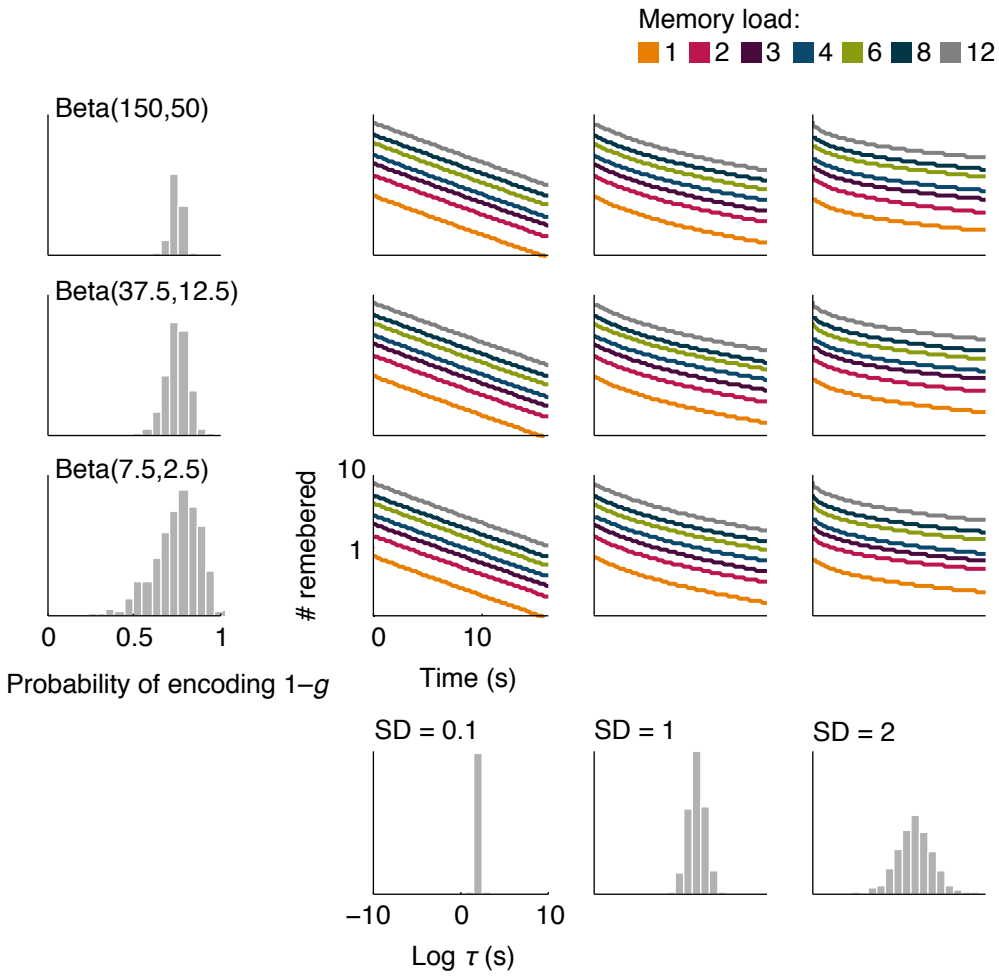


Figure 8: Individual differences in the pure death model. The leftmost column shows histograms of 1-g, the probability of successfully encoding an object in working memory. The bottommost row show histograms of τ, the mean lifetime. There are nine plots, one for each pair of distributions on 1-g and τ. Moving rightward, columns have greater individual differences in τ. Moving downward, rows have greater individual differences in 1-g. The y-axis is logarithmic to highlight shifts away from an exponential function (a straight line).

effects as it does in the pure death model and that variability in t_d softens the corner occuring at time points directly before and after the cutoff. As with the pure death model, these individual differences change the shape of the curves, but do not impact the relationship between them.

1.5.4 Variability in the evolutionary model

In the proposed evolutionary model, variability can arise in three ways: through individual differences in (1) the number of quanta N, (2) the duration of one time step t_{step}, or (3) the stability threshold s. Through simulation, we inject variability into each parameter and observe the effects on the predicted forgetting functions. Drawing N from a (discretized) normal distribution, we find that individual differences have a greater benefit to high memory loads than to low loads, thereby leading to a slight weakening of the crossover effect and load-dependent stability (Fig. 9). However, a crossover is seen even with high levels of individual differences (SD of $\pm$ 2 ln units). Next, drawing the stability threshold from a discrete uniform distribution, we again find that individual differences have a greater benefit to high memory loads than to low loads, with considerably less crossover, but only a miniscule effect on the presence of load-dependent stability (Fig. 10). Lastly, drawing t_{step} from a log normal distribution, we once again find the same result, with slight weakening of both load-dependent stability and crossover (Fig. 10). Together, these results suggest that the predictions of the proposed evolutionary model are tolerant to large individual differences in N, moderate individual differences in k, and large individual differences in t_{step}.

1.5.5 The effects of practice

Here, we consider the effects of practice by tracking performance as it changes over the course of the experiment's 70 trials (Fig. 13). The number of remembered objects dropped slightly (linear corre-

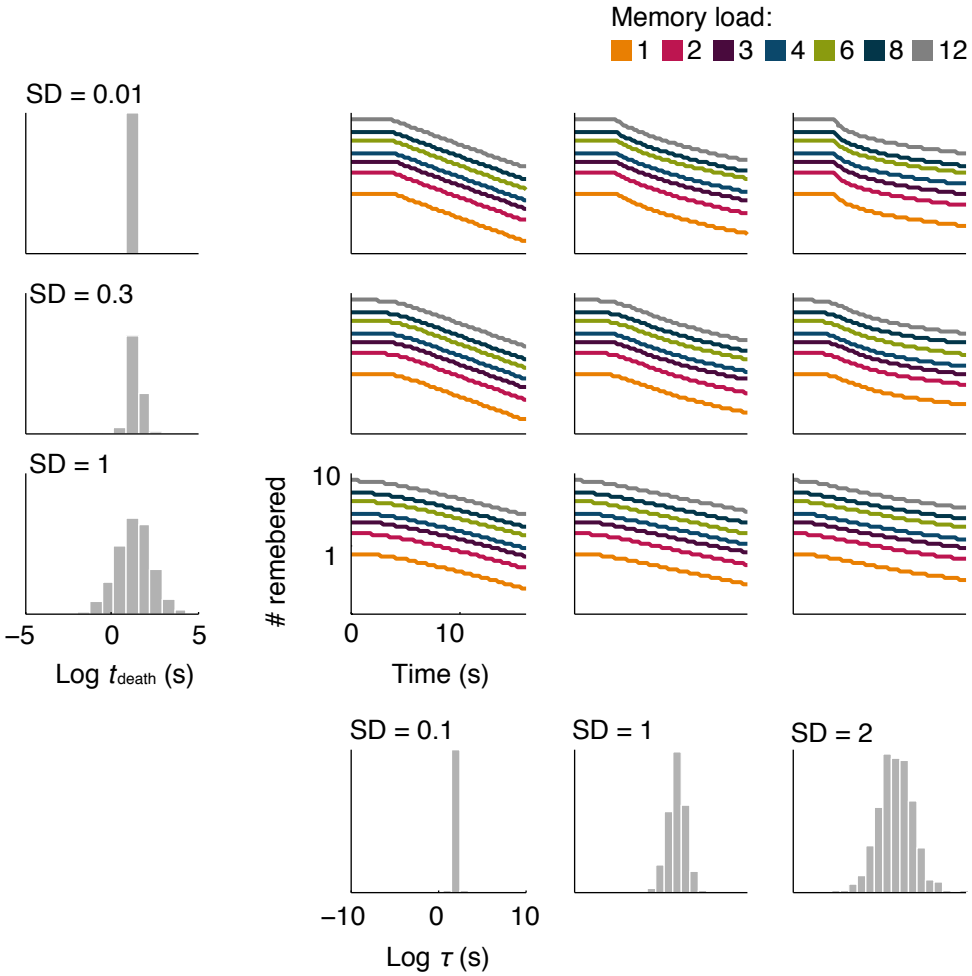


Figure 9: Individual differences in the sudden death model. The leftmost column shows histograms of t_{death}, the probability of successfully encoding an object in working memory. The bottommost row show histograms of τ, the mean lifetime. There are nine plots, one for each pair of distributions on t_{death} and τ. Moving rightward, columns have greater individual differences in τ. Moving downward, rows have greater individual differences in t_{death}. The y-axis is logarithmic to highlight shifts away from an exponential function (a straight line).

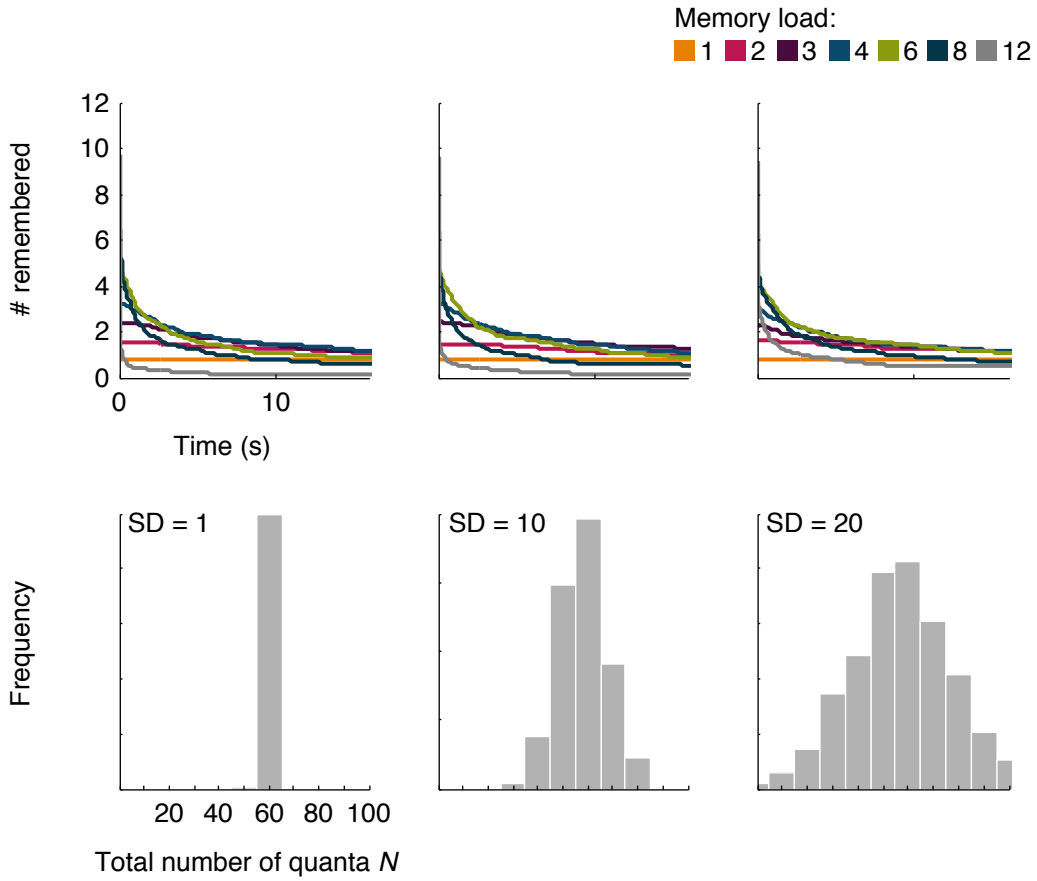


Figure 10: Individual differences in the number of quanta of the evolutionary model. The bottom row shows histograms of *N*, the total number of quanta. Moving rightward, columns have greater individual differences in *N*. The top row shows the corresponding forgetting functions.

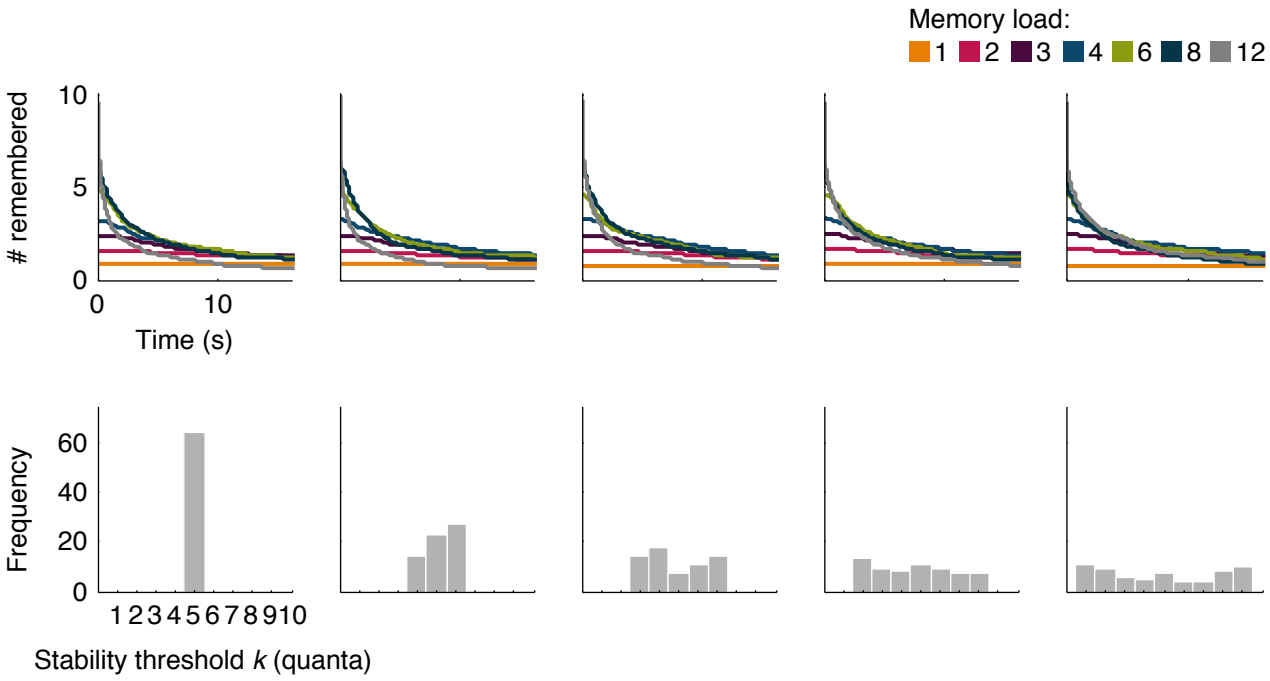


Figure 11: Individual differences in the stability threshold. The bottom row shows histograms of k, the stability threshold. Moving rightward, columns have greater individual differences in k. The top row shows the corresponding forgetting functions.

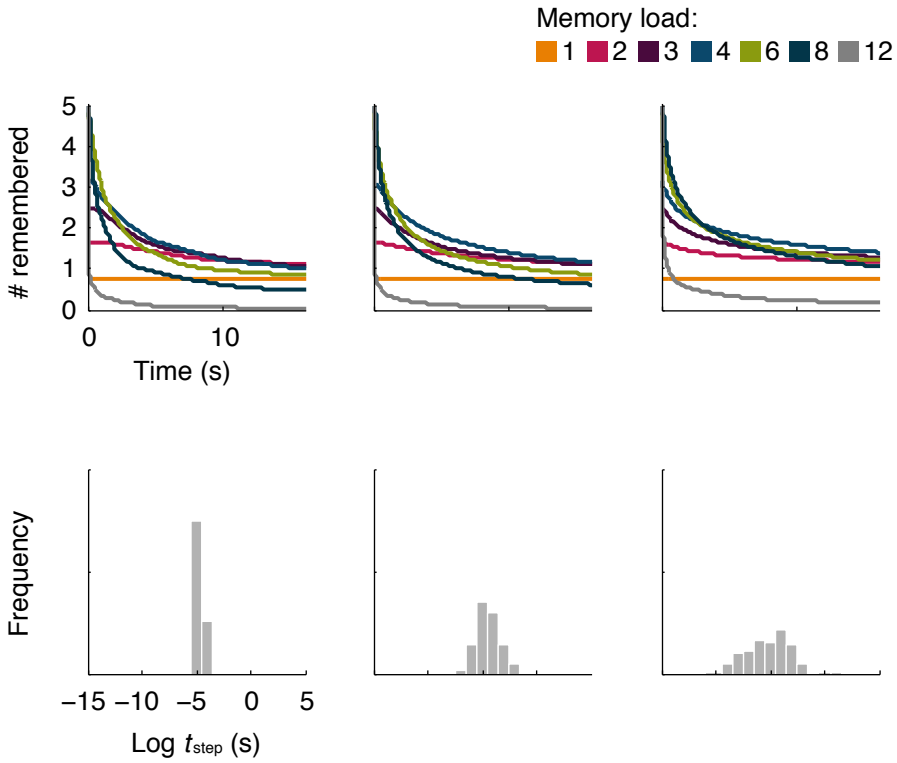


Figure 12: Individual differences in the rate of degradation. The bottom row shows histograms of t_{step}, the duration of a time step. Moving rightward, columns have greater individual differences in t_{step}. The top row shows the corresponding forgetting functions.

lation, r=–0.30, p = 0.013), roughly 0.01% per trial (slope of linear regression, –0.002 object/trial; intercept, 2.2 objects). There were no significant changes in memory quality (r=0.15, p=0.22) or bias (r=0.03, p = 0.81). This suggests that our training procedure was sufficient for participants to perform the task well.

1.5.6 Laboratory replication

We performed an in-laboratory replication of the main experiment in a group of six naive participants. The task used loads 1, 2, 3, and 6, and durations 0.125, 0.25, 0.5, 1, 4, and 10 s. Participants were each tested for 6-8 sessions of 360 trials, 15 trials per condition (pairing of duration and memory load), in random order. We observe similar results of load-dependent forgetting and overreaching, with a crossover for the highest load.

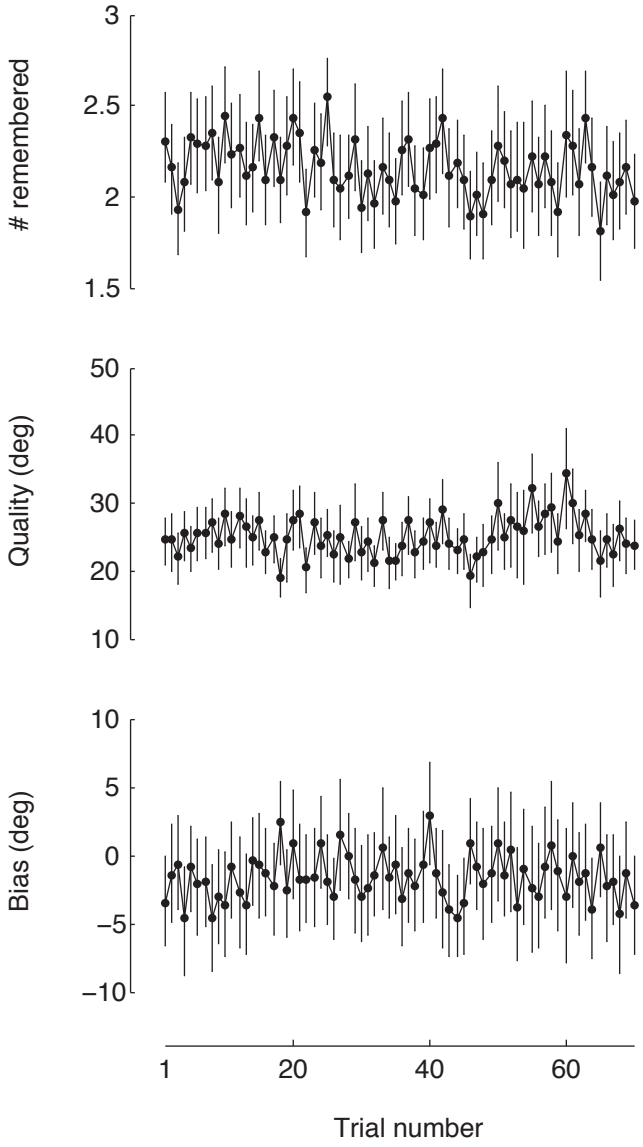


Figure 13: The effects of practice. Each subplots shows changes in performance as a function of trial number. The upper plot shows changes in the number of remembered objects. The middle plot shows changes in memory quality (lower is better). The lower plot shows changes in bias.

