THE UNIVERSITY OF CHICAGO

CHARACTERIZING THE LIMITS OF VISUAL WORKING MEMORY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF PSYCHOLOGY

BY

KIRSTEN CAROLINE SANDAGE ADAM

CHICAGO, ILLINOIS

JUNE 2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF PUBLICATIONS

Foster, J.J.* & Adam, K.C.S.* (2014). Is feature-based attention spatially global during visual search? *The Journal of Neuroscience, 34*(26), 8862-8864.

Adam, K.C.S.*, Mance, I.*, Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience*, *27*(8), 1601–1616. **This article is reproduced in full as Chapter 2.**

Adam, K.C.S. & Vogel, E.K. (2016). Reducing failures of working memory with performance feedback. *Psychonomic Bulletin & Review*, *23*(5), 1520-1527.

Adam, K.C.S. & Vogel, E.K. (2016). Tuning in by tuning out distractions. *Proceedings of the National Academy of Sciences, 113*(13), 3422-3423.

Adam, K.C.S. & Vogel, E.K. (2017). Confident failures: Lapses of working memory reveal a metacognitive blind spot. *Attention, Perception & Psychophysics*, *79*(5), 1506-1523.

Adam, K.C.S., Vogel, E.K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79-97. **This article is reproduced in full as Chapter 3.**

Xu, Z.*, Adam, K.C.S.* & Fang, X., & Vogel, E.K. (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods, 50*(2), 576-588.

Adam, K.C.S., Robison, M.K., & Vogel, E.K. (2018). Contralateral delay activity tracks fluctuations in working memory performance. *Journal of Cognitive Neuroscience*.

* Indicates that these authors contributed equally to the work

ACKNOWLEDGEMENTS

It's often said that "you are what you pay attention to," but I think a truer statement is that those who pay attention to you make you who you are. I'd like to thank all of those who have taken so much of their time to pay attention to me and help me grow.

To my lab-mates and departmental peers, both in Chicago and Eugene—Thank you for always being there to discuss research (and, more importantly, for being there to *not* discuss research). From scrambling up Spencer's Butte to navigating city high-rises, I've never been short of friends. Special thanks are due to David Sutterer and Joshua Foster – moving institutions was bound to be challenging, and I can't imagine having done it without you. I don't have space to list the names of my peers who have positively impacted my life during graduate school, and for this I feel lucky and grateful – thank you, all.

To my graduate advisors, Ed Vogel and Ed Awh—Thank you for guiding my research all these years, and for fostering a research environment that is collaborative, constructive, and fun.

To all the research assistants and support staff working "behind the scenes" – who make sure the facilities are working and reimbursements are processed and that grad students don't absent-mindedly forget deadlines – Thank you! This work wouldn't be possible without you.

Finally, I wouldn't have made it through without the loving support of my family. To my parents, Dean and Cheryl Adam, thank you for being so supportive of me as I've traversed the country (emotionally but also literally hauling boxes across the country and back again). To my sister, Manda Adam, thank you for always having a listening ear and knowing the perfect silly thing to brighten my day. To my extended family – thanks for your support and good humor in reminding me that they don't teach common sense in graduate school.

ABSTRACT

Visual working memory is a mental workspace used to temporarily maintain and manipulate visual information. Although we can perceive the visual world around us in seemingly infinite richness and detail, only a fraction of these details can be held in working memory at any given time. Thus, working memory is described as a capacity-limited system, and individual differences in working memory capacity strongly predict individual differences in fluid intelligence and academic achievement. Despite the importance of working memory capacity for cognition, the nature and definition of capacity limits are still under extensive debate. In this dissertation, I use novel behavioral and neural measures to interrogate the limits of visual working memory and to inform these ongoing debates. In Chapter 1, I summarize key questions related to the limitations of visual working memory, focusing on capacity limits in working memory and individual differences in these limits. In Chapter 2, I characterize the dynamic nature of visual working memory performance by measuring and modeling how fluctuations of attention influence working memory. In Chapter 3, I address critiques of capacity limits by theoretical models which propose that all items receive mnemonic resources (i.e., no truly uninformed guesses are necessary). I demonstrate how looking at within-trial mnemonic variability, rather than average data, can provide clear evidence against these capacity-unlimited models. Finally, in Chapter 4, I ask whether we can further bolster evidence for capacity limits by decoding the content of active working memory signals in EEG (the topography of alpha-band power). Together, these studies inform contemporary models of visual working memory, supporting a capacity-limited system whose efficacy varies dynamically under the influence of attention.

CHAPTER 1. INTRODUCTION TO KEY QUESTIONS ABOUT LIMITATIONS OF VISUAL WORKING MEMORY

Working memory is a cognitive workspace used to temporarily hold and manipulate information in mind. Although we can perceive the rich detail of the world around us, there are strict limits on the amount of information that we can actively maintain at any given time. Thus, working memory is said to be capacity-limited. Working memory capacity can be measured with stimuli from many domains, such as auditory (Prosser, 1995), visual (Luck & Vogel, 1997), tactile (Katus, Müller, & Eimer, 2015), verbal (Daneman & Carpenter, 1980), and numerical (Turner & Engle, 1989). Each of these domain-specific tasks predicts a common, domain-general working memory capacity (e.g., Unsworth, Fukuda, Awh, & Vogel, 2014), which is an important predictor of broader cognitive abilities. For example, working memory capacity is positively correlated with fluid intelligence (Engle, Kane, & Tuholski, 1999; Unsworth et al., 2014), standardized test scores (Engle, Tuholski, Laughlin, & Conway, 1999), and reading comprehension (Daneman & Carpenter, 1980). Conversely, working memory capacity is disrupted in clinical populations, such as in individuals with Schizophrenia (Gold et al., 2010) and Parkinson's (Lee et al., 2010). As such, understanding working memory capacity limits is critical for understanding the basic building blocks of cognition and for understanding the complex pathology of many neuropsychiatric disorders.