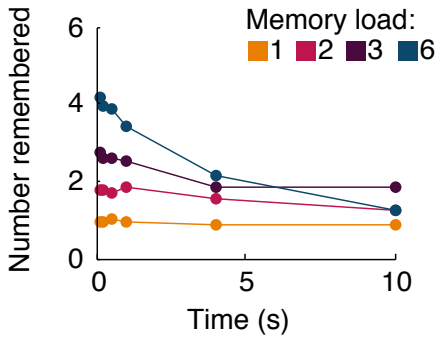


Figure 14: Replication in the lab. We replicated the online experiments in the lab with a group of six participants. The tested memory loads were 1, 2, 3, and 6. The tested durations were 0.125, 0.25, 0.5, 1, 4, and 10 s. Participants were each tested for 6-8 sessions of 360 trials, 15 trials per condition (pairing of duration and memory load), in random order. Data were fit with a hierarchical version of the 2-component model of Zhang & Luck (Zhang & Luck, 2008). The plotted data is the population mean. Data were fit using the MemToolbox 1.0.0. (Suchow et al., 2013)

2

Looking inwards and back: realtime monitoring of visual working memories

2.0 Abstract

Confidence in our memories is influenced by many factors, including beliefs about the perceptibility or memorability of certain kinds of objects and events, as well as knowledge about our skill sets, habits, and experiences. Notoriously, our knowledge and beliefs about memory can lead us astray,

causing us to be overly confident in eyewitness testimony or to overestimate the frequency of recent experiences. Here, using visual working memory as a case study, we designed a task that strips away all these potentially misleading cues, requiring observers to make confidence judgments by directly assessing the quality of their memory representations. We show that individuals can monitor the status of a memory as it degrades over time. Our findings suggest that people have access to information reflecting the existence and quality of their memories, and furthermore, that they can use this information to guide their behavior.

2.1 Introduction

METAMEMORY IS AN AWARENESS OF OUR MEMORIES and the systems that store them. We use metamemory to determine that we are uncertain ("I can't remember where I parked my car"), to ask for a reminder ("When's that appointment, again?"), and to form beliefs about our ability to remember certain kinds of information ("I'm good with names") (Flavell & Wellman, 1977). For better or for worse, judgments of confidence in our memories are influenced by many factors. These include general knowledge in the form of beliefs about ourselves and what we find memorable, as well as more specific knowledge derived from our previous experience with the task at hand (Koriat, 1997; Schwartz, 1994; Schwartz et al., 1997). However, the cognitive mechanisms underlying metamemory judgments are poorly understood.

Looking towards research in other areas of metacognition, where a variety of confidence mechanisms have been explored in detail, may provide a clue about the workings of metamemory. For example, in the case of perceptual discriminations, one simple mechanism for judging confidence is to use visual cues associated with uncertainty (e.g. faintness and blur), alone or in combination, as a proxy for confidence (Barthelmé & Mamassian, 2010). Then, when asked to identify an object that

appears blurry or faint, an observer using this mechanism will report having low confidence because blurriness and faintness are stimulus features typically associated with uncertainty. Importantly, cue-based mechanisms like this one draw on static information about the stimulus and decision-maker, rather than directly accessing an internal representation or decision making process. In contrast, an alternative class of mechanisms has been proposed that can compute realtime measures of perceptual confidence (Kepecs et al., 2008). Realtime mechanisms are notable because, rather than relying on externally observed cues, they monitor internal states as they change over time (Kepecs et al., 2008). These monitored states may be those of decision variables associated with the task, or those of underlying representations that store uncertainty explicitly — e.g., as probability distributions over past states of the environment (Barthelmé & Mamassian, 2010).

Here, using visual working memory as a case study, we designed a test to isolate realtime mechanisms underlying the evaluation of confidence in memory. Our task was a variant of a popular test of visual working memory in which participants view a display of colorful dots and then, shortly thereafter, report the color of a dot selected for them at random (Wilken & Ma, 2004). In our variant, however, instead of requiring participants always to report the color of a randomly selected dot ("random" condition), they were sometimes afforded the opportunity to report the color of the object they remembered best ("best" condition). Choosing the best-remembered object requires an inward-looking comparison of the relative quality of multiple memories, and is a within-trial analogue to the opt-out procedure used extensively in studies of human and animal metacognition (Smith et al., 1995, 1997; Tanaka & Funahashi, 2012). To strip away nearly all the usual sources of metamemory information — general knowledge, stimulus-based cues, and time-based fluctuations in attention and arousal — we compared memory for an object in a display when it was chosen by the participant as their best remembered object to when it was chosen by the experimenter at random. This procedure

enables us to isolate a form of monitoring whereby an individual tracks the status of a memory as it degrades over time.

2.2 Methods

2.2.0 Logic of the task: isolating realtime monitoring

Participants were asked to remember the colors of a set of colorful dots, and then either to report the color of a randomly selected dot, or to make an inward-looking decision by choosing the dot they remembered best and reporting its color. Because our goal was to isolate the contribution of realtime monitoring to this decision, the experimental procedure combined multiple techniques to eliminate confounding sources of metamemory cues:

Stimulus-based cues. Of principal interest was whether memory would be more accurate for the best-remembered object than for a randomly selected item. However, the best-remembered object might be preferred for reasons that do not require a realtime assessment of memory quality. For example, a participant may prefer a particular color (say, red), and pay more attention to it. Or perhaps they find it more memorable, preferring to select red objects whenever the chance arises (e.g., see Morey, 2011). This is a form of metamemory, but it does not reflect realtime monitoring.

To eliminate display factors such as these, we used a double-pass procedure (Burgess & Colborne, 1988; Gold et al., 2005; Green, 1964) where participants are asked to remember the same display (i.e., a particular color and arrangement of dots) multiple times. We then compare memory performance for a particular object when it was randomly selected by the experimenter versus when it was chosen by the participant as the object that was best remembered. Any advantage for the preferred object is thus unlikely to depend on stimulus-based factors, which are held constant across the conditions.

Tradeoffs in encoding or maintenance. When viewing the stimulus, a participant's attention might wander due to either an explicit strategy or accidental drift, causing one object to be encoded more vigorously than another. This could also happen during maintenance, shifting priority from one object to another after the stimulus has already disappeared. These imbalances, if known to the participant, could be used as a proxy for memory fidelity because an ignored object is unlikely to be remembered well. To avoid this, our approach was to design the experiment to minimize tradeoffs and then to perform a separate tradeoff detection procedure once the experiment was over.

Our design interleaved two different types of trials in random order. On half of the trials, participants reported the color of the dot they remembered best, while on the other half, they reported the color of a dot selected at random by the experimenter. Interleaving the trial types encourages participants to remember all the dots, because they do not know which dot will be tested.

This procedure mitigates the tradeoffs, but it is possible that tradeoffs were still present. To determine this, we used an additional tradeoff detection procedure: After reporting either the best-remembered object or a randomly selected one, the participant was then asked to report the color of a second dot on the display, selected at random. These two reports are used together to detect tradeoffs with the detection procedure introduced in Fougnie et al. (2012), which relies on the fact that tradeoffs introduce dependencies in the measured quality of representations of objects on a display: if one object is remembered particularly well, it comes at the expense of the others. Therefore, the detection procedure compares performance for the first-reported object in two conditions: where the absolute error for the second-reported item was (1) above or (2) below the median absolute error across all second reports. A tradeoff is revealed by first reports being more accurate when second reports are worse. The tradeoff analysis was limited to trials where the first object was randomly selected to avoid the dependency introduced by the participant's selection.

Fluctuations in attention or arousal. Attention, arousal, and effort can fluctuate from moment to moment (Kahneman, 1973). Most studies of metamemory ask for ratings or judgments about a particular memory at a particular moment, and so momentary fluctuations can affect performance and therefore contribute to metamemory decisions. In our task, we asked participants to make a relative judgment about the quality of simultaneously held memories, such that the confidence judgment could never depend on the overall state of attention or arousal, which would apply equally to all objects on the display. Similarly, we randomized the order of the two trial types, which prevents momentary fluctuations from systematically affecting one trial type over the other.

2.2.1 Implementation of the task

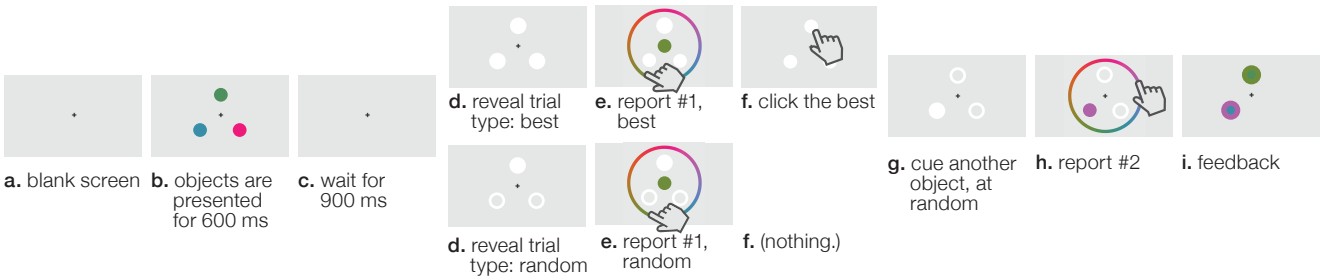


Figure 15: Timeline of one trial of the double-pass metamemory task. First, the participant sees a set of colorful dots and is asked to remember them for 900 ms (a–c). Then the type of trial is revealed: the participant will either choose, reporting the object they remember best (d, top, "best"), or mandatorily report the object highlighted for them (d, bottom, "random"). These two trial types are interleaved in random order. Thus the participant does not know the trial type until it is revealed in panel (d). Therefore the participant must encode all the items. Once the trial type is revealed, the participant reports the color by selecting it from the color wheel (e). A different item is then selected at random (g) and the participant reports its color (h). This second report is later used in an assay of strategic or accidental trade-offs (see Methods and Supplemental Information). Finally, the participant receives feedback (i). These steps, which constitute one trial, are repeated hundreds of times in two rounds. In the second round, the displays used in the two conditions ("best" or "random") are swapped, producing a double-pass procedure where, unbeknownst to the participant, in the second round the "randomly" chosen objects are in fact those chosen by the participant in the first round.

At the beginning of each trial, the participant fixated a small dot in the center of the screen. Then the stimulus (a set of three colorful dots) appeared for 600 ms. Next, the trial type was revealed to

the participant through a display that contained a cue in each of the locations of the test objects. If it was a trial where the participant was asked to report a specific object, that object was highlighted as a filled circle among open circles. On the other hand, if it was a trial where the participant reported the best-remembered object, all the objects appeared filled in. Then a color wheel with all the possible colors appeared and the color of the best-remembered object was reported. Finally, the participant used a mouse-controlled cursor to select which object was best remembered. After this first report, the participant was asked to report the color of a second object selected at random from the two that remained. The reporting procedure was the same. Feedback was provided at the end of each trial. The feedback screen, which appeared for 1000 ms, showed the actual color (inner ring) and the reported color (outer ring) for both of the tested objects (Fig. 1i).

2.2.2 Stimuli and presentation

Each dot had a radius of 0.4° of visual angle. The dots were arranged in a ring with a radius of 3.8° and centered on the display. The color of each dot was drawn uniformly from a circle cut out of the CIE 1976 $L^*a^*b^*$ color space, centered at $L = 54$, $a = 18$, $b = -8$, with the constraint that the magnitude of each display's mean hue vector was 0.35. This decreases grouping cues and reduces imbalances in appearance across displays. After the stimulus disappeared, there was a 900 ms retention interval. Stimuli were rendered by MATLAB with the Psychophysics toolbox (Brainard, 1997; Pelli, 1997), and presented on a 1920×1200 LCD screen at 60 Hz, 38 pixel/cm, positioned 60 cm from the participant.

2.2.3 Participants

Twelve people between the ages of 18 and 31 participated. They all had normal or corrected-to-normal visual acuity. The protocol, approved by the Committee on the Use of Human Subjects in Research under the Institutional Review Board for the Faculty of Arts & Sciences, was carried out in accordance with the provisions of the World Medical Association Declaration of Helinski.

2.2.4 Data analysis

To quantify memory performance, for each participant and condition we separately fit a variable-precision model to the data (Fougnie et al., 2012; van den Berg et al., 2012). This model supposes that each object on the display is either remembered or forgotten, and that the quality with which objects are remembered can vary. Specifically, the variable-precision model assumes that objects are remembered with some probability, and that if remembered, errors in recall are distributed according to a von Mises distribution centered around zero (though perhaps with some bias) and whose spread (s.d.), a measure of memory fidelity, is itself distributed according to some higher-order distribution, assumed here to be a zero-truncated normal.

We also considered a simpler fixed-precision model that did not allow memory quality to vary, as well as an extension to it that allows for the possibility that the participant will "swap" items, erroneously reporting an item that was not the target (Bays et al., 2009). Analysis was performed with MemToolbox 1.0.0 (Suchow et al., 2013). Analysis scripts and data are available as Supplementary Material.

2.3 Results

We found that participants can use realtime monitoring to make metamemory judgments. Figure 2 shows estimates of guessing rate (left panel) and precision (middle panel), averaged across participants. Individual participant results are shown (right panel) for items that were chosen as the best remembered (circles) versus those same items when they were selected at random (squares). Observers performed better in both guessing rate and memory precision when they chose to report the object, versus when the object was randomly selected. When asked to report the color of the best-remembered object, participants remembered it 92±2% (mean ± sem) of the time and with fidelity of 20.8±1°. When those same displays were presented in the second session and participants were forced to report the same object that they had previously picked, they performed worse, remembering it 71±3% of the time and with a fidelity of 23.8±2° (paired samples t-test, $t(11) = 7.5$, $p = 1.2 \times 10^{-5}$ and $t(11) = -2.5$, $p = 0.03$, respectively). This worsening across exposures happened despite overall performance being comparable in the two rounds (difference of 0.6° in fidelity from the first to second round, paired samples t-test, $t(11) = 0.51$, $p = 0.62$; difference of 0.4% in guess rate, paired samples t-test, $t(11) = 0.14$, $p = 0.89$).

Using the trade-off detection procedure described in Methods, we tested for trade-offs in the encoding or maintenance of items, but found none (guess rate for above vs. below median split: 20.5 vs. 19.4%, paired sample t-test, $t(11) = -0.68$, $p = 0.51$; fidelity: 22.1° vs. 22.4°, $t(11) = 0.42$, $p = 0.68$).

We performed additional analyses to determine whether our results are robust to assumptions about the structure of visual memory representations. Specifically, we repeated the above analysis using three other models: the two-component mixture model introduced by Zhang & Luck (2008), the "swap" model introduced by Bays et al. (2009), and a one-component model without guessing.

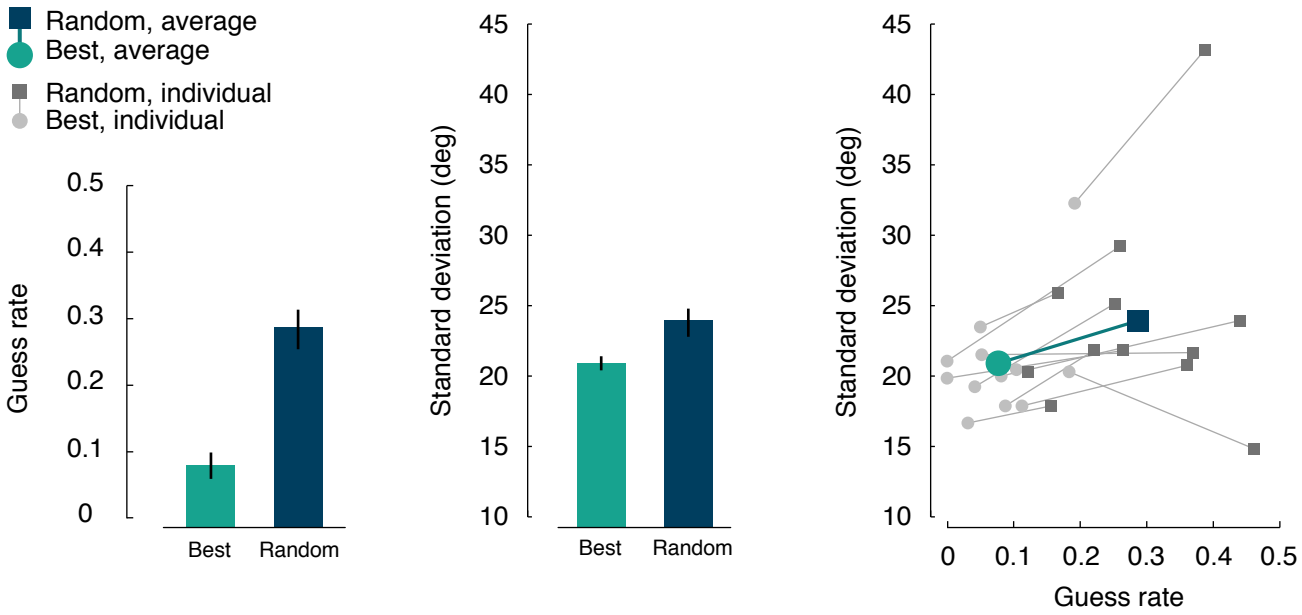


Figure 16: The contribution of realtime monitoring to judgments of confidence in memory. We compare performance across two conditions, one where the participant selects an item as the one that was best remembered from that display (squares), and another where the participant reports that same item because they were required to (circles). They performed better when they made the choice, which implies that participants can use realtime monitoring to guide their selections in the task, picking out the one they remember best.

2.3.0 Results: Two-component mixture model

We considered a two-component model without variability in precision. When asked to report the color of the best remembered item, participants remembered it 91±2% (mean ± sem) of the time and with a fidelity of 20.5±1° (Fig. S1). When those same displays were presented in the second session and participants were forced to report the same item that they had previously picked, they performed worse, remembering it 68±4% of the time and with a fidelity of 22.3±1° (paired samples t-test, $t(11) = 6.8$, $p = 2.8 \times 10^{-5}$ and $t(11) = -2.7$, $p = 0.02$, respectively). This across-exposure worsening happened despite the presence of small practice effects, which caused performance for the best-remembered items to improve from the first session to the second (a gain of 1.8° in fidelity, paired samples t-test, $t(11) = 2.12$, $p = 0.0572$; an insignificant 0.014 improvement in guess rate, paired samples t-test, $t(11) = 0.70$, $p = 0.50$). No tradeoffs were detected (guess rate for first object according to whether absolute error on second object was above vs. below median: 0.36 vs. 0.34, paired sample t-test, $t(11) = -0.77$, $p = 0.46$; fidelity: 23.8° vs. 25.0°, $t(11) = 0.62$, $p = 0.55$). Analysis included a bias parameter; its value did not differ significantly between the conditions (-0.47° vs. −0.48°, $t(11) = -0.02$, $p = 0.99$).

2.3.1 Results: Adding a swap component

When asked to report the color of the best-remembered item, participants remembered it 93±2% (mean ± sem) of the time and with a fidelity of 20.4±1° (Fig. S1a). When those same displays were presented in the second session, they performed worse, remembering the item 70±4% of the time and with a fidelity of 22.2±1.1° (paired samples t-test, $t(11) = -6.8$, $p = 2.8 \times 10^{-5}$ and $t(11) = -2.4$, $p = 0.03$, respectively). The swap rate, the probability of mistakenly reporting one of the uncued objects instead of the target, was 2% in both conditions and was no better for the best-remembered object

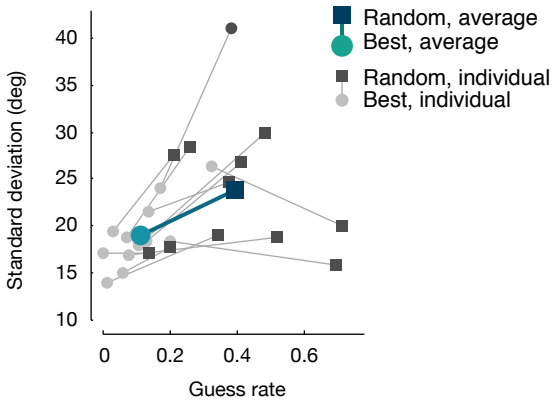


Figure 17: Comparing performance in the random probe and choose-the-best conditions. Performance is analyzed using the two-component model of Zhang & Luck (2008).