In this dissertation, I will use behavior, modeling, and neural measures to address ongoing debates about limitations of visual working memory. Below, I provide a brief background of the measurement of visual working memory and I outline key debates surrounding the structure and limitations of visual working memory.

Estimating capacity limits in visual working memory

The study of capacity limits in visual working memory blossomed following the popularization of a visual change-detection task (Luck & Vogel, 1997). In this task, participants briefly view an array of items (e.g., colored squares), remember these items across a delay, and then are tested on one of the items (e.g., is this square the same or different color from the one in the memory array?). To measure working memory capacity, participants view memory arrays with different numbers of items ("set sizes"). At set size 1, participants perform nearly perfectly (100% accuracy). At set size 10, they perform abysmally (near chance). A hallmark finding in Luck & Vogel (1997) was a precipitous drop in accuracy after the set size surpassed 3 items. This sudden drop in performance suggests that visual working memory capacity is limited to only 3 or 4 simple visual items. Subsequent work converting responses (hits, false alarms) to an estimate of the number of remembered items ("K"), revealed that this drop in raw accuracy is consistent with a plateau in the number of items stored (Cowan, 2001; Rouder, Morey, Morey, & Cowan, 2011).

Visual working memory tasks have been critical for better understanding the mechanisms underlying the temporary storage of information held in mind. With a short trial length and simple task instructions, change detection can be used in developmental samples (Cowan, Fristoe, Elliott, Brunner, & Saults, 2006; Isbell, Fukuda, Neville, & Vogel, 2015), clinical populations (Gold et al., 2010; Lee et al., 2010), and other species (Buschman, Siegel, Roy, & Miller, 2011; Gibson, Wasserman, & Luck, 2011), allowing for the rare opportunity to compare task performance and neural measures in the same task across species (Elmore, Magnotti, Katz, & Wright, 2012; Reinhart et al., 2012). The short trial-duration of the visual change detection task also allows for the collection of a large number of trials, yielding robust measurements using

single-unit electrophysiology, electroencephalogram (EEG), and functional magnetic resonance imaging (fMRI). In EEG, a measure of lateralized activity to a memory array (the contralateral delay activity, CDA) reaches a plateau at typical capacity limits, and this plateau correlates with individual differences in behavioral working memory capacity (Luria, Balaban, Awh, & Vogel, 2016; Vogel & Machizawa, 2004; Vogel, McCollough, & Machizawa, 2005). Likewise, in fMRI, univariate activity in the intraparietal sulcus (IPS) increases gradually as the number of remembered items increases, but reaches a plateau at typical capacity limits (Todd & Marois, 2004, 2005). Thus, both behavior and electrophysiological signals corroborate a working memory capacity that is limited to around 3-4 simple visual items.

Does attentional control determine individual differences in capacity?

While average visual working memory capacity is around 3 to 4 items, there are stable and significant individual differences in this number (Fukuda, Woodman, & Vogel, 2015; Z. Xu, Adam, Fang, & Vogel, 2017). Working memory capacity is often thought of as a "space" where we can put information, so individual differences in working memory capacity are typically assumed to measure the size of this available space. By analogy, if you own a bigger suitcase, then you can carry more items with you.

However, some work suggests that other factors besides capacity *per se* may contribute to individual differences in working memory performance. For example, in "filtering tasks" participants are asked to store a subset of items from a display and filter out irrelevant items (e.g., remember red bars but ignore blue bars). Individual differences in the efficacy of ignoring irrelevant items strongly predict individual differences in working memory capacity (McNab & Klingberg, 2008; Vogel et al., 2005). In addition, individuals with poor working memory ability have difficulty suppressing irrelevant visual information (Fukuda & Vogel, 2009; Zanto &

Gazzaley, 2009). By analogy, this work suggests that individual differences may be due not to the storage space available (e.g., the size of your nightclub) but rather to the efficiency of attentional control (e.g., how good your bouncers are; Awh & Vogel, 2008; McNab & Klingberg, 2008).

Because of the strong relationship between attention and working memory (Awh & Jonides, 2001; Chun, 2011; Engle, 2002), a central question is how dynamic fluctuations of attention impact working memory performance, and whether these fluctuations of attention may actually underlie individual differences in working memory capacity. To assess this, in Chapter 2 I will measure trial-by-trial fluctuations in working memory performance using a whole-report working memory task. Further, I will use previously established event-related potentials (ERPs) and oscillatory markers of distinct cognitive processes (e.g., encoding, executive control) to assess which of aspects of task performance are disrupted during failures of working memory performance.

Is visual working memory truly capacity-limited?

Thus far in this chapter, I have assumed a capacity limit in visual working memory. This assumption is based on work in which behavioral performance falls precipitously beyond 3 to-be-remembered items and neural signals which plateau at this same capacity limit. However, competing models of working memory challenge the claim that working memory is truly capacity limited (i.e., some items are not stored). Instead, these competing models posit that working memory is information-limited (i.e., some information is stored about all items, but this information may be very imprecise). Such an information-limited model is capacity-unlimited; in that it assumes that subjects never completely fail to store some information about an item.

The debate between capacity-limited and information-limited models (Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007) was accelerated by measures of visual working memory performance that assess fine-grained information about the quality of remembered items (Wilken & Ma, 2004; Zhang & Luck, 2008). In these continuous-report tasks, participants remember a memory array across a blank delay, and then are probed to freely recall the identity of one of the items from the array. For example, if the participants are asked to remember color, then they might choose a color from a color wheel that contains a circularly-wrapped, continuous color space (e.g., 360 colors, presented on a continuous rainbow). This task yields a continuous distribution of response errors for each set size. A key insight by Zhang and Luck (2008) was that this distribution of errors can be modelled by a two-state mixture model where some proportion of the time subjects fail to store an item and guess (yielding a uniform distribution of response errors) and other times they have information about the item in mind (yielding a normal distribution). This model yielded good fits, and changes in the guessing parameter predict individual differences in typical measures of working memory capacity and associated cognitive abilities (Unsworth et al., 2014). However, competing models have proposed that, instead, the response distribution for large set sizes can be accounted for by only a circular normal distribution or by a mixture of many different widths of normal distributions (Bays, Catalao, & Husain, 2009; Bays & Husain, 2008; van den Berg, Shin, Chou, George, & Ma, 2012).

Competing models have become quite sophisticated, yet have reached an empirical stalemate. The same distribution of behavioral errors can be described equally well by the most competitive capacity-limited and capacity-unlimited models of visual working memory (Donkin, Kary, Tahir, & Taylor, 2016; Nosofsky & Donkin, 2016b; van den Berg, Awh, & Ma, 2014). Thus, Chapters 3 and 4 of this dissertation circumvent this stalemate by developing behavioral

and neural measures that focus on a central remaining question in the debate – do participants ever fail to store items? Capacity-limited models predict that subjects must guess when they fail to store an item, whereas capacity-unlimited models predict that any response will have some information – however imprecise – about the true identity of the probed memory item. I explicitly pit these two models against each other by developing a novel behavioral task that assesses fine-grained information about the quality of all items from supra-capacity arrays. Then, I use EEG to bolster behavioral evidence of guessing, showing that failure to store an item is not an artifact of making response, but is evident in item-specific memory signals during the delay period.

Overview of experiments

In Chapter 2, I will present evidence that the waxing and waning of attention impacts working memory performance. Further, I will argue that most individuals share the same true capacity (e.g., a common limit on the amount of information that can be stored) but differ in the consistency with which they achieve this mostly-common capacity.

In Chapter 3, I will present new evidence in favor of capacity limits in visual working memory. To do so, I will obtain fine-grained estimates of the quality of multiple items held simultaneously in mind (rather than probing a single item). In addition, I will present simulation results which demonstrate that the poorest memories posited by capacity-unlimited models are best described as guesses (i.e., uniformly distributed errors), not as low-precision memories.

Finally, in Chapter 4, I will present novel methods for assessing capacity limits in visual working memory using neural measures of active working memory storage. Previous neural measures of capacity limits have relied on interpreting a "plateau" in a univariate signal. This approach is limited, because it does not definitively demonstrate that the information contained

in this signal is specific to only a subset of items from the array. To circumvent this issue, I will use multivariate methods in EEG to attempt to decode item-specific information about multiple items from the memory array.

CHAPTER 2. THE CONTRIBUTION OF ATTENTIONAL LAPSES TO INDIVIDUAL DIFFERENCES IN VISUAL WORKING MEMORY CAPACITY

Introduction

Individuals with low Working Memory (WM) capacity perform poorly on measures of aptitude such as fluid intelligence and academic achievement (Daneman & Carpenter, 1980; Daneman & Green, 1986; Engle, Tuholski, et al., 1999; Fukuda, Vogel, Mayr, & Awh, 2010; Turner & Engle, 1989; Unsworth et al., 2014) and extensive work has suggested that individual differences in capacity stem in part from variability in deploying attention (Fukuda & Vogel, 2009, 2011). Low capacity individuals have specific difficulties in tasks that require attentional control, suggesting that variability in effectively exerting these control mechanisms determines apparent capacity differences. However, because capacity measures are calculated from average performance across an entire session, a common alternative hypothesis remains: individual differences in capacity are the result of a mixture of "complete attention trials", in which subjects maximally allocate their available memory resources, and "lapse trials", in which subjects fail to allocate any available memory resources due to complete disengagement from the task. Thus, low capacity individuals may simply have more lapse trials than high capacity individuals and consequently show reduced average performance.

A lapse account of working memory performance is consistent with evidence suggesting that task engagement is associated with differences in working memory ability. First, low capacity individuals engage in mind wandering more frequently than high capacity individuals, particularly during cognitively challenging tasks (McVay & Kane, 2010; Mrazek et al., 2012; Smallwood & Schooler, 2006). Second, the slowest reaction time trials in choice reaction time

tasks are the most predictive of intelligence scores (i.e., the worst performance rule) suggesting that attentional lapses contribute to estimates of individual aptitude (Coyle, 2003). Finally, low capacity individuals exhibit periods of goal neglect, a performance failure in which task rules are explicitly understood, but are nevertheless not behaviorally executed (Duncan, Emslie, Williams, Johnson, & Freer, 1996; Duncan, Schramm, Thompson, & Dumontheil, 2012). Thus, low capacity subjects may experience an increased number of failures for a variety of reasons, from being captured by distracting internal thoughts to simply giving up on a difficult task.