(paired samples t-test, $t(11) = -0.06$, $p = 0.95$). The swap rates observed here, which are roughly three times lower than those measured in a previous study using similar methods (Bays et al., 2009), is perhaps due to differences in stimuli. Specifically, our stimuli were generated with a procedure that reduced grouping of similar colors (see Methods), which is likely to be a major contributor to swap errors. Analysis included a bias parameter; its value did not differ significantly between the conditions ($-0.36°$ vs. $-0.54°$, $t(11) = 0.27$, $p = 0.79$).

2.3.2 Results: No guessing

When asked to report the color of the best-remembered item, participants remembered the item with a fidelity of $3 1 \pm 1°$ (Fig. S2). When those same displays were presented in the second session and participants were forced to report the same item that they had previously picked, they performed worse, remembering the item with a fidelity of $54 \pm 1°$ (paired samples t-test, $t(11) = -8.3$, $p = 4.5 \times 10^{-6}$). Analysis included a bias parameter; its value did not differ significantly between the conditions ($-0.54°$ vs. $-0.98°$, $t(11) = 0.29$, $p = 0.78$).

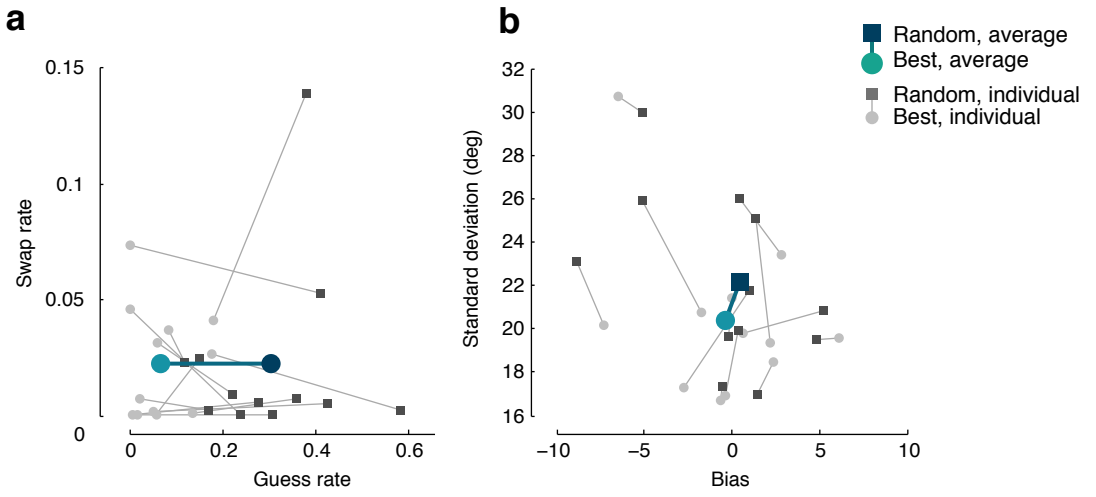


Figure 18: Comparing performance in the random probe and choose-the-best conditions. Performance is analyzed using the "swap" model of Bays, Catalao, & Husain (2009).

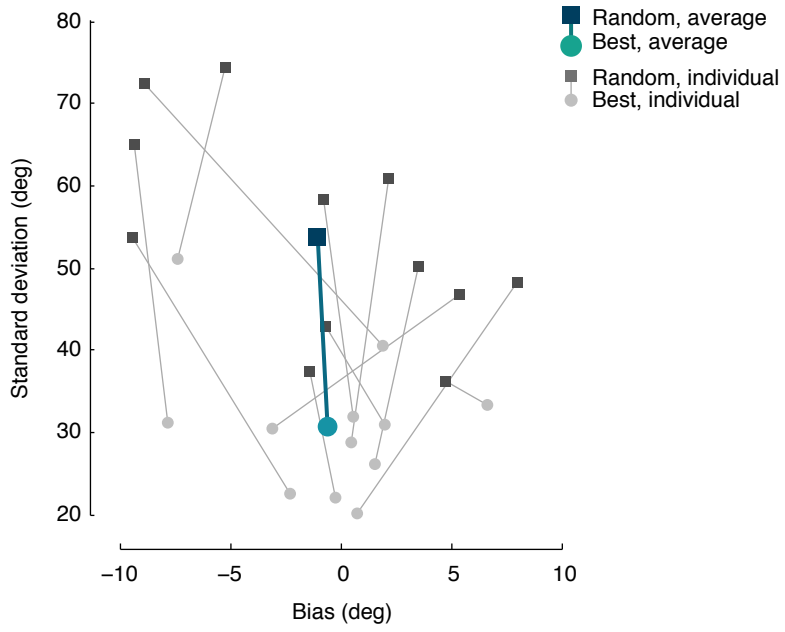


Figure 19: Comparing performance in the random probe and choose-the-best conditions. Performance is analyzed using a model without guessing.

2.4 Discussion

The results of this experiment suggest that realtime monitoring can be used to make judgments of confidence in working memory. We found that participants remembered an object's color more accurately when it had been chosen as the one they remembered best than when that same object, presented in the context of the same display, was selected at random by the experimenter. Even after eliminating other sources of metamemory information, such as stimulus-based cues and tradeoffs in encoding and maintenance, we found that observers were able to assess the quality of their memories in realtime and could use that information to guide their behavior. Thus, the present results reveal a strategy that can monitor both the existence and quality of representations in visual working memory.

This form of metamemory requires access to information that indexes the quality of memories. Though it is unclear what mechanism provides realtime access to memories, research on uncertainty in decision making may provide a clue:

A number of simple mechanisms have been proposed that might support realtime measures of confidence in perceptual judgments (Kepecs et al., 2008). These mechanisms involve accessing decision variables that contribute to a decision. For example, in a race model, where evidence simultaneously accumulates for each alternative choice (Gold & Shadlen, 2007), confidence can be estimated by measuring the difference in accumulated evidence for each alternative at the moment the choice is made. High confidence is appropriate when there is a big imbalance in accumulated evidence. Analogous mechanisms may be at play in the monitoring of visual working memories. For example, confidence in a memory could be computed by comparing the accumulated evidence for the winning decision (i.e., stimulus value) to the average of the others. Alternatively, monitoring may be accom-

plished through more indirect means, using a process akin to an availability heuristic. Suppose, for instance, that less precise memories are more difficult to access (e.g., see Brady et al., 2013). Then, the participant can use a metamemory routine that attempts to access a memory and terminates if nothing has been accessed after a fixed amount of time. The time to access then serves as a proxy for memory fidelity and can be used to inform confidence.

Whatever the mechanism might be, the present results demonstrate it is possible to access the current state of a memory and to use that information to guide behavior. The existence of realtime monitoring mechanisms has important implications not only for our understanding of metamemory, but also for theories of the representational format of visual working memory. Leading models of visual working memory assume that memory limits are determined purely by the availability of a limited commodity: once you run out of memory slots (Awh et al., 2007; Zhang & Luck, 2008; Luck & Vogel, 2013) or memory resources (Alvarez & Cavanagh, 2004; Bays et al., 2009; Wilken & Ma, 2004; Ma et al., 2014), you can no longer store additional objects in memory. However, in addition to possible commodity-based limits, there is emerging evidence that visual working memory is also limited by interference, degradation, or decay that leads to the gradual loss of information over time (Fougnie et al., 2013). This decrease in quality appears to reflect a process that operates independently across items (Fougnie et al., 2013). Such degradation leads to substantial variability in the quality of memories across objects, with some objects remembered very well, others remembered poorly, and others completely forgotten. The present results provide evidence for the presence of variability in memory quality (Fougnie et al., 2012; van den Berg et al., 2012) and show that this variability cannot be explained by stimulus differences or by differential allocation of attention within or across trials.

2.5 Conclusion

Most research on metacognition has focused on perception and long-term memory, exploring how people assess uncertainty about their current perceptions and distant memories. Theories of metamemory have thus focused on how multiple sources of information influence judgments of confidence, including several static factors such as how memorable the material is, or judgments about our own abilities. Thus, it has been difficult to assess whether and how participants have access to information that directly indexes the quality of a memory. In the present study, we developed a new method to strip away these static factors, enabling us to isolate realtime metamemory mechanisms, taking advantage of the fact that working memories appear to degrade stochastically over time. We found that observers appear to have access to the current state of their memories, and can use that information to guide their behavior in an ongoing task. These findings open the door to new explorations into the nature of the cues that enable realtime memory monitoring and into the impact of metamemory in complex cognitive processes that rely on working memory.

3

Controlling working memory maintenance

Memories can be remembered and forgotten intentionally through the process of directed forgetting and directed remembering, which prioritize some experiences over others for later access (Muther, 1965; Bjork et al., 1968). These directed processes are closely related to cognitive control and to the top-down processes that determine our conscious thoughts from moment to moment (Macrae et al., 1997). At times, these control processes can backfire, causing unwanted thoughts and

memories to linger despite our best intentions (Wegner, 2009).

One of the clearest demonstrations of directed remembering in short term visual memory comes from recent work by Williams et al. (2012). In the study, participants performed a task in which they were asked to remember the colors of one or two colorful squares. On trials when two objects were presented, a cue would sometimes appear 1 s into the retention interval, informing the participant of which object would be tested a few seconds later. This hint afforded participants the opportunity to alter their maintenance behavior accordingly (Fig. 20). Williams et al. found that participants performed better when the cue was present than when it was absent: the probability of remembering the tested object increased from 85% to 91% and fidelity improved from 12.5° to 10.0° (Fig. 21). These results show that visual memory maintenance is flexible — it can be shifted, prioritizing one stored representation over another.

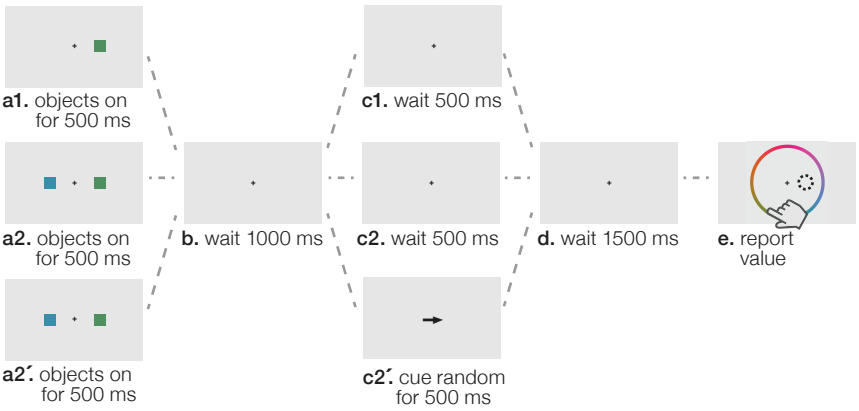


Figure 20: Method for revealing directed remembering in visual working memory (Williams et al., 2012). There are three conditions. In condition 1, one object is presented. In condition 2, two objects are presented. In condition 2′, two objects are presented and the tested object is cued early in the retention interval.

Directed remembering is one kind of flexible maintenance behavior, but other kinds are possible, too. Specifically, the results of Chapter 2, in which we showed that people can use metamemory to access the current strength of their visual memories, suggest the possibility of maintenance strategies

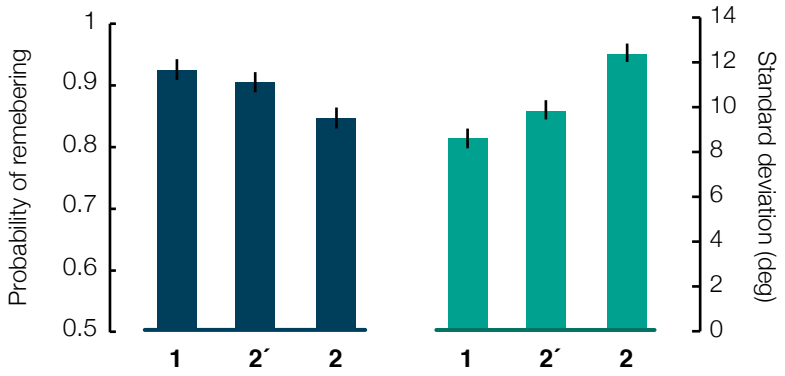


Figure 21: Directed remembering in visual working memory. The probability of remembering the tested object is highest when there is only one object (1), lower when there are two but one of them is cued (2′), and even lower when there are two but none is cued (2). (Right) Fidelity shows the same pattern of results (NB: here, a lower standard deviation means better performance). Data is replotted from (Williams et al., 2012).

that make use of metamemory. For example, one strategy that uses metamemory targets the weakest accessible memory. Another gives preference to memories that are already strongly represented. These are just two of many such maintenance strategies that make use of metamemory to guide maintenance. What is the space of possible maintenance strategies, and how successful is each of them in retaining visual information over short durations?

One factor affecting that success is the cost of directed maintenance: prioritizing one representation means neglecting others. This trade-off can be observed by comparing performance for prioritized vs. neglected representations (e.g., see Fig. 21). However, the results of Chapter 1 predict that there will also be a second cost associated with direct maintenance. Specifically, in Chapter 1 it was proposed that maintenance is limited by a stability threshold, such that only memory representations with a strength that surpasses the threshold are accessible to the maintenance process and can receive its benefits. Memory representations that are too weak can not be maintained. The second cost derives from this stability threshold: redistributing maintenance brings the neglected objects closer to the edge. This cost would be apparent when comparing the overall efficiency of directed remem-

bering across loads. Specifically, when there are few objects to maintain, there should be a relatively efficient trade-off between objects, such that the overall amount that is remembered remains constant, only being shifted from one object to another. However, as the number of objects increases, efficiency should drop because the act of shifting maintenance brings neglected representations closer and closer to the stability threshold and thus more susceptible to total loss.

The goal of this work is to define the space of possible maintenance strategies, considering both the constraints (e.g., the stability threshold) and the sources of information that are available (e.g., metamemory). In light of this goal, the plan of the chapter is as follows. Section 1 presents an experiment on the efficiency of directed remembering, showing that this efficiency depends on load. Section 2 presents an experiment on self-directed remembering, in which a person uses metamemory to guide selection of the target of maintenance. Section 3 describes a formal framework, the Markov decision process, which is useful for describing sequential decision making and will allow us to define the space of possible maintenance strategies and to specify the minimal mechanisms needed to produce the flexible maintenance behaviors that were observed. The section then presents two classes of maintenance policies — conditional and unconditional — and shows how conditional policies can produce directed and self-directed remembering. Finally, it extends the model to cases of imperfect metamemory, describing memory maintenance in a partially observable mind – i.e., situations when the maintenance system has incomplete or uncertain information about the current status of actively held memories.

3.1 Exp. 1: Efficiency of directed remembering

In the experiment that follows, we tested the prediction that the efficiency of directed remembering decreases with load. Participants performed a directed remembering task. We manipulated whether

the cue was present and whether it was *valid* — faithfully indicating which object would later be tested. We also manipulated load. The undirected condition (no cue) provides the baseline measurement of how much is remembered by default. The prioritized (i.e., validly-cued) and neglected (i.e. invalidly-cued) conditions allow us to measure how much is remembered when maintenance is directed. We use these measurements to compute the efficiency of directed remembering, which is the proportion of information remembered in the undirected task that is remembered when maintenance is directed.

3.1.1 Methods

Participants

Eight people between the ages of 18 and 26 participated. They all had normal or corrected-to-normal visual acuity, normal color vision, and were naive to the purpose of the experiment.

Stimuli

Stimuli were 1, 2, 3, or 4 cubes. Each cube had three visible sides, one white, one grey, and one black, viewed either from above or below. There were thus twelve configurations (Fig. 22). Cubes were positioned 5° above, below, to the left of, or to the right of a central fixation mark. Stimuli are adapted from Alvarez & Cavanagh (2004).

Procedure

After receiving instructions, the participant initiated the experiment by clicking a mouse. On each trial, the objects appeared for 300 ms and then disappeared. The retention interval was 2.80 seconds. On half of the trials, a cue appeared 0.600 s into the trial. The cue was a small grey dot (0.1 deg of vi-

sual angle) flashed in the location of one of the cubes. The cue was valid on 75% of trials. On valid trials, the cued cube was always tested. On invalid trials, the tested cube was chosen uniformly at random from the other (non-cued) cubes. The tested cube's location was marked with a small cue, identical to the one that was presented during the retention interval. Two hundred milliseconds after this cue, a response screen containing all 12 possible cubes appeared. The participant selected the cube that they remembered having seen in the location that was tested. Selecting a response automatically initiated the next trial. No feedback was provided.

Analysis

We compare the efficiency of directed remembering across different loads. Efficiency is measured as the proportion of the number of objects (or features) remembered in an undirected setting that are remembered in a directed setting. Efficient directed remembering happens when the participant is able to shift the balance of memory maintenance at whim while maintaining the same total amount of information that they would have had they not shifted maintenance at all. Inefficient directed remembering happens when shifting the balance comes at the cost of fewer remembered objects (or features) in total. For example, if the participant remembers 2 of 4 objects in an undirected setting, but only 1 of 4 in a directed setting, that participant's efficiency is 1/2.

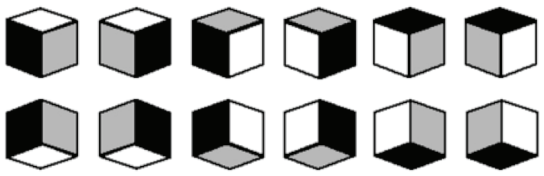


Figure 22: Cubes used in the directed remembering experiment.

Measuring efficiency requires converting performance on the task to a measure of how much was remembered. The specific conversion that is used depends on one's model of how participants store

memories. The most popular conversion assumes that the participant remembers k objects. If the tested object is among those that are remembered, the participant responds correctly. Otherwise the participant guesses blindly. This conversion assumes a "high-threshold" model of memory — either the participant remembers enough about the particular object to report it correctly, or they remember nothing about it at all (Luck & Vogel, 1997; Cowan, 2001; Rouder et al., 2008). Perfomance on the task is linear in k. It is also possible to relax this assumption by assuming that participants have graded memories of the presented objects. We suppose that the participant remembers each feature with some probability p; the participant responds by selecting randomly from among all possibilities that are consistent with what is stored in memory. Under these assumptions, we find that task performance is approximately linear in p (Fig. 23), such that efficiency measurements based either on the number of stored objects or on the number of stored features will be similar.

The efficiency of directed remembering is thus defined as the ratio $k_U/(k_P + k_N)$, where k_U, k_P, and k_N are the number of remembered objects in the undirected, prioritized (i.e., validly-cued) and neglected (i.e., invalidly-cued) conditions, respectively. The value of k is given by

$$k = L\left(\frac{P - 1/c}{1 - 1/c}\right), \tag{3.1}$$

where L is the load, c is the number of alternatives (i.e., possible cubes, here 12), and P is proportion correct on the task.

3.1.2 Results

Fig. 24 shows overall performance on the directed remembering task under various loads. As expected from previous studies, performance is best when there is only one object and declines with increasing load. A 4×3 repeated measures ANOVA was run with load (1, 2, 3, or 4 cubes) and di-

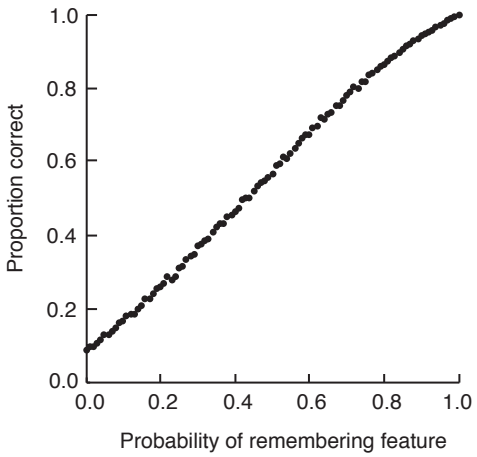


Figure 23: Task performance as a function of the probability of remembering each feature. Each point is the result of simulating 100,000 trials under the assumptions in the main text.

rectedness (neglected, undirected, or directed) as factors. There was a main effect of load ($F(1,11) = 14.08$, $p < 0.01$). Performance is nearly perfect (96% correct) when there is only one object, confirming that participants were attentive during the task. Estimates of capacity were roughly 0.9 across the whole task. Though these capacity estimates are uncharacteristically low for simple objects such as colorful squares and familiar shapes, they are typical of highly complex or unfamiliar objects (Alvarez & Cavanagh, 2004).