We have described a *lapse model* of inattention, in which inattention leads to complete disengagement from the task. However, inattention could also manifest as degraded attentional control rather than as a complete lapse. Under an *attentional control model* of inattention, impaired attention would lead to reduced performance, though not necessarily to total neglect of the task if the inattention is incomplete. From this view, the efficiency of attentional control may vary over the session in a graded fashion; differences between subjects could be viewed as a shift in the distribution of effective attentional engagement. Low capacity individuals would have a downward shift in this distribution, leading to more frequent periods of partial disengagement than high capacity subjects. Indeed, the evidence reviewed above from mind wandering frequency and reaction time variability would be consistent with either complete or graded failures of attention. However, these two models have not been evaluated directly against one another. Critically, most attention and working memory tasks rely on one of two metrics: accuracy and reaction times. Accuracy measures are binary (correct or incorrect), making it difficult to test for a graded attentional model. Conversely, reaction time measures are continuous, making it difficult to test for a complete lapse model. Thus, it is not surprising that the attention literature (predominately reaction time measures) has strongly assumed a graded

model of attention, while the working memory capacity literature (predominately signal detection measures) has tested only coarse lapse parameters.

For example, in the change detection paradigm, a randomly chosen item from each memory array is probed and the subject must indicate whether it is the same as the originally presented item. Here, an individual's capacity is estimated across a series of trials by determining the probability of having stored the probed item on a given trial. Consequently, performance on an individual trial is not informative on its own because an error could be the result of either a complete memory failure (0 items stored) or a successful memory (4 items stored) that was unlucky because the probed item did not happen to be stored. At present, only the complete lapse model has been directly tested in the working memory literature (Morey, 2011; Rouder et al., 2008, 2011; Sims, Jacobs, & Knill, 2012; van den Berg et al., 2014). For example, Rouder et al. (2008, 2011) found that adding a lapse parameter substantially improved model estimates of working memory capacity in a change detection task, particularly by accounting for why errors occur on sub-capacity memory arrays. This work demonstrates that an "all or nothing" lapse account could plausibly explain individual differences in capacity. However, due to the aggregate nature of the data, a graded attentional control account could not be evaluated.

Here, we measure how WM performance fluctuates within a session to determine whether performance failures are better explained by a lapse model (coarse failures) or by an attentional control model (graded failures). To do so, we employ a novel whole-report task that provides a trial-by-trial measurement of the total number of correctly reported items from each array, allowing us to examine the distribution of levels of success on each trial. Our novel whole-report measure is both discrete (each item is correct or incorrect) and continuous (the number of objects correct can fluctuate), allowing us to uniquely distinguish between these two hypothetical

models.

Measuring performance fluctuations with continuous whole-report allows us to test distinct predictions made by the competing attentional control and lapse accounts of under-performance. First, while a lapse model predicts that lower capacity individuals would show higher lapse rates than high capacity individuals regardless of task load, an attentional control model predicts that substantial performance failures related to individual capacity would only be observed under task loads that necessitate attentional control, such as for supracapacity displays. Second, while a lapse model predicts a bimodal distribution of performance (i.e., lapse vs. full attention), an attentional control model predicts a graded downward shift in performance distributions for low capacity individuals. After distinguishing between lapse and attentional control models of performance fluctuations, we test several hypotheses about the mechanisms of performance fluctuations. These mechanisms include changes in task engagement over time, task non-compliance, sensory encoding, and attentional control.

In three experiments, we establish the validity of a novel working memory measure and then test hypotheses that differentiate a lapse model and an attentional control model of individual differences in WM capacity. In a discrete whole-report measure of working memory, subjects view a display of brightly colored items, remember the display, and then are asked to identify the color of each item from a fixed set of color choices. Task accuracy is determined by counting the number of correctly identified items. Thus, a subject's level of performance can be calculated for every trial in the experiment. This trial-by-trial measurement of performance is critical for investigating novel hypotheses about the nature of trial-by-trial task engagement.

In Experiment 2-1a, we tested whether set-size affects the number of performance failures. A coarse lapse model would predict that the amount of time spent disengaged from the task does

not vary across set-sizes (Rouder et al., 2008, 2011). Alternatively, an attentional control model would predict an increased rate of performance failures (few items correct) when the set-size exceeds capacity. We had participants complete a change detection memory task and a whole-report memory task for multiple set-sizes (two to six items). We include a change detection measure for two reasons: (1) to initially validate our novel whole-report measure of working memory and (2) as an independent measure that allows us to more closely examine the contributions of within- and between-subject variability in working memory performance while minimizing issues of circularity.

In Experiment 2-1b, we collected a large number of supra-capacity trials in order to test lapse and attentional control models and to examine the consistency of performance failures over time. Subjects completed a large number of exclusively set-size six trials (along with a standard change detection measure). By continuously repeating the same challenging set-size, we could look for fluctuations in engagement over time without confounding carryover effects from conditions with different levels of difficulty.

Finally, we tested whether neural measures of sensory encoding and attentional control predict trial-by-trial working memory performance. In Experiment 2-2, a new set of participants performed the same tasks as in Experiment 2-1b while EEG and EOG activity was recorded. Using only the set-size six condition was also ideal for EEG analyses, as we could examine fluctuations in neural signals while holding physical stimulation constant. To examine markers of low-performance trials, we analyzed the P1/N1 visual-evoked response and behavioral accuracy for artifact-rejected trials. To examine the potential contribution of attentional control, we measure frontal theta power and posterior alpha power across all time points in the trial.

Materials & Methods

Participants

All subjects gave written informed consent according to procedures approved by the University of Oregon institutional review board. Subjects were compensated for participation with course credit or monetary payment ($8/hour for behavior, $10/hour for EEG). A unique set of subjects participated in each experiment, with 40 in Experiment 2-1a, 45 in Experiment 2-1b, and 26 in Experiment 2-2. One subject from Experiment 2-1b was excluded from analyses for failure to comply with task instructions. After artifact rejection, 3 subjects were excluded from Experiment 2-2 analyses for artifact rates in excess of 25% or fewer than 300 remaining trials; 1 subject did not complete the change detection measure, so they were excluded from between-subjects analyses using change detection, but included in within-subjects analyses.

Stimuli

Stimuli were generated in MATLAB (The MathWorks, Natick, MA) using the Psychophysics toolbox (Brainard, 1997; Pelli, 1997), and presented on 21-inch CRT monitors. Subjects were seated approximated 60 cm from the monitor. In Experiment 2-1a, stimuli (2.50° visual angle) for both whole-report and change detection tasks consisted of 8 colors (RGB values: Red=255 0 0; Green= 0 255 0; Blue= 0 0 255; Magenta=255 0 255; Yellow= 255 255 0; Cyan= 0 255 255; White=255 255 255; Black= 0 0 0), presented on a gray background (RGB=128 128 128). In Experiment 2-1b and Experiment 2-2, one additional color (RGB: Orange = 255 128 0) was added to the potential memory set colors.

Tasks

Change detection task. The change detection task used in all experiments followed standard procedures (Luck & Vogel, 1997). In Experiment 2-1a, subjects were presented with

arrays of two to six colored squares for 150 ms (memory array), which disappeared for 900 ms (retention period), followed by the presentation of one colored square (test probe) at the location previously occupied by an item in the memorandum. In Experiment 2-1b and Experiment 2-2, subjects were presented with arrays of four, six, or eight items, and trials used a 250 ms stimulus array and 1000 ms retention period. On 50% of trials, the test item was the same as the item presented in the memory array, and in the remaining 50% of trials the test item was different. Participants were instructed to make an unspeeded button press to indicate whether the color of the probe had changed. The next trial began after an inter-trial interval (ITI) of either 900 ms (Experiment 2-1a), or 1000 ms (Experiment 2-1b and Experiment 2-2).

Whole-report task. The whole-report procedure was similar to the change detection task with the exception that individuals recalled each item shown in the memory array. At response, individuals were shown a three by three matrix of colors over each location of memory array items; they were instructed to use a mouse to click on the color square corresponding to the memory array item at each location. In Experiment 2-1a, individuals were encouraged to respond to as many items as they could remember, and advanced to the next trial by pressing the spacebar. In Experiments 2-1b and 2-2, individuals were required to respond to all items in the memory array. The next trial began when all responses were made (Experiment 2-1b) or after the subject clicked to indicate they were ready for the next trial (Experiment 2-2). Stimulus timing parameters were the same as for the respective change-detection task except for Experiment 2-2. In Experiment 2-2 the retention interval and ITI periods were increased to 1300 ms to provide a larger time window for oscillatory analyses.

Procedures

Experiment 2-1a. Participants completed two blocks of 150 trials of the change detection task and the whole-report task across several set-sizes (two to six items). Set-sizes were intermixed within blocks for all experiments. For each of the two tasks, subjects completed 150 trials, for a total of 30 trials per set-size.

Experiment 2-1b. Participants performed five blocks of 36 trials of the change-detection task, for a total of 60 trials per set-size. For the whole-report task, participants completed 10 blocks of 30 trials (300 trials total), and all arrays were set-size 6.

Experiment 2-2. Participants performed the same tasks as in Experiment 2-1b, while we recorded EEG activity. Participants performed five blocks of 36 trials of the change-detection task, for a total of 60 trials per set-size. For the whole-report task, participants completed 15 to 18 blocks of 30 trials (450 to 540 trials total), and all arrays were set-size six.

EEG Data Collection

EEG activity was recorded at 20 electrodes mounted in an elastic cap (ElectroCap International) using our standard recording and analysis procedures (McCollough, Machizawa, & Vogel, 2007). The International 10/20 sites F3, FZ, F4, T3, C3, CZ, C4, T4, P3, PZ P4, T5, T6, O1 AND O2 were used along with five nonstandard sites: OL midway between T5 and O1; OR midway between T6 and O2; PO3 midway between P3 and OL; PO4 midway between P4 and OR; and POz midway between PO3 and PO4. All sites were recoded with a left-mastoid reference, and the data were re-referenced off-line to the algebraic average of the left and right mastoids. Horizontal electrooculogram (EOG) was recorded from electrodes placed ~1 cm to the left and right of the external canthi of each eye to measure horizontal eye movements. To detect blinks, vertical EOG was recorded from an electrode mounted beneath the left eye and

referenced to the left mastoid. The EEG and EOG signals were amplified with an SA Instrumentation amplifier with a bandpass of 0.01 – 80 Hz and were digitized at 250 Hz in Labview 6.1 running on a PC.

Trials including blocking, blinks, or large (>1°) ocular movements were rejected and excluded from further analysis. For ERP analyses, we baselined the signal over the 200 ms prior to the timelocking event (onset of the memory array); trials were filtered with a low-pass finite impulse response filter with a cutoff of 40 Hz. For oscillatory analyses, we bandpass filtered the raw EEG using a two-way, least-squares finite impulse response filter using the eegfilt.m filter function from the EEGLAB Toolbox (Brainard, 1997; Delorme & Makeig, 2004). We applied the MATLAB Hilbert transform (hilbert.m) to extract the instantaneous power values. For spectrograms, power data was calculated separately for each 1 Hz band. For band-specific analyses, power data was calculated for typically defined frequency bands (Theta: 4-7Hz; Alpha: 8-12Hz; Beta: 13-22Hz).