Critically, we find that the efficiency of directed remembering decreases as the load increases (Fig. 25). Efficiency was lower with 4 objects than with 3 (paired t-test, $t(7) = 3.0$, $p = 0.02$) and lower with 3 objects than with 2 (paired t-test, $t(7) = 4.6$, $p = 0.001$), declining from 0.87 to 0.55.

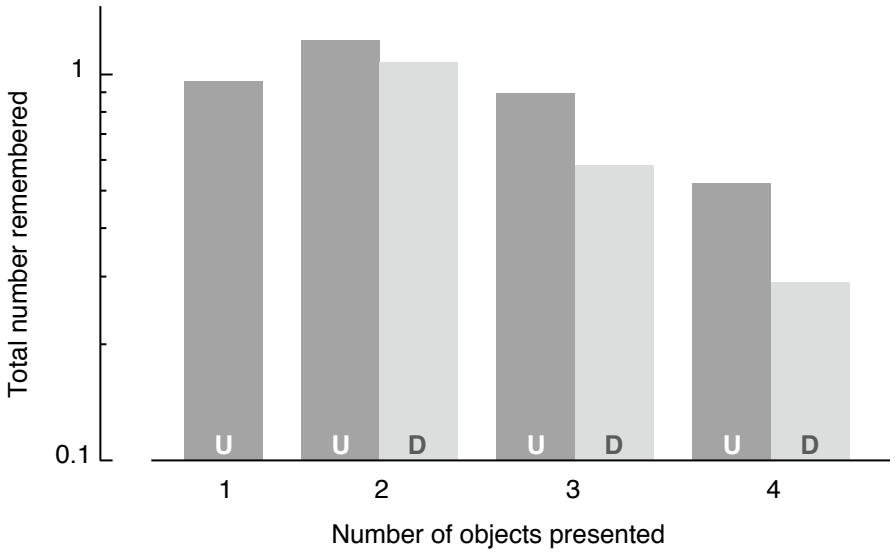


Figure 24: Performance on the directed remembering task. Numbers below the bars are the load, with the letters U and D marking the conditions where maintenance was undirected and directed, respectively.

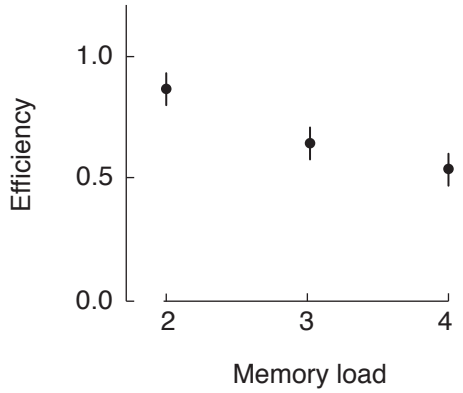


Figure 25: Efficiency of directed remembering compared across loads. Efficiency is defined only for loads greater than one. Error bars are 95% confidence intervals.

3.1.3 Interim discussion

The results of Experiment 1 showed that the efficiency of directed remembering depends on the memory load, revealing a second cost associated with directed maintenance. When multiple objects are held in mind, the act of biasing maintenance towards one of them is more detrimental to the others if there are many objects rather than only one or two. Our account of this effect lies in a stability threshold that governs memory maintenance. The greater the load, the weaker the individual memory. Thus at greater loads, directing maintenance away from a representation (and thus allowing it to weaken further) makes it more likely that it will cross the threshold of inaccessibility.

The drop in efficiency with load has implications for the selection of maintenance strategies that are well suited to a given task. When one representation is known to be the only one that will be relevant to future behavior, directing maintenance to it fully is the most sensible thing to do because the aforementioned costs are irrelevant — they affect only the neglected representations, not the one that is prioritized. However, when there is uncertainty about what information will be relevant to future behavior, both costs become relevant and it becomes sensible to weigh the benefits of directing maintenance towards the representations mostly likely to be useful against the costs of worsening the neglected representations and remembering less in total.

3.2 Exp. 2: Self-directed remembering

Experiment 1 introduced a constraint on directed remembering: its efficiency is limited by the number of objects over which maintenance operates. The effectiveness of memory maintenance also depends on the kinds of information that are available to it. Certain kinds of information, such as metamemory, can in principle be used to direct maintenance and to help people to remember more.

For example, prioritizing the representation that is currently weakest may promote remembering many weak representations. In contrast, prioritizing the representation that is currently strongest may promote remembering a small number of strong representations. Can participants use the relative strength of memories as a cue for what to prioritize, selectively maintaining the best- or worst-remembered object?

In this experiment, we extend the phenomenon of directed remembering beyond external cues. Specifically, we asked participants to engage in self-directed remembering — shifting the balance of memory maintenance according to an internally measured metamemory signal. This is accomplished by generalizing the directed remembering paradigm to include an internally generated direction, with a cue informing participants that at the end of the trial they will report the color of whichever object is remembered best. There are thus four conditions. In condition **1**, a cue appears at the end of the retention interval, telling the participant which object to report. This is the usual partial report paradigm, used in Chapter 2. In condition **2**, a cue appears at the end of the retention interval telling the participant to report the best-remembered object. This is a simplified version of the metamemory paradigm used in Chapter 1. In condition **3**, a cue appears early in the retention interval, again telling the participant which object to report the color of. This is the directed remembering paradigm. Finally, in condition **4**, a cue appears early in the retention interval telling the participant to report the best-remembered object. This is a test of self-directed remembering. The experiment is thus a 2×2 design: an early vs. late cue to a random vs. the best-remembered object.

3.2.1 Methods

Participants

Eight people between the ages of 18 and 25 participated. They all had normal or corrected-to-normal visual acuity, normal color vision, and were naive to the purpose of the experiment.

Stimuli

The stimulus was four colorful dots (radius 0.4° of visual angle), arranged in a circle centered on the display, with each object 4° from the center. At the center was a small black fixation mark. The color of each dot was drawn uniformly from a circle cut out of the CIE 1976 $L^*a^*b^*$ color space, centered at $L = 54$, $a = 18$, $b = -8$. Stimuli were rendered by MATLAB with the Psychophysics toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007), and presented on a 1920×1200 LCD screen at 60 Hz, 38 pixel/cm, positioned 60 cm from the participant.

Procedure

A schematic diagram of the procedure is found in Fig. 26 on pg. 70. Each trial began with a blank screen, followed by the presentation of the stimulus for 250 ms. There was then a 1000 ms retention interval with a blank screen. In conditions with an early cue, the cue then appeared for 250 ms. Otherwise the screen remained blank for the same amount of time. In all conditions there was then a 4000 ms retention interval. Then, in conditions with a late cue, the cue appeared for 250 ms. Otherwise the screen remained blank for the same amount of time. After a 250 ms blank period, on every trial the participant then reported the color of the appropriate object and its location.

Analysis

For each condition, we separately fit a hierarchical model to the data from all the participants. The model assumes that errors are distributed according to a circular normal distribution centered around 0 deg with a standard deviation that is normally distributed in the population. The model was fit to the data using MemToolbox 1.0.0 (Suchow et al., 2013).

3.2.2 Results

Fig. 27 compares performance on the memory task across the four conditions. Two comparisons are critical:

First, comparing early cues for the best vs. random condition replicates the basic metamemory result from Chapter 1 and a previous publication (Fougnie et al., 2012). Performance was better for the best-remembered object than for a randomly chosen object (early cue: $42 \pm 6°$ vs. $31 \pm 7°$, standard error, paired t-test, $t(7) = 5.8$, $p = 6.3 \times 10^{-4}$; late cue: $53 \pm 18°$ vs. $35 \pm 8°$, standard error, paired t-test, $t(7) = 3.9$, $p = 0.006$). This confirms that participants are successfully using metamemory to guide selection of structures from working memory.

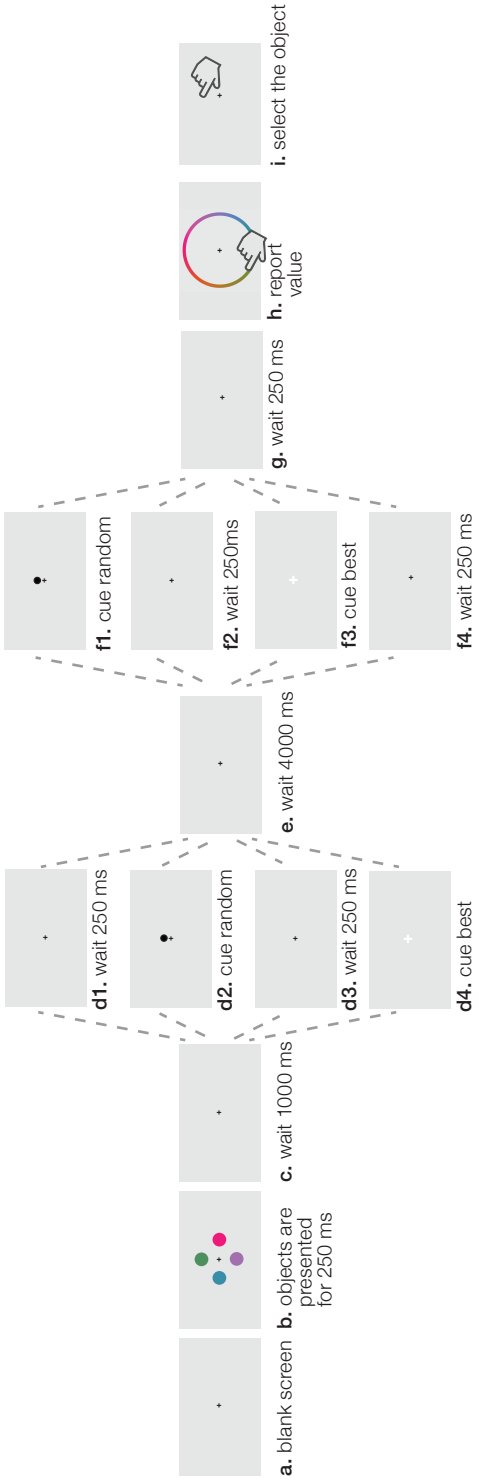
Second, comparing early vs. late cues for the best-remembered condition provides a test of self-directed remembering. We found that performance was better with an early cue (s.d. 30.9°) than with a late cue (s.d. 35.0°; paired t-test, $t(7) = -4.6$, $p = 0.002$; Fig. 27).

3.2.3 Interim discussion

Experiment 2 elaborated directed remembering to include internally-generated cues derived from metamemory — "self-directed remembering". We found that people were able to direct maintenance to the representation that was currently remembered best, preferentially maintaining it.

Figure 26 *(following page)*: Experimental design for a test of self-directed remembering in visual working memory. The screen starts out empty, with a small fixation dot in the center (**a**). In (**b–c**), four objects are briefly presented and then removed. Then a cue appears, either early or late in the retention interval (**d–f**). After a brief pause (**g**), the participant reports the color of the appropriate object (**h**) and then confirms the location of the reported object (**i**).

Figure 26: (continued)



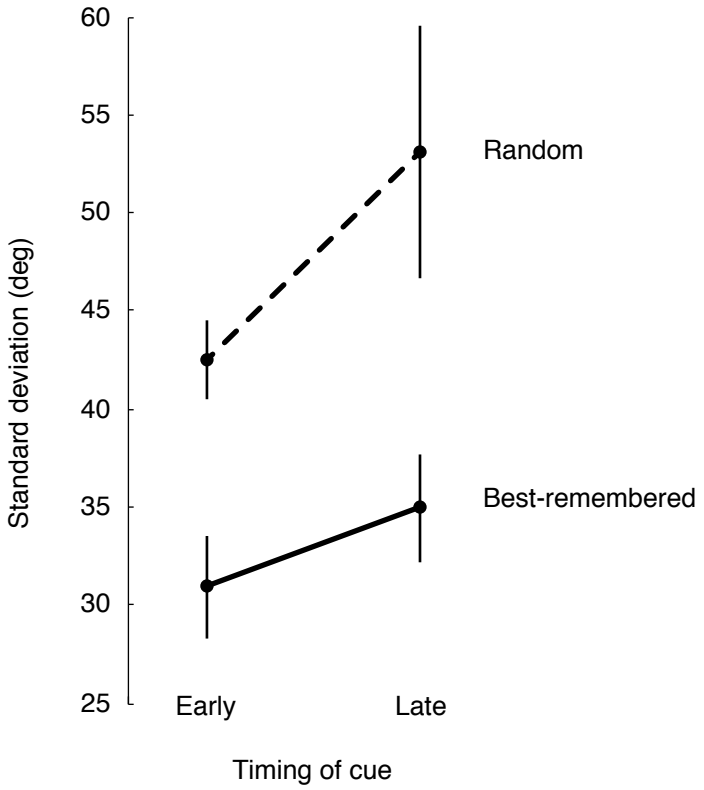


Figure 27: Performance differed when participants reported the best-remembered object versus one that was selected at random. Performance was better when the cue came early in the retention interval than when it came late.

The ability to control working memory in a way that depends not only on external cues but also on the current memory state opens up the possibility for new maintenance strategies that make use of metamemory to remember more. Two such strategies, mentioned above, at each moment prioritize representation that is currently weakest or strongest. Other strategies are possible too, e.g. giving graded preference to representations that are weakly represented, much like a conservationist might allocate resources to species on the brink of extinction.

The time course of directed and self-directed remembering remains a crucial open question, and the answer to it will determine the effectiveness of self-directed maintenance as a strategy for remembering more. Our participants made a single metamemory decision and then directed maintenance to the selected representation for the remainder of the trial. More sophisticated maintenance strategies might use multiple metamemory measurements, updating the target of maintenance to always reflect, for example, the strongest or weakest memory. The act of using metamemory, selecting a target of maintenance, and redirecting maintenance to the selected target all presumably take time. The longer it takes to redirect maintenance using metamemory, the less feasible such an approach will be.

3.3 Computational framework: Markov decision process

The goal of this work is to define the space of possible maintenance strategies, considering both the constraints and the sources of information that are available. Experiment 1 showed that the efficiency of directed remembering depends on the load, as would be predicted from a stability threshold that determines the weakest memory that can be maintained. Experiment 2 showed that maintenance can be directed by information from metamemory. Here, we extend the framework from Chapter 1 to accommodate flexibility in working memory — i.e., control over the contents of memory according to the demands of the task — as seen in both directed and self-directed remembering.

Specifically, to describe these flexible maintenance behaviors, we extend the model presented in Chapter 1 by drawing on tools from decision theory, likening the act of working memory maintenance to a sequential decision process in which, at each moment, the maintenance process decides which mental representations to prioritize. We focus on a particular kind of sequential decision process known as the Markov Decision Process, or MDP (Puterman, 1994), which provides an abstract mathematical framework for describing decision making in a setting that is partly under control of the decision maker (here, the control of maintenance) and partly under control of the environment (here, the degradation process). Besides being well-suited to describing the problem of memory maintenance, considering the Markov Decision Process has the added benefit of it being one of the most well-understood domains in the mathematics and psychology of reinforcement learning. Thus, once the connection has been established, a multitude of existing concepts and tools from reinforcement learning and decision theory can be brought to bear on the dynamics of memory maintenance.

An MDP is defined by a state space, a set of possible actions, a transition model, and a reward function, each explained in turn below.

3.3.1 State space

We suppose that there is a memory-supporting commodity divided into N quanta, each assigned to a particular memory structure. The more of the commodity assigned to a structure, the stronger and more robust the memory structure is. The state of memory is then an assignment of these quanta to structures. A structure may receive the entire commodity, only a portion of it, or perhaps none at all.

One way to visualize the state space S for such a system is to consider the various ways that N quanta can be distributed among K structures. [Mathematically, this corresponds to a $(K-1)$ regular discrete simplex — which, in the case of $K = 3$, consists of the points on the interior of a gridded

triangle whose vertices represent full allocation to a single memory structure, see Fig. 28. There is a point on the simplex that corresponds to the initial allocation.]

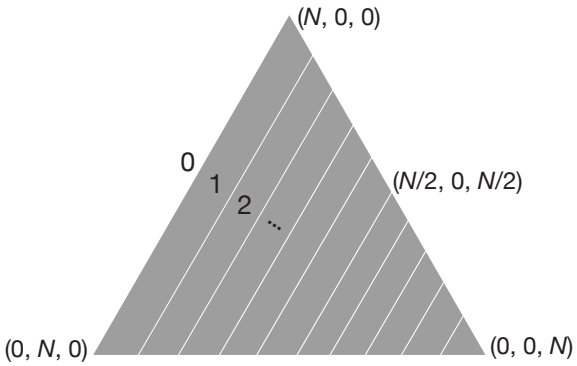


Figure 28: The 2-simplex. Each point on the simplex represents a specific allocation of the commodity. At the vertices, the entire commodity is allocated to a single memory structure. At the edges, one structure receives none of the commodity.

3.3.2 Set of possible actions

At each time step, we allow the maintenance process to act by selecting a quantum as the target of maintenance. Thus the set of possible actions A is of size N, one action per quantum, and does not depend on the state.

3.3.3 Transition model

The transition model specifies the probability of moving from one memory state to another. We will make use of the transition model proposed in Chapter 1 — a generalized Moran process with a stability threshold. This transition model assumes that at each time step, a quantum degrades because another quantum interferes with it or replaces it. If any memory structure is so weak that it has fewer than s quanta assigned to it, all those quanta become free floating and unassigned. [This defines a

transition model $\Pr(s' \mid s, a)$, which gives the probability of landing in state s' given that the agent took action a while in state s (Sutton & Barto, 1998; Ewens, 2004). It is a formal model of memory degradation.]

3.3.4 Reward function

Finally, there is the reward function. By definition, the agent's goal is to maximize the total reward that is received. The reward function is a mapping from states to an amount of reward that is received for landing in that state. In the case of a visual working memory task with a retention interval that is known to the participant, the reward function is time-varying, taking on a value of zero everywhere until the moment of the test, at which point it becomes positive for some states and (possibly) zero for others. The specifics of the reward function inevitably depend on the demands of the task and are usually implicit in the experiment's design and mechanism of feedback. For example, tasks using the continuous partial report paradigm require participants to hold information in mind for a fixed duration, e.g., 2000 ms, with reward provided in proportion to the similarity between the participant's response and the true value. Other tasks provide all-or-none feedback.

Three reward functions are commonly used in experiments on visual working memory. The first applies to tasks with an all-or-none design in which the participant receives full credit (e.g., +1) for having remembered anything at all about the cued object (i.e., having at least one quantum assigned to it at the time of the test) and otherwise receives no reward. This reward function is appropriate when scoring performance using a high-threshold model (Luck & Vogel, 1997; Cowan, 2001; Rouder et al., 2008, 2011), considering only the probability of remembering while ignoring accuracy. The second reward function applies to tasks in which the participant receives credit in proportion to the closeness of the reported value to the actual value (e.g., proportional to $180 - |x|$, where x is the

error in degrees in a circular dimension, such as hue), but with no penalty for blind guessing (e.g., by providing an opt-out button or by filtering out guesses using a statistical procedure). The third reward function applies to tasks in which both accuracy is important and guessing is penalized, modifying the previous reward function by scoring all the responses, even those that are blind guesses.