Statistical Analyses

Change detection capacity. Each participant's change detection accuracy was transformed into a *K* estimate using Cowan's formula $K=N*(H-FA)$; here *N* represents the set-size, *H* is the hit-rate, and *FA* is the false alarm rate (Cowan, 2001). In Experiment 2-1a, the average of set-size four, five, and six arrays was used to estimate each participant's change detection capacity. In Experiment 2-1b and Experiment 2-2, all set-sizes (four, six, and eight were used to estimate capacity).

Whole-report accuracy. For the whole-report procedure, we estimated accuracy in two ways. First, we calculated the average number of correctly reported items per set-size. Second, we split performance into the proportion of trials in which participants correctly reported zero,

one, two, etc. for each set-size. This method allowed us to measure the proportion of trials during which participants exhibited impaired WM performance, and how failures varied as a factor of set-size and WM capacity. By examining trial-by-trial accuracy, we can observe the impact of performance fluctuations on overall WM ability.

Simulation Analyses

The greater number of trials in Experiment 2-1b allowed us to perform a finer analysis of trial-by-trial performance. First, we characterized the expected performance outcome if subjects had a complete attentional lapse. We ran 30 iterations of a simulation where the computer guessed colors (without replacement) for the six items in the display across 300 trials. We also wanted to characterize the effect of guessing inflation on performance, especially for high-performance trials. To do so, we assigned three correct objects to each trial in the simulation and examined the effect of guessing colors without replacement for the remaining three items. Finally, we ran simulations to test whether a complete lapse model or a graded attentional control model could explain trial-by-trial fluctuations in WM performance. The two hypothetical models both predict that trial-by-trial performance in the whole-report task can be modeled by (1) an upper limit on total available WM resources (2) a probability of allocating the available WM resources. The lapse and attentional control models differ only in the higher order distribution used to predict the probability of allocating WM resources on a trial-by-trial basis. Importantly, the parameter values for the models were chosen only with respect to mean performance; the model-fitting procedure was blind to the underlying distribution of trials. After the model that best fit the mean was chosen (minimum difference between true and simulated mean), we used the residuals of the model fits (RMSE) to test the fit of each lapse model across subjects.

Lapse model. The lapse model was based on the assumption that individuals are either fully on task or completely disengaged from the task (Morey, 2011; Rouder et al., 2008, 2011; van den Berg et al., 2014). This model predicts that during full-engagement trials, participants will be able to maximally allocate WM resources (maximum capacity), with guessing inflation accounting for the trials in which the number of items correct exceeds this capacity. For fully disengaged trials, the model predicts that responses will only be based on guessing. The higher-order distribution of the complete lapse model is Bernoulli-distributed, a distribution with only two values, zero and one. The proportion of zeros in the distribution represents disengaged trials, and the proportion of ones represents fully-engaged trials. On each simulated trial, one value is pulled from the higher-order distribution and multiplied by the maximum capacity parameter to yield the performance outcome. For example, if the maximum capacity parameter is three items, then on a trial where a "1" is pulled, the subject achieves 1 x 3, or three items. Guessing is accounted for the remaining items in the set-size (in this case, three guesses). For each subject and parameter level, we simulated 300 trials for each subject. During each run of the simulation, maximum capacity was initially held constant at three items, while the proportion of lapses (0's) in the higher-order distribution was parametrically increased from 0 to 100% in steps of 1%. We allowed for guessing by randomly drawing without replacement from the nine possible colors. For example, for a lapse parameter of 20%, we sampled only from the random guessing distribution for 20% of the trials, while for the remaining 80% of trials, capacity was set to three items plus three guesses. The simulated mean number correct for each run was compared to each individual's empirical mean number correct in order to determine the best-fit lapse parameter. This was repeated for each subject in the data set. The lapse parameter that best fit the mean experimental performance for each subject was used to generate the underlying response

distributions for the model. The aggregate of the generated response distributions was then used to test the fit of the complete lapse model.

Attentional control model. The attentional control model was similar to the complete lapse procedure with the exception that the higher order distribution was beta-distributed. Like a Bernoulli distribution, a beta distribution has values bound between zero and one. However, the beta distribution contains many graded outcomes between zero and one. In the attentional control model, one represents maximal attentional engagement and zero represents minimal attentional disengagement. We used a beta distribution since we could model the graded probability of maximally allocating WM resources on each trial. The α parameter of the beta distribution was parametrically varied from zero to six in steps of .01, while the β parameter was constrained to one. Thus, only a single parameter was varied to shift the relative proportion of more-engaged and less-engaged trials. For each cycle of α values, we randomly generated a value from a beta distribution of $[\alpha,1]$ for each trial. On each trial of the simulation, one value was pulled from the beta distribution (e.g., 0.5), multiplied by the maximum capacity parameter (e.g., 0.5 x 3 = 1.5), and rounded to the nearest discrete outcome (e.g., 2). As in the complete lapse model, for each subject in the data set, the α value that best-fit the subject's observed data was selected. These best-fit α values were then used to generate a distribution of responses predicted by the model, which were then used to test the fit of the model by calculating the root mean square error (RMSE).

Results

Mean whole-report performance corresponds with change detection capacity

The task and results from Experiment 2-1a are shown in Figure 2-1. For both tasks, WM performance increases with set-size, reaching a stable plateau around 3 to 4 items. Averaged

across set-sizes, the mean change detection capacity estimate (Cowan's K) was 2.62 ($SD = 0.72$), and mean number of items correct on the whole-report task was 2.91 ($SD = 0.51$). We ran separate repeated measures ANOVAs to verify the change in performance across set-sizes for change detection and for whole-report performance. Change detection performance was significantly different across set-sizes, $F(2.48, 96.68) = 18.13$, $p < .01$. Here and for other cases in which Mauchly's test indicated a violation of the assumption of sphericity, we report Greenhouse-Geisser corrected values. We also ran planned simple contrasts (comparing all smaller set-sizes to the largest set-size) to check for a plateau in performance; we found that the only significant contrast was between set-size 2 and set-size 6, $F(1, 39) = 18.13$, $p < .01$. The comparisons between set-size 6 and the other set-sizes (3, 4, and 5) were not significant, all comparisons $p > .20$, suggesting that performance reached a plateau at set-size 3. Whole-report performance was also significantly across different across set-sizes, $F(1, 39) = 89.54$, $p < .01$. We again ran planned simple contrasts comparing all lower set-sizes to set-size 6. We found that set-size 6 performance was significantly higher than set-size 2, $F(1, 39) = 119.49$, $p < .01$, and set-size 3, $F(1, 39) = 9.34$, $p < .01$, but not for other set-sizes ($p > .30$), suggesting that performance reached a plateau around set-size 4.

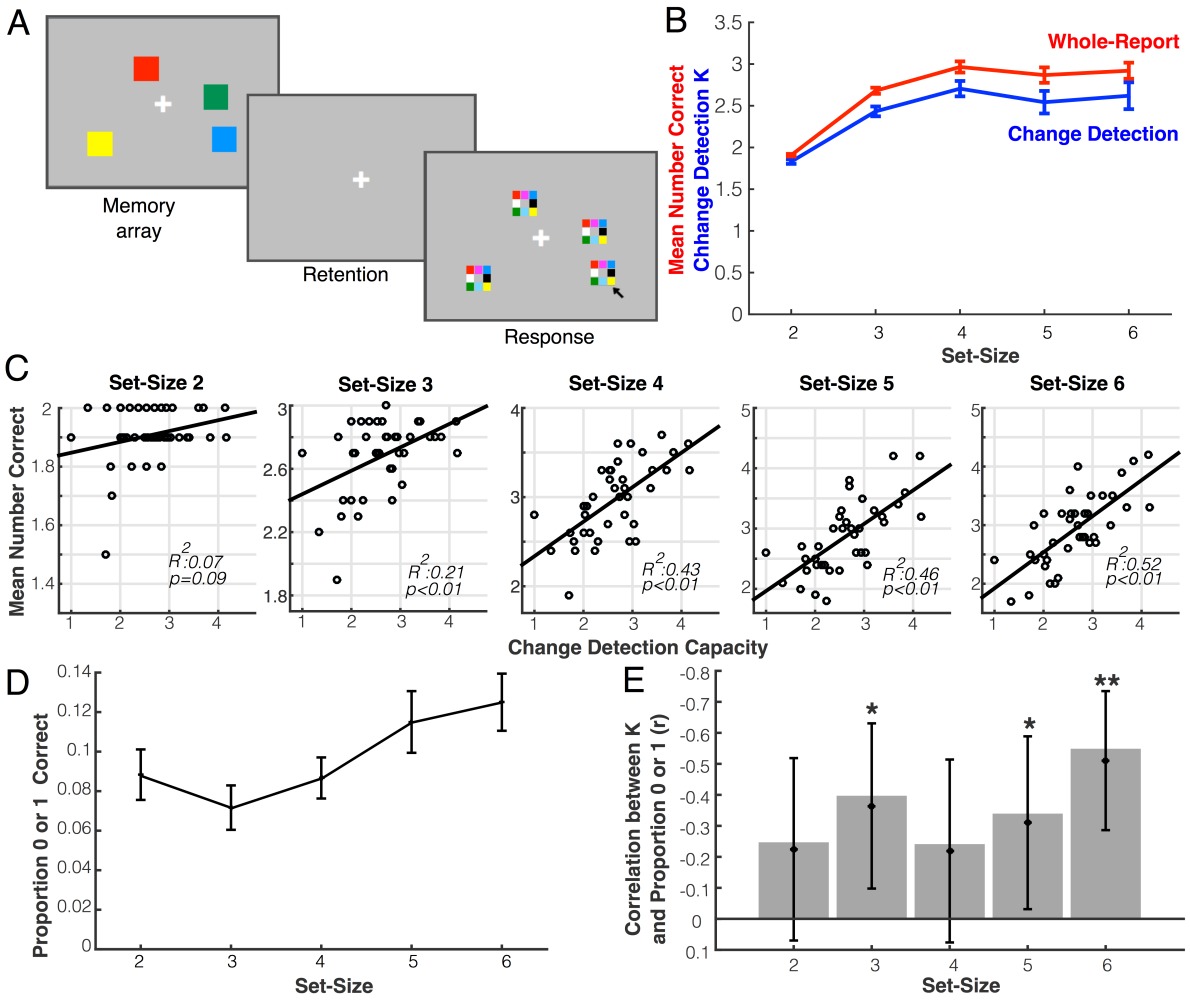

Figure 2-1 Results from Experiment 2-1a
(a) Illustration of the task design and stimuli in Experiment 2-1a. (b) Overall performance changes in a similar manner for change detection (*blue*) and whole-report (*red*) across set-sizes. (c) The correlation between mean whole-report performance and change detection capacity at each set-size. The proportion of performance failures in whole-report (0 or 1 correct) increases across set-size (d) and explains more variance in capacity across set-size (e). Error bars represent Standard Error of the Mean (SEM; b & d) or 95% confidence intervals (e).