3.3.5 Policies: unconditional

The Markov decision process is a general framework for describing the problem of sequential decision making, but it does not specify the particular strategy used by the agent to make a decision. That strategy is defined by a policy, a function that specifies an action (or probability distribution over actions) for each possible state. Much of modern research on MDPs focuses on finding the optimal policy, one that maximizes the (possibly time-discounted) reward (Monahan, 1982; Puterman, 1994; Sutton & Barto, 1998; Kearns & Singh, 2002; Todorov, 2009).

The simplest maintenance policies do not depend on the current state of memory. Rather, they produce the same behavior in every state. Borrowing terminology from game theory, in which a player can adopt a strategy that does not depend on the behavior of the opponent (e.g., a player in the Prisoner's Dilemma who always defects), we call these maintenance policies *unconditional* (Rapoport, 1965; Axelrod & Hamilton, 1981). An example of an unconditional maintenance policy is `all-i`, which selects the *i*th quantum as the target of maintenance. Another example of an unconditional strategy is `random`, which selects a target at random, uniformly over all quanta. This maintenance policy is equivalent to the neutral process from Chapter 1, and thus reproduces the full range of experimental results from that chapter.

3.3.6 Policies: conditional

Conditional policies depend on the state. There are many conditional policies that are possible, but we will focus on three that are needed to produce the behavioral results presented in Exps. 1 and 2: `all-j` and `Luce`.

One example of a conditional policy is `all-j`, which selects a quantum uniformly from among those assigned to structure *j* if one exists, otherwise choosing randomly among all the quanta. We find that the `all-j` class of conditional policies is sufficient to produce directed remembering, with performance being better for the prioritized object than for the neglected object, like it was in Exp. 1. Specifically, we simulate the effects of a participant who uses `random` until the cue is revealed and then switches to `all-j`, where *j* is the memory structure that corresponds to the cued object (Fig. 29).

A particularly interesting class of policies is inspired by a psychological principle known as Luce's choice axiom (Luce, 1959; Herrnstein, 1961). According to the axiom, when faced with a choice among alternatives, a decision maker will exhibit 'matching behavior', selecting options with probability proportional to their value. Matching behavior was originally studied in the context of learning theory, where value is defined as the expected reward (Estes, 1957; Sutton & Barto, 1998). Thus if two levers offer rewards in a ratio of 2:1, an individual that displays matching behavior will press the more rewarding lever twice as often. Here, value is akin to memory strength and is defined by the number of quanta dedicated to a mnemonic structure.

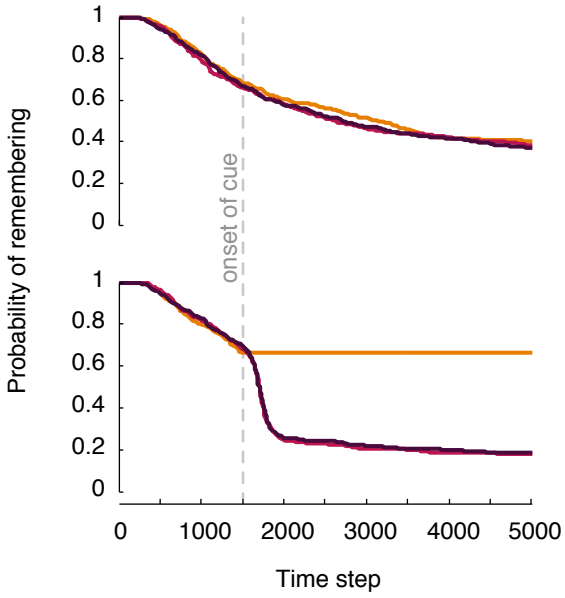


Figure 29: Reproducing directed remembering by simulating a conditional maintenance policy. Both panels show forgetting functions for each of three objects that were presented (purple, yellow, and red lines), averaged over 10,000 trials. The dashed vertical line marks the onset of the cue. In the upper panel, the participant uses the `random` policy, which gives equal priority to all three objects. In the lower panel, the participant uses the `all-j` policy, giving full priority to yellow and none to purple or red. Before the cue, the participants behave identically. After the cue, the behavior diverges, with yellow faring better than the other two. Simulations were run with parameters $N = 64$ and $K = 3$ for 5000 steps, matching the best-fit parameters from Chapter 1. The cue appeared at time step 1500.

In practice, it is common to consider a generalization of matching behavior in which a real-valued parameter L determines the decision-maker's sensitivity to the signal. In this so-called "softmax" generalization of matching behavior, the probability of selecting option a from the set of alternatives A is given by

$$p(a) = \frac{v(a)^L}{\sum_{b \in A} v(b)^L},$$

where $v(x)$ is the strength of the signal generated by x and where L determines the decision maker's sensitivity to the signal (Sutton & Barto, 1998; Vul, 2010). Other softmax generalizations of matching behavior are possible, many of which, like the Fermi function and Boltzmann distribution, are borrowed from the field of statistical physics and provide a formal link between selection in various domains (Ayala & Campbell, 1974; Blume, 1993; Sutton & Barto, 1998; Barabási & Albert, 1999; Pan, 2010; Traulsen et al., 2007).

Five values of L are particularly significant for the process of memory maintenance. When $L = 0$, the process is unconditional (i.e., insensitive to the signal). This corresponds to the neutral process developed in Chapter 2. When $L = 1$, the process gives preference to objects in proportion to how strongly they are currently represented. When $L \to \infty$, the winner takes all. In contrast, when $L = -1$, the process gives preference to objects in proportion to how *weakly* they are currently represented, and in the limit $L \to -\infty$, the loser takes all.

The `Luce` family of maintenance policies can reproduce the effects of self-directed remembering, observed in Exp. 2 (Fig. 30).

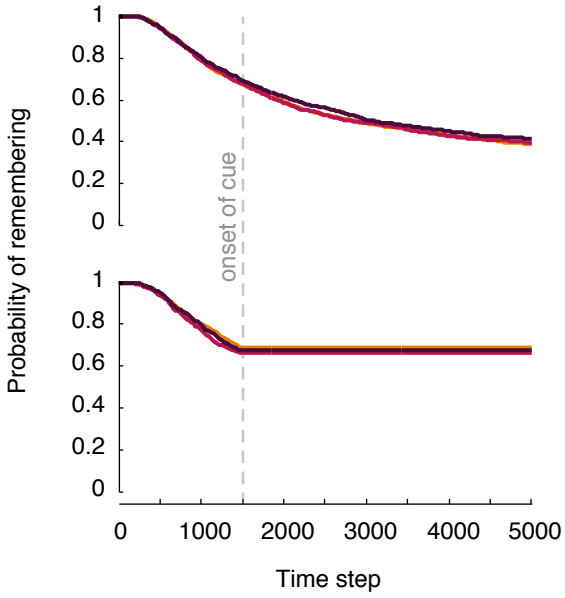


Figure 30: Reproducing self-directed remembering by simulating conditional maintenance policies. Once again, both panels show forgetting functions for each of three objects that were presented (purple, yellow, and red lines), averaged over 10,000 trials. The dashed vertical line again marks the onset of the cue. In the upper panel, the participant uses the `random` policy, which gives equal priority to all three objects. Here, in the lower panel, the participant uses the Luce policy with L=-1, assigning priority scores in proportion to how weakly the object is currently represented. Before the cue, the participants behave identically. After the cue, the behavior diverges, with the Luce policy faring better. Simulations were run with parameters $N = 64$, $K = 3$, and $L \in 0, -1$ for 5000 steps. The cue appeared at time step 1500.

3.3.7 Partially observable minds

The framework of a Markov decision process makes a strong commitment to the accessibility of the memory state to the memory maintenance system: it assumes perfect, real-time, no-cost metamemory. However, the experiment reported in Chapter 1 gives us reason to believe that metamemory may be imperfect. In that experiment, participants were more likely to remember an object when it was chosen as best-remembered on that trial ($92\% \pm 2$) than when it was selected at random from three objects ($71\% \pm 3$), even though the objects and displays were identical across the two conditions (Fig. 16). From these data we can estimate the efficiency of realtime metamemory. Specifically, if participants remembered a randomly cued object 71% of the time, then we can infer that the chance of remembering at least one of them (i.e., not having forgotten all three) is $1 - (1-0.71)^3 = 0.98$. When asked to pick out the best-remembered object, an ideal metamemorizer (one who performs optimally given the information that is available) would be able to retrieve an object 98% of the time. Our participants' inability to reach this level of performance suggests that at least some of them have less than ideal metamemory.

By generalizing the MDP to a partially observable world, we can accommodate situations of imperfect or costly metamemory. A partially observable world is one in which the agent does not know exactly what state it is in, making it impossible to carry out conditional policies that depend on the state without further information or assumptions. Often the agent has some instrument (a "sensor") for measuring or sensing the state. In the case of memory maintenance, the sensor is metamemory. The agent uses the sensor to update its beliefs about the state. Thus the POMDP extends the MDP through the introduction of a sensor model, which describes the information about the state that is provided by each observation, and a belief state, which is a probability distribution over the state space that embodies the agent's beliefs about the current state (Monahan, 1982; Kaelbling et al., 1996). One

possible belief state, a Dirichlet distribution with concentration parameters (4, 4, 4), is visualized here as a surface on the simplex (Fig. 31).

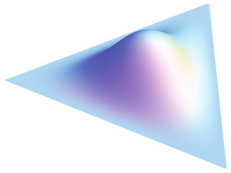


Figure 31: A Dirichlet distribution with concentration parameters 4, 4, and 4, representing one possible state of metamemory beliefs about resource allocation. This corresponds to a situation where the memorizer believes that the commodity is fairly evenly spread among memories 1, 2, and 3. The belief state is one component of metamemory, capturing what the memorizer believes about the current state of its memory. The sensor model is another component, capturing the information gained by metamemory observations.

In a partially observable mind, inefficiencies of metamemory limit the efficacy of flexible maintenance behaviors. This is because in a world where the future depends on the past, one who does not even know the present cannot suitably plan for what is to come. We demonstrate this dependence by defining a simple metamemory agent and then simulating its behavior with different levels of efficiency. Metamemory observations made by the agent come in the form of object labels sampled with probability proportional to their strength (that is, the number of quanta assigned to them). This defines the sensor model. The agent is initially unaware of the allocation of the commodity, represented by a belief state initially set to a Dirichlet distribution with concentration parameters 1, 1, and 1, which is equivalent to a uniform distribution over all possible allocations. At each time step, the agent makes m observations. We assume that the metamemory system has no memory of its own and thus considers only the observations made at the current time step (see below for a brief discussion of optimal filtering, in which the metamemory system also considers past observations). The Dirichlet distribution is convenient for describing states of uncertainty about multinomial data both because the Dirichlet is the conjugate prior for multinomial data and because its parameters act as pseudocounts, such that the a posteriori beliefs can be computed simply by incrementing the pseudocounts

by the number of observations made of each type. To avoid the problems caused by sampling zero quanta of a certain type, we use additive smoothing by adding one to all the counts. These counts are used by the Luce policy, with exponent 1. The efficiency of metamemory can be varied by altering the number of observations made at each time step. This formulation makes it possible to vary efficiency between two extremes. At one extreme, $m = 0$ and the agent gains no information about the state. At the other extreme, in the limit $m \to \infty$, the agent has perfect information about the state. Intermediate efficiencies lead to intermediate performance (Fig. 32).

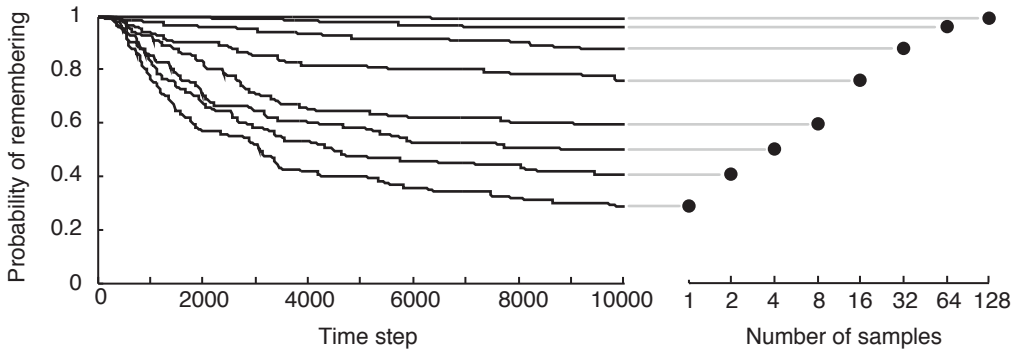


Figure 32: Inefficiencies of metamemory limit the efficiency of memory maintenance. On the left are forgetting functions for a simulated agent whose memory is only partially observable. At each time step the agent draws m quanta (with replacement) and observes their tags. Selection happens according to the procedure in the main text. On the right, performance increases with the number of samples taken. Simulations were run with settings $N = 128$, $K = 12$, and $L = -1$.

We have defined a simple agent that uses metamemory to control memory maintenance. More sophisticated approaches use the theory of optimal filtering. *Filtering* is the problem of estimating the current state of a time-varying system given noisy observations of its past and present (Anderson & Moore, 2012). Kalman filters and particle filters are two popular solutions to the filtering problem that take into account the system's dynamics and the statistical structure of the problem (Kalman et al., 1960; Van Der Merwe et al., 2000).

3.4 Discussion

This chapter began with two experiments on directed remembering. In the first experiment, we found that the efficiency of directed remembering depends on the memory load, with larger loads leading to less efficient directed remembering. In the second experiment, we found that participants were able to use metamemory to direct their maintenance behavior, which we call "self-directed remembering". We then extended the neutral model from Chapter 1 to describe flexible memory maintenance behaviors, including directed remembering and self-directed remembering. This was accomplished by likening memory maintenance to a sequential decision problem in which, at each moment, the agent decides which mental representations to prioritize. The strategy used by the agent to solve the problem of memory maintenance is defined by a policy. Different policies lead to different outcomes: the `all-i` family of policies lead to a neutral process; the `all-j` family allows for directed remembering; and the `Luce` family allows for self-directed remembering. Each policy was able to explain the pattern of results seen in experiments. We then considered memory maintenance in a partially-observable mind.

Accessibility

The theory of working memory has not traditionally included much of a role for the concept of accessibility because working memories are thought to be actively maintained, such that a retrieval stage seems unnecessary (Baddeley, 1992; Brady et al., 2013; Fougnie et al., 2014). However, there is a growing body of evidence for a role of accessibility and retrieval in visual working memory (Fougnie et al., 2014; Brady et al., 2013). For example, real-world objects can be stored in working memory with various degrees of visual detail, and Brady et al. (2013) found that the lowest possible quality in

working memory is similar to that found in long-term memory, suggestive of a common limit — a threshold of accessibility. The present work proposes a similar threshold of accessibility, though one that operates during maintenance, not retrieval. In Chapter 1 this threshold helped us to explain the crossover effect in the forgetting functions of visual memory, and here it helps us to explain why the efficiency of directed remembering depends on the load.

The maintenance problem

Sims et al. (2012) propose that the problem of visual working memory is that of transmitting visual information from one point in time to another. In this conceptualization of the problem of working memory, solutions take the form of encoders and decoders that maximize performance on a task. Sims et al. found that formulating the problem of working memory in this way does an excellent job of predicting behavioral performance and its dependence on the statistics of the visual input.

The present work views the problem of working memory differently because it takes seriously the idea that maintenance is an active and controllable process. In this light, the problem of visual working memory is that of encoding visual information, maintaining it over a short duration, and then later accessing it to perform a task well. The framework of the Markov decision problem makes it possible to carefully specify the structure of the task (the environment and goals) and its possible solutions.

Connections to other resource allocation tasks

Perhaps the biggest payoff that comes from connecting the problem of working memory maintenance to existing frameworks in mature fields such as decision theory and reinforcement learning is the set of new questions that it makes possible to ask.

For example, having delineated the set of possible maintenance policies, it becomes sensible to ask which policy is optimal. As mentioned earlier, a considerable body of work in the field of reinforcement learning explores methods for finding the optimal policy given the constraints of a particular task (Monahan, 1982; Puterman, 1994; Sutton & Barto, 1998; Kearns & Singh, 2002; Zilli & Hasselmo, 2008; Todorov, 2009). In the context of working memory, for example, policies that preferentially maintain strong memories result in fewer representations that each have greater strength; these policies are thus best suited for tasks that reward remembering much information about a few items. In contrast, policies that preferentially maintain weak memories result in a greater number of representations that are each weak; these policies are thus best suited for tasks that reward remembering a little about everything. Determining the optimal maintenance strategy for the reward structures commonly found in studies of visual working memory is a ripe area for future research.

One might also ask where maintenance policies come from. Specifically, how are they learned? Methods such as temporal difference learning have emerged as candidate learning mechanisms used in the brain to learn policies that guide behavior, and it has become popular to relate this particular class of learning algorithms to known reward circuitry in the brain (Kaelbling et al., 1996; Hollerman & Schultz, 1998; O'Doherty et al., 2003; O'Doherty, 2004; Maia, 2009; Schultz, 2010; Dayan & Niv, 2008). Particularly relevant is the work of Todd et al. (2008), who discuss methods for learning to use working memory by temporal difference methods. Specifically, they showed how temporal difference learning can be used to shape representations in the prefrontal cortex so that they are useful for working memory Todd et al. (2008). Also relevant is the work of (O'Reilly & Frank, 2006), who developed an "actor/critic" model of the neural substrates of working memory and cognitive control. They showed that an active gating mechanism that controls the contents of working memory can be learned through learning mechanisms from reinforcement learning (O'Reilly & Frank, 2006).

Finally, it may be useful to consider other resource allocation tasks that are similar in structure to that of memory maintenance — e.g., scheduling and queuing. Much of the original work on these problems came from the field of operations research, which originated from military planners in WWII and which today considers the optimal solutions to decision making and resource allocation tasks in a variety of settings, often in the context of organizational behavior (Graff, 1953; Taha, 2007) or electronic systems (Åström & Wittenmark, 2011; Silberschatz et al., 2013). Having made the link to these related problems, it may be fruitful to consider known solutions as candidate psychological mechanisms. For example, queuing theory is a set of tools for considering resource allocation tasks that feature the continuous arrival of entities that require the resource (e.g., callers to a company's customer support center) (Kleinrock, 1975). Most of the popular working memory tasks are episodic, with information arriving all at once and then being discarded at the end of the trial. Our visual experience is not always so episodic; rather, it is sometimes necessary to update the contents of working memory with new information or redirecting maintenance in light of new goals (Sandberg et al., 2000; Matthey et al., 2012). Looking towards queuing theory, for example, may provide insight into this problem of maintenance in the face of continuously-arriving information.

4

Conclusion

This dissertation has argued that the maintenance of visual memories is akin to breathing: it is a dynamic process with a default behavior that explains much of its usual workings, but which can be observed, overridden, and controlled.

Chapter 1 showed that the default behavior of maintenance is well described by an evolutionary process operating over the units of a memory-supporting cognitive commodity, like attention. In the chapter, we extended the classic depiction of the dynamics of visual memory maintenance to include competitive interactions between memories, fluid reallocation of a memory-supporting commodity,

and a stability threshold that determines the weakest memory that can still be maintained. The proposed model, based on these principles, can be understood as an evolutionary process with memories competing over a limited memory-supporting commodity. The model reproduced the time course of visual working memory observed in a series of experiments. Notable features of this time course included load-dependent stability and overreaching, in which the act of trying to remember more information causes people to forget faster, and to remember less, respectively. Our results demonstrated that evolutionary models provide quantitative insights into the mechanisms of memory maintenance.

Chapter 2 showed that memories can be observed through real-time metamemory, an inward-looking process that tracks the status of a memory as it degrades over time. Confidence in our memories is influenced by many factors, including beliefs about the perceptibility or memorability of certain kinds of objects and events, as well as knowledge about our skill sets, habits, and experiences. Notoriously, our knowledge and beliefs about memory can lead us astray, causing us to be overly confident in eyewitness testimony or to overestimate the frequency of recent experiences. We designed a task that strips away all these potentially misleading cues, requiring observers to make confidence judgments by directly assessing the quality of their memory representations. We show that individuals can monitor the status of a memory as it degrades over time. Our findings suggest that people have access to information reflecting the existence and quality of their memories, and furthermore, that they can use this information to guide their behavior.

Chapter 3 showed that the default maintenance behavior can be overridden and controlled by the processes of directed forgetting and self-directed remembering, which redirect maintenance in accordance with the goals. We found that the efficiency of directed forgetting depends on the memory load. This follows naturally from the proposed stability threshold. Specifically, when there are few representations, directed forgetting is efficient, shifting maintenance across objects without much

cost. However, when there are many representations, directed forgetting is inefficienct, such that shifting maintenance to prioritize one object comes at a big cost: fewer remembered objects in total. The chapter then combined the default behavior, metamemory, and directed forgetting into a formal framework borrowed from the field of reinformement learning. We extended the neutral model from Chapter 1 to describe flexible memory maintenance behaviors, including directed forgetting and self-directed remembering. This was accomplished by likening memory maintenance to a sequential decision problem in which, at each moment, the agent decides which mental representations to prioritize. The strategy used by the agent to solve the problem of memory maintenance is defined by a policy. Different policies lead to different outcomes: the `all-i` family of policies lead to a neutral process; the `all-j` family allows for directed forgetting; and the `Luce` family allows for self-directed remembering. And the efficiency of metamemory limits the efficacy of conditional maintenance policies in a partially observable mind.

The two dominant approaches to studying the processes that underly working memory — psychophysics and cognitive neuroscience — conceive of the problem of memory maintenance differently. Psychophysics studies how performance on a task changes as the spatiotemporal properties of the stimulus are adjusted (Fechner, 1860; Stevens, 1975). In this light, memory maintenance is defined by properties of the stimulus that affect it. Cognitive neuroscience studies the neural substrates of cognition and how those substrates together produce experience and behavior (Gazzaniga, 2004). In this light, maintenance is defined by its neural mechanisms.

This dissertation combined the techniques of psychophysics with a third approach — the computational approach — which delineates three levels of analysis that are needed to explain any cognitive phenomenon: the computational, the algorithmic, and the physical (Marr, 1982; Chater & Oaksford, 1998). Most relevant here are the computational and algorithmic levels, which describe the structure

of the task at hand, the logic of possible solutions, and the rules by which those solutions are carried out. Each chapter of the dissertation aimed at applying this kind of thinking to some aspect of our ability to maintain visual memories. Working memory is a flexible system, capable of many behaviors. The aspiration here is that by posing the problem of working memory in this way — as that of an agent who decides what to prioritize in a partially observable mind — it will become possible to discover the correspondingly rich and varied solutions taken to solve the maintenance problem.

A

Forgetting functions of visual memory

In the following derivations, we suppose that the participant is asked to remember a set of K things (the memory load), stored as objects, features, or hierarchical bundles of features (hereafter, "memory structures" or just "structures"). We further suppose that visual memory is limited and imperfect, such that only $Y \leq K$ of the structures are stored. The quantity Y is allowed to vary as a function of the time t since the offset of the stimulus. Then, for each model we can define a *forgetting function* that relates the expected number of stored structures to time. For each of the four models of visual mem-

ory compared in the main text, we derive expressions for its forgetting function.

A.0 Model #1: Classic

The classic model of the time course of visual memory, still used in modern applications (Lu et al., 2005; Kuhbandner et al., 2011; Hahn et al., 2011), emerged in the 1960s from research using the partial report paradigm (Sperling, 1960). That work revealed the existence of iconic memory, a storage system with a high capacity and whose contents is short-lived, typically fading within a second (Sperling, 1960). Under the classic model, working memory and iconic memory are together responsible for behavioral performance. The contribution of working memory is at most its full capacity β, which is unchanging over time. The contribution of iconic memory above and beyond that of working memory is often called the "partial report superiority effect" and is at most all of the remaining $K - \beta$ things that were not stored in working memory. The partial report superiority effect has been found to decline exponentially as a function of time, and so the forgetting function of the classic model is given by

$$\mathrm{E}[Y(t)] = \begin{cases} \beta + (K - \beta)e^{\frac{-t}{\tau}} & \text{if } \beta \leq K \\ K & \text{if } \beta > K, \end{cases} \tag{A.1}$$

where τ is the mean lifetime of an item held in iconic memory.

A.1 Model #2: Pure death

The previous model assumed that working memory is stable over time. But working memory is known to degrade (Zhang & Luck, 2009; Yang, 1999). For simplicity, we assume that degradation

in working memory is a pure death process in which structures are lost independently over time and independently of each other, each having a mean lifetime of τ_2. First consider the case of $\beta \leq K$, where working memory is exhausted. In this case, the probability that a randomly chosen structure is stored in working memory is $\frac{\beta}{K} e^{\frac{-t}{\tau_2}}$. The probability that it is stored in iconic memory is $e^{\frac{-t}{\tau}}$. Thus the forgetting function, which tracks the expected number of objects held in at least one of the two systems, is given by

$$\mathrm{E}[Y(t)] = K - K\left(1 - \frac{\beta}{K} e^{-t/\tau_2}\right)\left(1 - e^{-t/\tau}\right). \tag{A.2}$$

In the case of $\beta > K$, where working memory has room to spare, the term $\frac{\beta}{K}$ is replaced by unity because every structure is guaranteed a place. In the limit $\tau_2 \to \infty$, the pure death model reduces to the classic model.

A.2 Model #3: Sudden death

In 2009, Zhang & Luck proposed a "sudden death" model where after a window of initial stability lasting approximately four seconds, entire objects are lost over time, but the quality of those that survive is constant (Zhang & Luck, 2009). A reasonable way to formalize this model is to equip the pure death process with an initial grace period that lasts until time t_{death}. The forgetting function for this sudden death model is then governed by the classic model when $t < t_{\text{death}}$ and by the pure death model when $t \geq t_{\text{death}}$. Note that, because of the initial grace period, when used in the sudden death model, the term $\frac{t}{\tau_2}$ in Eq. 2.2 must be replaced by $\frac{t - t_{\text{death}}}{\tau_2}$.

A.3 Model #4: Evolutionary model

Here, we derive the forgetting function of the evolutionary model with the stability threshold set to $s = 1$, i.e., the Moran process (Moran, 1958, 1962). For $s \geq 1$, we determined the forgetting functions numerically.

We consider N quanta, each of which is assigned to one of the K structures at any given time. We suppose that the structures stored in these quanta undergo a process of neutral drift, modeled as a continuous-time Moran or pairwise comparison process. It is convenient to scale time so that one time unit corresponds to N "generations" of this process, so that the contents of each quantum is updated once per unit time, on average.

Decay of founding lineages

When stimuli are first presented to a subject, each quantum is immediately assigned a single structure. We consider this to be the "founding generation" of structures stored in memory. At any subsequent time, the contents of each quantum will be a copy (or a copy-of-a-copy, etc.) of a member of this founding generation. Over time, the lineages (copies and copies-of-copies, etc.) of this founding generation may grow or disappear through random drift. Eventually only one lineage will remain.

We first ask how many lineages from the founding generation will survive to time $t > 0$. This question can be addressed using results from population genetics. We represent the number of founding lineages that persist at time t by the random variable $X(t)$. The expectation of this random variable is (Tavaré, 1984):

$$\mathrm{E}[X(t)] = 1 + \sum_{\ell=2}^{M} (2\ell - 1) \frac{\binom{M}{\ell}}{\binom{M+\ell-1}{\ell}} e^{-\binom{\ell}{2} t}.$$

The forgetting function

We now consider the forgetting function—that is, the expected number of distinct structures that survive in memory at a given time. We suppose that, at time $t = 0$, each quantum is assigned randomly to one of K structures. We represent the the number of structures remaining at time $t \geq 0$ by the random variable $Y(t)$. The expected number of structures remembered at time t can be written as

$$\mathrm{E}[Y(t)] = 1 + \sum_{\ell=2}^{N} C_\ell^{N,K} e^{-\binom{\ell}{2} t}, \tag{A.3}$$

with the coefficients $C_\ell^{N,K}$ given by

$$C_\ell^{N,K} = (-1)^\ell (2\ell - 1) \frac{\binom{N}{\ell}}{\binom{N+\ell-1}{\ell}} \frac{K-1}{K} \, {}_2F_1\left(\ell + 1, 2 - \ell, 2, \frac{K-1}{K}\right). \tag{A.4}$$

Above, ${}_2F_1$ is the hypergeometric function. The derivation of the forgetting function is given in the next two sections.

In the limit $N \to \infty$ (that is, if memory is regarded as a continuous resource) the forgetting function converges to

$$\mathrm{E}[Y(t)] = 1 + \sum_{\ell=2}^{\infty} C_\ell^{K} e^{-\binom{\ell}{2} t},$$

with

$$C_\ell^{K} = (-1)^\ell (2\ell - 1) \frac{K-1}{K} \, {}_2F_1\left(\ell + 1, 2 - \ell, 2, \frac{K-1}{K}\right).$$

Trinomial coefficients

Our derivation of the forgetting function Eq. 2.3 relies on identities involving trinomial coefficients. For nonnegative integers M, i, j with $i + j \leq M$, the corresponding trinomial coefficient is defined as

$$\binom{M}{i \quad j \quad M-i-j} = \frac{K!}{i!j!(K-i-j)!}.$$

Trinomial coefficients arise as coefficients in the expansion of $(x + y + z)^M$. In particular, we have

$$(-x + y + 1)^M = \sum_{\substack{i+j=M \\ i\geq 0, j\geq 0}} (-1)^i \binom{M}{i \quad j \quad M-i-j} x^i y^j.$$

From the above expansion, we can derive the following relations:

$$\begin{aligned}
\sum_{i=0}^{M-j} (-1)^i \binom{M}{i \quad j \quad M-i-j} &= \frac{1}{j!} \frac{\partial^j}{\partial y^j} (-x + y + 1)^M \Big|_{(x,y)=(1,0)} \\
&= \binom{M}{j} (-x + y + 1)^{M-j} \Big|_{(x,y)=(1,0)} \\
&= \begin{cases} 1 & j = M \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{A.5}$$

$$\begin{aligned}
\sum_{i=0}^{M-j}(-1)^i i\binom{M}{i \quad j \quad M-i-j} &= \frac{1}{j!}\frac{\partial}{\partial x}\frac{\partial^j}{\partial y^j}(-x+y+z)^M\Big|_{(x,y,z)=(1,0,1)} \\
&= \binom{M}{j}\frac{\partial}{\partial x}(-x+y+z)^{M-j}\Big|_{(x,y,z)=(1,1,0)} \\
&= -\binom{M}{j}(M-j)(-x+y+z)^{M-j-1}\Big|_{(x,y,z)=(1,1,0)} \\
&= \begin{cases} -M & j = M-1 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{A.6}$$

Combining the previous two identities yields a third:

$$\begin{aligned}
\sum_{k=j}^{M}(-1)^{k-j}k\binom{M}{M-k \quad k-j \quad j} &= \sum_{i=0}^{M-j}(-1)^i(i+j)\binom{M}{M-i-j \quad i \quad j} \\
&= \sum_{i=0}^{M-j}(-1)^i i\binom{M}{i \quad j \quad M-i-j} \\
&\quad + j\sum_{i=0}^{M-j}(-1)^i\binom{M}{i \quad j \quad M-i-j} \\
&= \begin{cases} -M & j = M-1 \\ M & j = M \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{A.7}$$

Derivation of the forgetting function

We now derive the forgetting function for the evolutionary model. First we suppose that n of the N founding lineages remain after time t; that is, $X(t) = n$. Since neutral drift does not favor any structure over any other, we can regard these n lineages as being assigned randomly among the K structures. This situation thus reduces to the classical probability problem of randomly partitioning a set of n elements into K or fewer subsets.

For $k \leq n$, the probability that k of the K items are represented in these n lineages is

$$\Pr[Y(t) = k | X(t) = n] = \frac{\binom{K}{k} \left\{ n \atop k \right\} k!}{K^n}. \tag{A.8}$$

Above, $\left\{ n \atop k \right\}$ denotes the (n, k)th Stirling number of the second kind—that is, the number of ways to partition a set of n elements into k non-empty subsets. This Stirling number can be obtained by the formula

$$\left\{ n \atop k \right\} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n. \tag{A.9}$$

Combining these equations yields

$$\Pr[Y(t) = k | X(t) = n] = \binom{K}{k} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \left(\frac{j}{K} \right)^n,$$

or equivalently, upon rearranging,

$$\Pr[Y(t) = k | X(t) = n] = \sum_{j=0}^{k} (-1)^{k-j} \binom{K}{K-k \quad k-j \quad j} \left(\frac{j}{K}\right)^n. \tag{A.10}$$

The trinomial coefficient arises via the relation

$$\binom{K}{k}\binom{k}{j} = \binom{K}{K-k \quad k-j \quad j}.$$

Now we consider the overall expected number of items remembered at time t by summing Eq. 2.10 over values of n weighted by their probabilities:

$$\begin{aligned} \mathrm{E}[Y(t)] &= \sum_{k=1}^{K} k \sum_{j=0}^{k} (-1)^{k-j} \binom{K}{K-k \quad k-j \quad j} \sum_{n=1}^{N} \left(\frac{j}{K}\right)^n \Pr[X(t) = n] \\ &= \sum_{k=1}^{K} k \sum_{j=0}^{k} (-1)^{k-j} \binom{K}{K-k \quad k-j \quad j} G(j/K; t), \end{aligned} \tag{A.11}$$

Above, $G(x; t)$ is the probability generating function of $X(t)$:

$$G(x; t) = \sum_{n=1}^{N} x^n \Pr[X(t) = n].$$

We use a previously discovered (Tavaré, 1984) formula for this generating function:

$$G(x; t) = x + x(1-x) \sum_{\ell=2}^{N} (2\ell - 1)(-1)^{\ell+1} \frac{\binom{N}{\ell}}{\binom{N+\ell-1}{\ell}} \, {}_2F_1(\ell + 1, 2 - \ell, 2, x) \, e^{-\binom{\ell}{2} t}. \tag{A.12}$$

Substituting in Eq. 2.11, we obtain

$$\mathrm{E}[Y(t)] = \sum_{k=1}^{K} k \sum_{j=0}^{k} (-1)^{k-j} \binom{K}{K-k \quad k-j \quad j} \times \left[\frac{j}{K} + \frac{j}{K}\left(1 - \frac{j}{K}\right) \sum_{\ell=2}^{N} (2\ell - 1)(-1)^{\ell+1} \frac{\binom{N}{\ell}}{\binom{N+\ell-1}{\ell}} \, {}_2F_1(\ell + 1, 2 - \ell, 2, j/K) \, e^{-\binom{\ell}{2}t} \right] \quad \text{(A.13)}$$

Using identity 2.6, we can simplify the term that is linear in j/K:

$$\sum_{k=1}^{K} k \sum_{j=0}^{k} \frac{j}{K} (-1)^{k-j} \binom{K}{K-k \quad k-j \quad j} = 1.$$

Eq. 2.13 therefore reduces to

$$\mathrm{E}[Y(t)] = 1 + \sum_{k=1}^{K} k \sum_{j=0}^{k} (-1)^{k-j} \binom{K}{K-k \quad k-j \quad j} \times \frac{j}{K}\left(1 - \frac{j}{K}\right) \sum_{\ell=2}^{N} (2\ell - 1)(-1)^{\ell+1} \frac{\binom{N}{\ell}}{\binom{N+\ell-1}{\ell}} \, {}_2F_1(\ell + 1, 2 - \ell, 2, j/K) \, e^{-\binom{\ell}{2}t}.$$

In summary, the expected number of items remembered can be written as

$$\mathrm{E}[Y(t)] = 1 + \sum_{\ell=2}^{M} C_{\ell}^{N,K} e^{-\binom{\ell}{2}t},$$

with

$$C_\ell^{N,K} = (-1)^{\ell+1}(2\ell-1)\frac{\binom{N}{\ell}}{\binom{N+\ell-1}{\ell}} \times \sum_{k=1}^{K} k \sum_{j=0}^{k} (-1)^{k-j} \binom{K}{K-k \quad k-j \quad j} \frac{j}{K}\left(1-\frac{j}{K}\right) \times {}_2F_1(\ell+1, 2-\ell, 2, j/K). \quad \text{(A.14)}$$

To simplify this expression for $C_\ell^{N,K}$ we reorder sums:

$$\begin{aligned} &\sum_{k=1}^{K} k \sum_{j=0}^{k} (-1)^{k-j} \binom{K}{K-k \quad k-j \quad j} \frac{j}{K}\left(1-\frac{j}{K}\right) {}_2F_1(\ell+1, 2-\ell, 2, j/K) \\ =&\sum_{j=0}^{K} \frac{j}{K}\left(1-\frac{j}{K}\right) {}_2F_1(\ell+1, 2-\ell, 2, j/K) \sum_{k=j}^{K} (-1)^{k-j} k \binom{K}{K-k \quad k-j \quad j}. \end{aligned} \quad \text{(A.15)}$$

Simplifying the second (nested) sum according to identity Eq. 2.7, we obtain Eq. 2.4.

B

Data analysis with the MemToolbox

B.1 Abstract

The MemToolbox is a collection of MATLAB functions for modeling visual working memory. In support of its goal to provide a full suite of data analysis tools, the toolbox includes implementations of popular models of visual working memory, real and simulated data sets, Bayesian and maximum likelihood estimation procedures for fitting models to data, visualizations of data and fit, validation routines, model comparison metrics, and experiment scripts. The MemToolbox is released under the

permissive BSD license and is available at memtoolbox.org.

B.2 Introduction

Working memory is a storage system that actively holds information in mind and allows for its manipulation, providing a workspace for thought (Baddeley, 1986). Its strikingly limited capacity has inspired a slew of research aimed at characterizing those limits in terms of the spatiotemporal properties of the stimulus and the age, intelligence, tiredness, and mental health of the individual.

A handful of experimental paradigms are predominant in the study of working memory. These include the delayed match-to-sample task used in studies of animal cognition (Blough, 1959) and the span tasks used in studies of verbal working memory (Daneman & Carpenter, 1980). Research on visual working memory relies primarily on two tasks: partial report and change detection (Fig. 33). In a partial report task, the participant is shown a set of letters, shapes, or colorful dots, and then after a brief delay is asked to report the properties of one or a subset of the items (Sperling, 1960; Wilken & Ma, 2004; Zhang & Luck, 2008). In a change detection task, the participant is shown a pair of displays, one after the other, and is asked a question that requires comparing them, such as whether they match (Luck & Vogel, 1997; Phillips, 1974; Pashler, 1988).

Formal models have been proposed that link performance in change detection and partial report tasks to the architecture and capacity of the working memory system. These include the item limit model (Pashler, 1988), the slot model (Luck & Vogel, 1997; Cowan, 2001), the slots+averaging model (Zhang & Luck, 2008), the slots+resources model (Awh et al., 2007), the continuous resource model (Wilken & Ma, 2004), the resources+swaps model (Bays et al., 2009), the ensemble statistics+items model (Brady & Alvarez, 2011), and the variable-precision model (van den Berg et al., 2012; Fougnie et al., 2012). Each model specifies the structure of visual memory and the decision

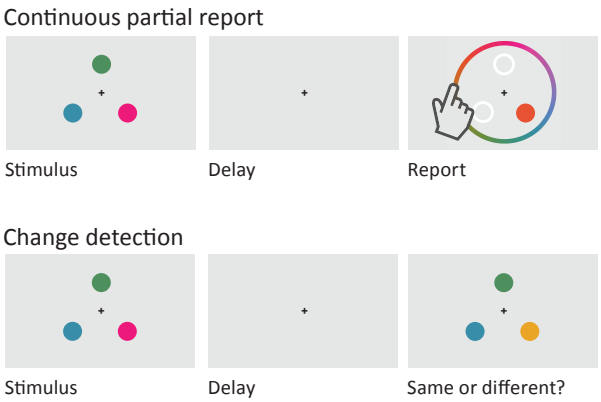


Figure 33: (a) A continuous partial report task. The observer sees the stimulus display, and then after a delay is asked to report the exact color of a single item. (b) A change detection task. The observer sees the stimulus display, then after a delay is asked to report whether the test display matches.

process used to perform the task.

Having been fit to the data, these models are used to make claims about the architecture and capacity of memory. For example, using an item limit model to fit data from a change detection task, Luck & Vogel (1997) showed that observers can remember the same number of objects regardless of whether they store one or two features per object (e.g., only color vs. both color and orientation), and from this inferred that the storage format of visual working memory is integrated objects, not individual features. Using data from a continuous partial report task, Brady & Alvarez (2011) showed that when items are presented in a group, memory for an individual item is biased towards the group average, and from this inferred that working memory is hierarchical, representing both ensembles of items and individual items.

B.3 The MemToolbox

We created the MemToolbox, a collection of MATLAB functions for modeling visual working memory. The toolbox provides everything needed to perform the analyses commonly used in studies of

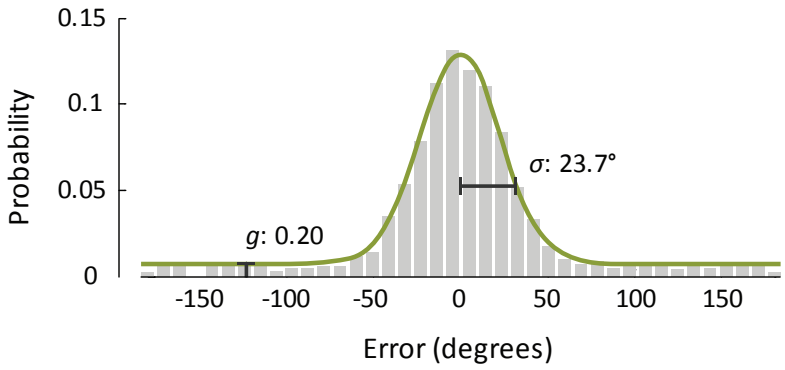


Figure 34: An example of a model's fit to continuous partial report data. Most responses fall within a relatively small range around the target item's true value, but some response are far off. These data are fit using the mixture model of Zhang & Luck (2008), which attributes the gap in performance to differences in the state of memory for the target item: with probability *g* the observer remembers nothing about the item and guesses randomly; with probability 1–*g* the observer has a noisy representation of the item, leading to responses centered at the true value and with a standard deviation *sd* that reflects the quality of the observer's memory.

visual working memory, including model implementations, maximum likelihood routines, and data validation checks, although it defaults to a Bayesian workflow that encourages a deeper look into the data and the models' fits. In the following sections, we highlight the toolbox's core functionality and describe the improvements it offers to the standard workflow. We begin by reviewing the standard workflow and its implementation in the toolbox.

B.3.1 The standard workflow

The experimenter first picks a model. Then, to fit the model to experimental data using probabilistic methods, a likelihood function is defined that describes the model's predictions for each possible setting of its parameters. (Formally, given a model M with free parameters θ, the model's likelihood function specifies a probability distribution $P(D|\theta)$ over possible datasets D.) With the likelihood function in hand, an estimator is used to pick the parameter settings that provide the best fit to the data. A popular choice is the maximum likelihood estimator, which selects the parameter values that

maximize the model's likelihood function given the data (Dempster et al., 1977; Lagarias et al., 1998). Typically, this procedure is performed separately for each participant and experimental condition, resulting in parameter estimates that are then compared using traditional statistical tests (e.g., Zhang & Luck, 2008).

The MemToolbox uses two MATLAB structures ("structs") to organize the information needed to analyze data using the standard workflow: one that describes the data to be fit, and another that describes the model and its likelihood function.

Fitting a set of data with a model is then as simple as calling the built-in `MLE()` function. For example, if data was obtained from a continuous color report task where an observer made errors of -89 degrees, 29 degrees, etc, a model could be fit using the following workflow:

```
>> model = StandardMixtureModel();
>> data.errors = [-89,29,-2,6,-16,65,43,-12,10,0,178,-42];
>> fit = MLE(data, model)
```

This will return the maximum likelihood parameters for this observer's data, allowing for standard analysis techniques to be used.

Thus, with little effort, the MemToolbox can be used to simplify (and speed-up) existing workflows by allowing for straightforward fitting of nearly all of the standard models used in the visual working memory literature. In support of this goal, the toolbox includes descriptive models such as the `StandardMixtureModel` (that of Zhang & Luck, 2008) and `SwapModel` of Bays et al. (2009), as well as several explanatory models, such as `VariablePrecisionModel` (e.g., van den Berg et al., 2012; Fougnie et al., 2012). For more information about a particular model `m`, type `help m` at the MATLAB prompt. For example, to access the help file for `StandardMixtureModel`, run `help StandardMixtureModel`. It is also possible to view the full code for a model by running `edit m`. (In fact, this applies to any function in the toolbox.) For example, to view the code for the

swap model, type `edit SwapModel`, which will show the model's probability distribution function, the parameters' ranges, and the specification of priors for the model's parameters.

The toolbox also includes a number of wrapper functions that extend existing models and make them more robust. For example, the wrapper function `WithBias()` adds a bias term, `WithLapses()` adds an inattention parameter, and `Orientation()` alters a model so that it uses a 180 deg error space, appropriate for objects that are rotationally symmetric, e.g., line segments. Inattention parameters are particularly important because deciding whether to include such parameters has an inordinate influence on parameter estimation and model selection (Rouder et al., 2008). Many of the standard models in the toolbox (e.g., `StandardMixtureModel`) already include such inattention parameters.

B.3.2 The Bayesian workflow

By default, instead of returning the maximum likelihood estimate, the toolbox uses a Bayesian workflow that constructs a full probability distribution over parameter values. This probability distribution describes the reasonableness of each possible parameter setting after considering the observed data, in light of prior beliefs. In doing so, it strongly encourages a thorough examination of model fit. The Bayesian workflow is implemented as `MemFit`, the toolbox's primary fitting function.

```
>> fit = MemFit(data, model)
```

Bayesian inference provides a rational rule for updating prior beliefs ("the prior") based on experimental data. The prior, $P(\theta)$, conveys which parameter values are thought to be reasonable, and specifying it can be as straightforward as setting upper and lower bounds (for example, bounding the guess rate between 0 and 1). Analysts add value through judicious selection of priors that faithfully reflect their beliefs. Because a prior can have arbitrarily large impact on the resulting inference, it is important both to carefully consider which distribution is appropriate, and, when communicat-

ing results that depend on those inferences, to report exactly the choice that was made. For the purposes of exploratory data analysis, it is common to use a noninformative or weakly informative prior that spreads the probability thinly over a swath of plausible parameter values (e.g., the Jeffreys prior, a class of noninformative priors that are invariant under reparameterization of the model; Jeffreys, 1946; Jaynes, 1968) to avoid an inordinate influence of the prior on inferences.

Once specified, beliefs are then updated (according to Bayes' rule) to take into account the experimental data. Bayes' rule stipulates that after observing data D, the posterior beliefs about the parameters ("the posterior") are given by

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta),$$

which combines the likelihood of the data given the parameters with the prior probability of the parameters.

Estimating the full posterior distribution is harder than finding the maximum likelihood estimate. For some models it is possible to derive closed-form expressions for the posterior distribution, but for most models this is intractable and so sampling-based algorithms are used to approximate it. One such algorithm, the Metropolis-Hastings variant of Markov Chain Monte Carlo (MCMC), is applicable to a wide range of models and is thus the one used by the toolbox (Metropolis et al., 1953; Hastings, 1970). The algorithm chooses an initial set of model parameters, and then, over many iterations, proposes small moves to these parameter values, accepting or rejecting them based on how probable the new parameter values are in both the prior and the likelihood function. In this way, it constructs a random walk that visits parameter settings with frequency proportional to their probability under the posterior. This allows the estimation of the full posterior of the model in a reasonable amount of time, and is theoretically equivalent to the more straightforward (but much slower) technique of evaluating the model's likelihood and prior at every possible setting of the parameters (implemented in

the `GridSearch` function). For an introduction to MCMC, we recommend Andrieu et al. (2003). The MemToolbox includes an implementation of MCMC that attempts, as best as possible, to automate the process of sampling from the posterior of the models that are included in the toolbox.

With the posterior distribution in hand, there are a number of ways to analyze and visualize the results. First, the maximum of the posterior distribution (the *maximum a posteriori* or MAP estimate) can be used as a point-estimate that is analogous to the maximum likelihood estimate, differing only in the MAP's consideration of the prior (This estimate can be calculated directly using the MAP function). However, visualizing the full posterior distribution also provides information about how well the data constrain the parameter values and whether there are trade-offs between parameters (Fig. 35).

Fig. 35 shows a posterior distribution for data analyzed with the standard mixture model of Zhang & Luck (2008). The plots on the main diagonal are histograms of values for each parameter, the so-called "marginals" of the posterior distribution; the plots on the off-diagonals reveal correlations between parameters. Note that in the standard mixture model, there is a slight negative correlation between the standard deviation parameter and the guess rate parameter: data can be seen as having either a slightly higher guess rate and lower standard deviation, or a slightly lower guess rate and higher standard deviation. Examining the full posterior reveals this tradeoff, which remains hidden when using maximum likelihood fits. This is important when drawing conclusions that depend on how these two parameters relate. For example, Anderson et al. (2011) found correlations between a measure based on guess rate and another based on standard deviation, and used this to argue that each observer has a fixed personal number of discrete memory slots. However, because the parameters trade off, such correlations are meaningful only if the estimates are derived from independent sets of data; otherwise the correlations are inflated by the noise in estimating the parameters (Brady et al., 2011a). Thus, understanding the full posterior distribution is critical to correctly estimating parameters and their relationships to each other.

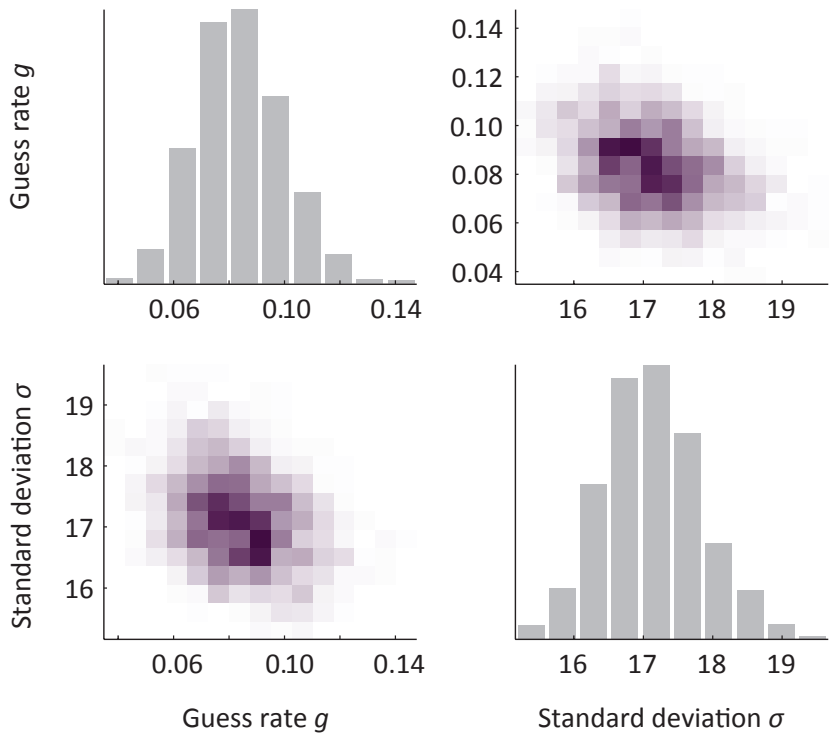


Figure 35: An example of the full posterior of the standard mixture model of Zhang & Luck (2008), where g is the guess rate and sd is the standard deviation of observers' report for remembered items. On the diagonal are plots that show the posterior for an individual parameter – e.g., the distribution for guess rate (g) is plotted in the top left corner. We can see that the data make us quite confident that the guess rate is between 0.08 and 0.11. On the off-diagonals are the correlations between the parameters – for example, the top right axis shows guess rate (y-axis) plotted against standard deviation (x-axis). Each row and each column corresponds to a parameter, e.g., the x-axis for all the plots in the second column corresponds to standard deviation.

B.3.3 Posterior predictive checks

Another technique applied by the MemToolbox is the automatic use of posterior predictive checks. Sometimes a whole class of models performs poorly, such that there are no parameter settings that will produce a good fit. In this case, maximum likelihood and maximum a posteriori estimates are misleading: they dutifully pick out the best, even if the best is still quite bad. A good practice then is to check the quality of the fit, examining which aspects of the data it captures and which aspects it misses (Gelman et al., 1996, 2004). This can be accomplished through a posterior predictive check,

which simulates new data from the posterior fit of the model, and then compares the histograms of the actual and simulated data (Fig. 36). `MemFit` performs posterior predictive checks by default.

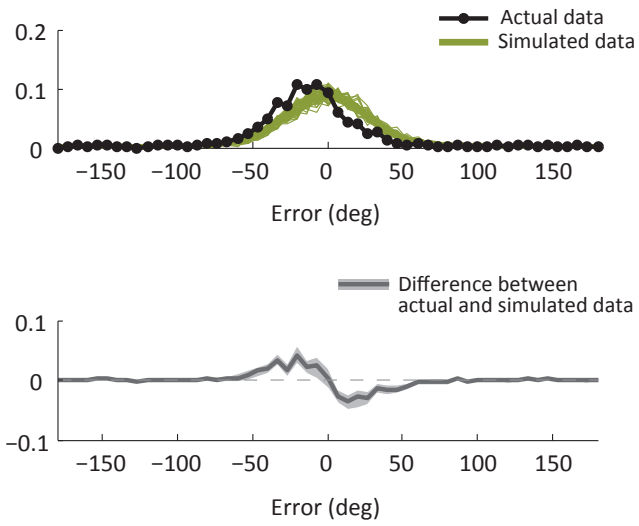


Figure 36: (a) Simulated data from the posterior of the model (green), with the actual data overlaid (black). The mismatch between the two is symptomatic of a poor fit. (b) The difference between the simulated and real data, bounded by 95% credible intervals. If at any spot the credible interval does not include zero, it is an indication that the model does not accurately fit the data.

A model that can accurately fit the data in a posterior predictive check does not necessarily provide a good fit. For example, the model may fit the averaged data but fail to fit observers' data from individual displays, perhaps because of reports of incorrect items (Bays et al., 2009) or because of the use of grouping or ensemble statistics (Brady & Tenenbaum, 2013). In addition, a good fit does not necessarily indicate a good model: an extremely flexible model that can mimic any data always provides a good fit, but this provides no evidence in favor of that model (Roberts & Pashler, 2000). However, models that systematically deviate in a posterior predictive check nearly always need improvement.

B.3.4 Hierarchical modeling

Typically, the question of interest in working memory research is not about a single observer, but a population: Do older individuals have reduced working memory capacity? Do people guess more often when there is more to remember? When aggregating results from multiple participants to answer such questions, the standard technique is to separately fit a model to the data from each participant (using, for example, maximum likelihood estimation) and to combine parameter estimates across participants by taking their average or median. Differences between conditions are then examined through *t*-tests or ANOVAs. This approach to analyzing data from multiple participants allows generalization to the population as a whole, since participant variance is taken into account by treating parameters as random effects (Daw, 2011). One flaw with this approach is that it entirely discards information about the reliability of each participant's parameter estimates. This is particularly problematic when there are differences in how well the data constrain the parameters of each participant (e.g., because of differences in the number of completed trials), or when there are significant trade-offs between parameters (as in the parameters of the standard model), in which case analyzing them separately can be problematic (Brady et al., 2011a). For example, the standard deviation parameter of the standard mixture model is considerably less constrained at high guess rates than at low guess rates. Thus, even with the same number of trials, our estimate of the standard deviation will be more reliable for participants with lower guess rates than those with higher guess rates.

A better technique, although one that is more computationally intensive, is to fit a single hierarchical model of all participants (e.g., Morey, 2011; Rouder et al., 2003; Rouder & Lu, 2005). This treats each participant's parameters as samples from a normally distributed population and uses the data to infer the population mean and SD of each parameter. This technique automatically gives more weight to participants whose data give more reliable parameters estimates and causes "shrinkage" of each participant's parameter estimates towards the population mean, sensibly avoiding extreme val-

ues caused by noisy data. For example, using maximum likelihood estimates, participants with high guess rates are sometimes estimated to have guess rates near zero but standard deviations of 3000 deg (resulting in a nearly flat normal distribution). This problem is avoided by fitting participants in a hierarchical model.

By default, when given multiple data sets, one per participant, `MemFit` will separately fit the model to each participant's data. Hierarchical modeling is performed by passing an optional parameter, 'UseHierarchical', to `MemFit`:

```
>> data1 = MemDataset(1);
>> data2 = MemDataset(2);
>> model = StandardMixtureModel();
>> fit = MemFit({data1,data2}, model, 'UseHierarchical', true)
```

Fitting such models is computationally more difficult, and so you should ensure the estimation procedure has correctly converged (e.g., using the `PlotConvergence` function provided by the toolbox) before relying on the parameter estimates to make inferences.

B.3.5 Model comparison

Which model best describes the data? Answering this question requires considering both the resemblance between the model and the data and also the model's flexibility. Flexible models can fit many data sets, and so a good fit provides only weak evidence of a good match between model and data. In contrast, a good fit between an inflexible model and the data provides stronger evidence of a good match. To account for this, many approaches to model comparison penalize more flexible models; these include the Akaike Information Criterion with correction for finite data (AIC_c; Akaike, 1974; Burnham & Anderson, 2004), the Bayesian Information Criterion (BIC; Schwarz, 1978), the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), and the Bayes factor (Kass & Raftery, 1995). It is also possible to perform cross-validation — fitting and testing separate data — to elim-

inate the advantage of a more flexible model. Implementations of some of these model comparison techniques are provided by the MemToolbox, and can be accessed by passing multiple models to the `MemFit` function:

```
>> model1 = StandardMixtureModel();
>> model2 = SwapModel();
>> modelComparison = MemFit(data, {model1, model2})
```

This will output model comparison metrics and describe them, including which model is preferred by each metric.

Despite the array of tools provided by the MemToolbox, we do not wish to give the impression that model selection can be automated. Choosing between competing models is no easier or more straightforward than choosing between competing scientific theories (Pitt & Myung, 2002). Selection criteria like AIC_c are simply tools for understanding model fits, and it is important to consider their computational underpinnings when deciding which criterion to use — before performing the analysis. For example, the criteria included in the toolbox are calibrated differently in terms of how strongly they penalize complex models, with criteria such as AIC_c having an inconsistent calibration that penalizes complex models less than criteria such as BIC, which penalizes complex models in a way that depends on their functional form, taking into account correlations between parameters. DIC is the only method appropriate in a hierarchical setting.

In addition to choosing an appropriate model comparison metric, we recommend computing the metric for each participant independently and looking at consistency across participants to make inferences about the best fitting models. Importantly, by fitting independently for each participant, you can take into account participant variance, and are thus able to generalize to the population as a whole (for example, by using an ANOVA over model likelihoods). By contrast, computing a single AIC_c or BIC value across all participants does not allow generalization to the population, as it ignores participant variance; one participant that is fit much better by a particular model can drive the entire AIC_c or

BIC value. This kind of *fixed effects* analysis can thus seriously overstate the true significance of results (Stephan et al., 2010). As in the case of estimating parameters, this technique of estimating model likelihoods for each participant and then performing an ANOVA or *t*-test over these parameters is only an approximation to the fully Bayesian hierarchical model that considers the evidence simultaneously from each participant (Stephan et al., 2009); however, in the case of model comparison, the simpler technique is likely sufficient for most visual working memory experiments.

To facilitate this kind of analysis, the MemToolbox performs model comparison on individual participant data and `MemFit` calculates many of the relevant model comparison metrics, so that you may choose the appropriate comparison for your theoretical claim.

B.4 Availability, contents, & help

The MemToolbox is available on the web at memtoolbox.org. To install the toolbox, place it somewhere sensible and then run the included Setup.m script, which will add it to MATLAB's default path. The distribution includes source code, demos, and a tutorial that reviews all of the toolbox's functionality. It is released under a BSD license, allowing free use for research or teaching. The organization of the toolbox's folder structure is outlined in the file `MemToolbox/Contents.m`. Detailed descriptions of each function (e.g., `MCMC`) can be found in the help sections contained in each file. To access the help section for some function `f` from the MATLAB prompt, run `help f`.

B.5 Conclusion

We created the MemToolbox for modeling visual working memory. The toolbox provides everything needed to perform the analyses routinely used in visual working memory, including model implementations, maximum likelihood routines, and data validation checks. In addition, it provides tools that offer a deeper look into the data and the fit of the model to the data. This introduction gave a

high-level overview of its approach and core features. To learn to use the toolbox, we recommend the tutorial, available at memtoolbox.org.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Alvarez, G. A. & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, 15(2), 106–111.

Anderson, B. D. O. & Moore, J. B. (2012). *Optimal Filtering*. Dover Publications.

Anderson, D. E., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *The Journal of Neuroscience*, 31(3), 1128–1138.

Anderson, J. R. (1989). A rational analysis of human memory. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 195–210).

Anderson, J. R. & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4), 703–719.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2(6), 396–408.

Anderson, R. B. (2001). The power law as an emergent property. *Memory & Cognition*, 29(7), 1061–1068.

Anderson, R. B. & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, 25(5), 724–730.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, 50(1), 5–43.

Åström, K. J. & Wittenmark, B. (2011). *Computer-controlled systems: Theory and design*. Courier Dover Publications.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, 18(7), 622–628.

Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211, 1390–1396.

Ayala, F. J. & Campbell, C. A. (1974). Frequency-dependent selection. *Annual Review of Ecology and Systematics*, 5, 115–138.

Baddeley, A. (1992). Working memory. *Science*, 255, 556–559.

Baddeley, A. D. (1986). *Working Memory*. Oxford University Press.

Barabási, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509–512.

Barthelmé, S. & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences*, 107(48), 20834–20839.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10), 1–27.

Bays, P. M. & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, 321, 851–854.

Becker, M. W., Pashler, H., & Anstis, S. M. (2000). The role of iconic memory in change-detection tasks. *Perception*, 29(3), 273–286.

Belson, K. (November 4, 2013). Mutai bides his time for a repeat victory. *The New York Times*, (pp. F 3).

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3), 351–368.

Bjork, R. A. (1972). Theoretical implications of directed forgetting. In A. W. Melton & E. Martin (Eds.), *Coding Processes in Human Memory* (pp. 217–235).

Bjork, R. A., Laberge, D., & Legrand, R. (1968). The modification of short-term memory through instructions to forget. *Psychonomic Science*, 10, 55–56.

Bjorklund, D. F. & Harnishfeger, K. K. (1990). The resources construct in cognitive development: Diverse sources of evidence and a theory of inefficient inhibition. *Developmental Review*, 10(1), 48–71.

Block, R. A. (1971). Effects of instructions to forget in short-term memory. *Journal of Experimental Psychology*, 89(1).

Blough, D. S. (1959). Delayed matching in the pigeon. *Journal of the Experimental Analysis of Behavior*, 2, 151–160.

Blume, L. E. (1993). The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 5, 387–424.

Bowman, H. & Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological Review*, 114(1), 38–70.

Brady, T. F. & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: ensemble statistics bias memory for individual items. *Psychological Science*, 22(3), 384–392.

Brady, T. F., Fougnie, D., & Alvarez, G. A. (2011a). Comparisons between different measures of working memory capacity must be made with estimates that are derived from independent data [Response to Anderson et al.]. *Journal of Neuroscience*, Oct. 14th.

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

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, 24(6), 981–990.

Brady, T. F. & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher-order regularities into working memory capacity estimates. *Psychological Review*, 120(1), 85–109.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, 109(2), 204–223.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3–5.

Burgess, A. E. & Colborne, B. (1988). Visual signal detection. IV. Observer inconsistency. *Journal of the Optical Society of America A*, 5(4), 617–627.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference. *Sociological Methods & Research*, 33(2), 261–304.

Burwitz, L. (1974). Proactive interference and directed forgetting in short-term motor memory. *Journal of Experimental Psychology*, 102(5), 799–805.

Chater, N. & Oaksford, M. (1998). *Rational models of cognition*. Oxford University Press.

Chee, M. W. L. & Choo, W. C. (2004). Functional imaging of working memory after 24 hr of total sleep deprivation. *The Journal of Neuroscience*, 24(19), 4560–4567.

Cheuvront, S. N. & Haymes, E. M. (2001). Thermoregulation and marathon running. *Sports Medicine*, 31(10), 743–762.

Coltheart, M. (1980). Iconic memory and visible persistence. *Perception & Psychophysics*, 27, 183–228.

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, 7(12), 547–552.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(01), 87–114.

Cowan, N. (2005). *Working memory capacity*. Psychology Press.

Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4), 450–466.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision Making, Affect, and Learning, Attention and Performance XXIII* (pp. 3–38). Oxford University Press.

Dayan, P. & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2), 185–196.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

Dobbs, A. R. & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, 4(4), 500–503.

D'Esposito, M. & Postle, B. R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia*, 37(11), 1303–1315.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. Teachers College, Columbia University.

Emrich, S. M., Riggall, A. C., LaRocque, J. J., & Postle, B. R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *The Journal of Neuroscience*, 33(15), 6516–6523.

Epstein, W. (1972). Mechanisms of directed forgetting. In G. Bower (Ed.), *Psychology of Learning and Motivation: Volume 6: Advances in Research and Theory* (pp. 147–191). Academic Press.

Estes, W. K. (1957). Of models and men. *American Psychologist*, 12(10), 609–617.

Ewens, W. J. (2004). *Mathematical population genetics: I. Theoretical introduction*, volume 27. Springer.

Fechner, G. T. (1860). *Elemente der Psychophysik (Elements of Psychophysics)*. Breitkopf & Härtel.

Feldman, J. L. & McCrimmon, D. (2008). Neural control of breathing. In L. R. Squire (Ed.), *Fundamental Neuroscience*. San Diego: Academic Press.

Ferber, S., Humphrey, G. K., & Vilis, T. (2005). Segregation and persistence of form in the lateral occipital complex. *Neuropsychologia*, 43(1), 41–51.

Flavell, J. H. & Wellman, H. M. (1977). Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–33). Erlbaum.

Fougnie, D. & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, 11(12), 1–12.

Fougnie, D., Brady, T. F., & Alvarez, G. A. (2014). If at first you don't retrieve, try, try again: The role of retrieval failures in visual working memory. In *Annual Meeting of the Vision Sciences Society*.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, 3, 1229.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2013). Gradual decay and death by natural causes in visual working memory. *Journal of Vision*, 13(9), 19.

Franconeri, S. L., Alvarez, G. A., & Cavanagh, P. (2013). Flexible cognitive resources: Competitive content maps for attention and memory. *Trends in Cognitive Sciences*, (pp. 134–141).

Fudenberg, D., Nowak, M. A., Taylor, C., & Imhof, L. A. (2006). Evolutionary game dynamics in finite populations with strong selection and weak mutation. *Theoretical Population Biology*, 70(3), 352–363.

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, 17(5), 673–679.

Gardner, J. L., Merriam, E. P., Movshon, J. A., & Heeger, D. J. (2008). Maps of visual space in human occipital cortex are retinotopic, not spatiotopic. *The Journal of Neuroscience*, 28(15), 3988–3999.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, 40(2), 177–190.

Gazzaniga, M. S. (2004). *The cognitive neurosciences*. MIT press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. CRC press.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–759.

Gold, J. I. & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535–574.

Gold, J. M., Carpenter, C., Randolph, C., Goldberg, T. E., & Weinberger, D. R. (1997). Auditory working memory and wisconsin card sorting test performance in schizophrenia. *Archives of General Psychiatry*, 54(2), 159.

Gold, J. M., Murray, R. F., Sekuler, A. B., Bennett, P. J., & Sekuler, R. (2005). Visual memory decay is deterministic. *Psychological Science*, 16(10), 769–774.

Goldman-Rakic, P. S. (1994). Working memory dysfunction in schizophrenia. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 6, 348–357.

Graff, S. (1953). *Work simplification applied to the office*. Master's thesis, City College School of Business and Civic Administration.

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review*, 71(5), 392–407.

Hahn, B., Kappenman, E. S., Robinson, B. M., Fuller, R. L., Luck, S. J., & Gold, J. M. (2011). Iconic decay in schizophrenia. *Schizophrenia Bulletin*, 37(5), 950–957.

Harnishfeger, K. K. & Bjorklund, D. F. (1993). The ontogeny of inhibition mechanisms: A renewed approach to cognitive development. In *Emerging themes in cognitive development* (pp. 28–49). Springer.

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

Hartman, M. & Hasher, L. (1991). Aging and suppression: Memory for previously relevant information. *Psychology and Aging*, 6(4), 587–594.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4(3), 267–272.

Holcombe, A. O. & Chen, W. (2013). Splitting attention reduces temporal resolution from 7 hz for tracking one object to < 3 hz when tracking three. *Journal of Vision*, 13(1), 1–19.

Hollerman, J. R. & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1(4), 304–309.

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. *NYU Center for Digital Economy Research Working Paper CeDER-10-01*.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4, 227–241.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, 186(1007), 453–461.

Jeopardy Productions (2014). *Jeopardy!* Did you know... http://www.jeopardy.com/.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, 59, 193–224.

Joormann, J. & Gotlib, I. H. (2008). Updating the contents of working memory in depression: interference from irrelevant negative material. *Journal of Abnormal Psychology*, 117(1), 182–192.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 237–285.

Kahneman, D. (1973). *Attention and effort*. Prentice Hall.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kalman, R. E. et al. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1), 35–45.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Katz, A. & Sahlin, K. (1988). Regulation of lactic acid production during exercise. *Journal of Applied Physiology*, 65(2), 509–518.

Kearns, M. & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3), 209–232.

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455, 227–231.

Keysers, C., Xiao, D.-K., Földiák, P., & Perrett, D. (2005). Out of sight but not out of mind: The neurophysiology of iconic memory in the superior temporal sulcus. *Cognitive Neuropsychology*, 22(3-4), 316–332.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, 36(14), 1–1.

Kleinrock, L. (1975). *Queueing systems. Volume 1: Theory*. Wiley.

Komarova, N. L. & Nowak, M. A. (2003). Language dynamics in finite populations. *Journal of Theoretical Biology*, 221(3), 445–457.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370.

Kuhbandner, C., Spitzer, B., & Pekrun, R. (2011). Read-out of emotional information from iconic memory. *Psychological Science*, 22(5), 695–700.

Kyllonen, P. C. & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, 14(4), 389–433.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1), 112–147.

Lieberman, E., Hauert, C., & Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature*, 433, 312–316.

Lothian, J. A. (2011). Lamaze breathing: What every pregnant woman needs to know. *The Journal of Perinatal Education*, 20(2), 118–120.

Lu, Z.-L., Neuse, J., Madigan, S., & Dosher, B. A. (2005). Fast decay of iconic memory in observers with mild cognitive impairments. *Proceedings of the National Academy of Sciences*, 102(5), 1797–1802.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

Luck, S. J. & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390, 279–281.

Luck, S. J. & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, 17(8), 391–400.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, 17(3), 347–356.

MacLeod, C. M. (1975). Long-term recognition and recall following directed forgetting. *Journal of Experimental Psychology: Human Learning and Memory*, 1(3), 271–279.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Ford, R. L. (1997). On regulation of recollection: The intentional forgetting of stereotypical memories. *Journal of Personality and Social Psychology*, 72(4), 709–719.

Magnussen, S. (2000). Low-level memory processes in vision. *Trends in Neurosciences*, 23(6), 247–251.

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience*, 9(4), 343–364.

Manoach, D. S. (2003). Prefrontal cortex dysfunction during working memory performance in schizophrenia: reconciling discrepant findings. *Schizophrenia Research*, 60(2), 285–298.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co.

Martinussen, R., Hayden, J., Hogg-Johnson, S., & Tannock, R. (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 44(4), 377–384.

Mason, W. & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 1–23.

Matthey, L., Bays, P., & Dayan, P. (2012). Probabilistic palimpsest memory: Multiplicity, binding and coverage in visual short-term memory. In COSYNE (pp. III–48).

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(3), 817–835.

Metcalfe, J. E. & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing.* The MIT Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.

Monahan, G. E. (1982). State of the art—a survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1), 1–16.

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1), 60–71.

Moran, P. A. P. (1962). *The statistical processes of evolutionary theory*. Clarendon Press.

Morey, R. D. (2011). A bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, 55(1), 8–24.

Muther, W. S. (1965). Erasure or partitioning in short-term memory. *Psychonomic Science*, 3, 429–430.

Neisser, U. (1967). *Cognitive psychology.* Appleton-Century-Crofts.

Nowak, M. A. (2006). *Evolutionary dynamics: Exploring the equations of life*. Belknap Press of Harvard University Press.

Nowak, M. A., Sasaki, A., Taylor, C., & Fudenberg, D. (2004). Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428, 646–650.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, 14(6), 769–776.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2), 329–337.

O'Reilly, R. C. & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2), 283–328.

Pan, J. (2010). *A comparative study of non-Markovian stochastic processes in marketing*. PhD thesis, Emory University.

Pannu, J. K. & Kaszniak, A. W. (2005). Metamemory experiments in neurological populations: A review. *Neuropsychology Review*, 15(3), 105–130.

Pashler, H. (1988). Familiarity and the detection of change in visual displays. *Perception & Psychophysics*, 44, 369–378.

Passolunghi, M. C. & Siegel, L. S. (2001). Short-term memory, working memory, and inhibitory control in children with difficulties in arithmetic problem solving. *Journal of Experimental Child Psychology*, 80(1), 44–57.

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, 35(4), 1.

Pelli, D. G. (1997). The Videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4), 437–442.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception and Psychophysics*, 16, 283–290.

Pitt, M. A. & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10), 421–425.

Postle, B. R., Berger, J. S., & D'Esposito, M. (1999). Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. *Proceedings of the National Academy of Sciences*, 96(22), 12959–12964.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc.

Rapoport, A. (1965). *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan Press.

Rensink, R. A. (1999). The magical number one, plus or minus zero. *Investigative Ophthalmology & Visual Science*, 40, 52.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358.

Rouder, J. N. & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, 12, 573–604.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, 105(16), 5975–5979.

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review*, 18(2), 324–330.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68(4), 589–606.

Salthouse, T. A. & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, 27(5), 763.

Sandberg, A., Lansner, A., Petersson, K. M., & Ekeberg, Ö. (2000). A palimpsest memory based on an incremental Bayesian learning rule. *Neurocomputing*, 32, 987–994.

Schira, M. M., Tyler, C. W., Spehar, B., & Breakspear, M. (2010). Modeling magnification and anisotropy in the primate foveal confluence. *PLoS Computational Biology*, 6(1), e1000651.

Schultz, W. (2010). Review dopamine signals for reward value and risk: Basic and recent data. *Behavioral and Brain Functions*, 6, 1–9.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, 1(3), 357–375.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, 6(5), 132–137.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Silberschatz, A., Galvin, P. B., & Gagne, G. (2013). *Operating system concepts*, volume 8. Wiley.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, 119(4), 807–830.

Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2008). Are there multiple visual short-term memory stores? *PLoS ONE*, 3(2), e1699.

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., & Erb, L. (1995). The uncertain response in the bottlenosed dolphin (Tursiops truncatus). *Journal of Experimental Psychology: General*, 124(4), 391–408.

Smith, J. D., Shields, W. E., Schull, J., & Washburn, D. A. (1997). The uncertain response in humans and animals. *Cognition*, 62(1), 75–97.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, 74.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & der Linde, A. V. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(4), 583–616.

Stanovich, K. E. & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, 23(05), 701–717.

Steenari, M.-R., Vuontela, V., Paavonen, E. J., Carlson, S., Fjällberg, M., & Aronen, E. T. (2003). Working memory and sleep in 6-to 13-year-old schoolchildren. *Journal of the American Academy of Child & Adolescent Psychiatry*, 42(1), 85–92.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, 46(4), 1004–1017.

Stephan, K. E., Penny, W. D., Moran, R. J., Den Ouden, H. E. M., Daunizeau, J., & Friston, K. J. (2010). Ten simple rules for dynamic causal modeling. *Neuroimage*, 49(4), 3099–3109.

Stevens, S. S. (1975). *Psychophysics*. Transaction Publishers.

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, 13(10), 9.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An Introduction)*. The MIT Press.

Taha, H. A. (2007). *Operations research: An introduction*. Pearson/Prentice Hall.

Tanaka, A. & Funahashi, S. (2012). Macaque monkeys exhibit behavioral signs of metamemory in an oculomotor working memory task. *Behavioural Brain Research*, 233(2), 256–270.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2), 119–164.

Todd, J. J. & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, 428(6984), 751–754.

Todd, M. T., Niv, Y., & Cohen, J. D. (2008). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In *Advances in Neural Information Processing Systems* (pp. 1689–1696).

Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28), 11478–11483.

Traulsen, A., Pacheco, J. M., & Nowak, M. A. (2007). Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of Theoretical Biology*, 246(3), 522–529.

Unsworth, N. & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between operation span and raven. *Intelligence*, 33(1), 67–81.

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, 109, 8780–8785.

Van Der Merwe, R., Doucet, A., De Freitas, N., & Wan, E. (2000). The unscented particle filter. In *NIPS* (pp. 584–590).

Vergauwe, E., Barrouillet, P., & Camos, V. (2009). Visual and spatial working memory are not that dissociated after all: A time-based resource-sharing account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 1012–1028.

Vogel, E. K. & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory. *Nature*, 428, 748–751.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, 438(7067), 500–503.

Voytek, B. & Knight, R. T. (2010). Prefrontal cortex and basal ganglia contributions to visual working memory. *Proceedings of the National Academy of Sciences*, 107(42), 18167–18172.

Vul, E. (2010). *Sampling in human cognition*. PhD thesis, Massachusetts Institute of Technology.

Wegner, D. M. (2009). How to think, say, or do precisely the worst thing for any occasion. *Science*, 325, 48–50.

Wegner, D. M. & Erber, R. (1992). The hyperaccessibility of suppressed thoughts. *Journal of Personality and Social Psychology*, 63(6), 903–912.

Wellman, H. M. (1977). Tip of the tongue and feeling of knowing experiences: A developmental study of memory monitoring. *Child Development*, 48, 13–21.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition*, 2(4), 775–780.

Wilken, P. & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, 4(12), 1120–1135.

Williams, M., Hong, S. W., Kang, M. S., Carlisle, N. B., & Woodman, G. F. (2012). The benefit of forgetting. *Psychonomic Bulletin & Review*, 20(2), 348–355.

Wixted, J. T. (1990). Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(5), 927–935.

Wixted, J. T. & Carpenter, S. K. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science*, 18(2), 133–134.

Wixted, J. T. & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, 2(6), 409–415.

Wood, J. N. (2009). Distinct visual working memory systems for view-dependent and view-invariant representation. *PLoS One*, 4(8), e6601.

Woodward, A. E. & Bjork, R. A. (1971). Forgetting and remembering in free recall: Intentional and unintentional. *Journal of Experimental Psychology*, 89(1), 109–116.

Xu, Y. & Chun, M. M. (2005). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, 440, 91–95.

Yang, W. (1999). *Lifetime of human visual sensory memory: Properties and neural substrate*. PhD thesis, New York Univeristy.

Zacks, R. T. (1989). Working memory, comprehension, and aging: A review and a new view. *Psychology of Learning & Motivation*, 22, 193–225.

Zhang, W. & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453, 233–235.

Zhang, W. & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science*, 20(4), 423–428.

Zilli, E. A. & Hasselmo, M. E. (2008). The influence of Markov Decision Process structure on the possible strategic use of working memory and episodic memory. *PLoS ONE*, 3(7), e2756.

This dissertation was typeset with LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice and issued by Adobe in 2007. A template that can be used to format a dissertation with this look & feel has been released under the permissive MIT (X11) license, and can be retrieved online at github.com/suchow/Dissertate or from its author at suchow@post.harvard.edu.