In addition to examining how within-task performance changes with load, we can also examine if the relationship between whole-report performance and change detection K is consistent across loads. For each whole-report set-size, we calculated the correlation with a typical measure of change detection capacity (set-sizes 4, 5, and 6) and the mean number of items correctly reported for each set-size. The relationship between change detection K and

whole-report performance across set-sizes is shown in Figure 2-1b. At the lowest set-size (2 items) the relationship between the two measures was non-significant; the relationship became significant at set-size 3 and increased in strength as subjects became more over-loaded with items (Figure 2-1c). However, the lack of correlation between change detection capacity and whole-report performance at set-size two could be due to ceiling effects; most subjects were nearly perfect at responding to the set-size 2 trials. In sum, whole-report and change detection estimates of average working memory performance were strongly related to each other. Next, we investigated whether specific performance outcomes (e.g., 0 items correct) change across working memory loads and whether performance outcomes also predict individual differences in WM capacity.

The relationship between performance failures and K is set-size dependent

To assess performance failures, we measured the proportion of trials in which a given number of items was correctly identified on each trial. We defined performance failures as trials in which subjects scored 0 or 1 items correct out of 6 possible items, since a simulation of guessing yielded 0 or 1 correct on 85% of trials (see simulation results, Figure 2-3a). In particular, we were interested in quantifying whether the proportion of extreme performance failures was constant or variable across set-sizes. As illustrated in Figure 2-1c, we found that the proportion of performance failures increased across set-sizes. Previous lapse models have assumed that the rate of performance failures is constant across set-size. Instead, a repeated measures ANOVA revealed that there was a significant difference in performance failures between set-sizes, $F(3.33, 130.02) = 5.16$, $p = .001$, with performance failures increasing across set-size (Figure 2-1d). We ran planned simple contrasts to test how the rate of performance failures at earlier set-sizes compared to the rate at the highest set-size (6 items). We found that

the only non-significant contrast was for the rates at set-size 5 and set-size 6, $F(1, 39) = .44$, $p = .51$. All other set-sizes had failure rates lower than set-size 6, minimum difference $p < .03$. Additionally, the relationship between change detection capacity and set-size 6 performance failures was the strongest, $r = -0.55$, $p < 0.01$, 95% CI [-0.73, -0.29], (Figure 2-1e).

This set of correlations revealed that, while performance failures occurred at all set-sizes, they were consistently diagnostic of individual differences in capacity only for supra-capacity set-sizes. Thus, though all participants perform very poorly on a subset of trials, low capacity individuals display much greater proportions of poor performance trials. Furthermore, this difference between high- and low- capacity subjects emerges only for supracapacity arrays, supporting an attentional control model over a lapse model. Given these findings, we next examined performance for a task where subjects repeatedly performed set-size 6 trials; this allowed for a more precise characterization in performance distributions and how performance may change over time.

Fluctuations in whole-report performance predict change detection capacity

In Experiment 2-1b, a new sample of subjects completed 300 trials of set-size 6, allowing us to examine fluctuations in task performance that are independent of trial-by-trial variability in task difficulty. The mean change detection K was 2.90 ($SD = 0.98$) and mean whole-report accuracy was 2.87 ($SD = 0.49$). Again, we found a strong positive relationship between change detection K and overall whole-report performance, $r = .55$, $p < .01$ 95% CI [0.30, 0.73] (Figure 2-2a).

We initially examined individual differences in a coarse manner by splitting subjects into 3 groups based on their change detection performance (Figure 2-2b). Subjects in the low K group had change detection scores more than one standard deviation below the mean K score. Likewise

subjects in the high K group had change detection scores more than one standard deviation above the mean K score. All other subjects were placed in the middle K group. As can be seen by the distributions, the prevalence of performance failures (0, 1, or 2 correct) increased across these performance groups. A simple lapse model of performance would predict bimodality, with large proportions of trials at 0 and at typical capacity values. Here, the low-capacity group had more complete failures (0 or 1 correct) than the high-capacity group, but neither group showed bimodality. Instead, this difference in performance failures was part of an overall shift in performance distributions that is more consistent with an attentional control model of individual differences. We observed a downward shift in performance distributions for low K subjects, and an upward shift for high K subjects. However, while using an extreme groups split is useful for summarizing gross differences between groups, such an approach often creates statistical problems (Conway et al., 2005). As such, we also wanted to examine the fine-grained, correlational differences between subjects' distributions.

To better visualize how subtle distributional shifts correspond with visual working memory capacity, we plotted distributions from all subjects in a single heat map and sorted the rows by change detection K (Figure 2-2c). The heat map depicts the distribution of number correct for all subjects. Each horizontal line represents a different subject; the lines are arranged along the y-axis according to each subject's change detection score. The x-axis represents different trial outcomes, between 0 and 6 correct. Intensity of the heat scale represents proportion of trials that all into each score category. Here, we can see a strong, dark band at 3 items, indicating that most subjects had a larger proportion of trials in which they scored 3 correct. Again, none of the subjects showed a pattern consistent with a bimodal lapse model of performance. To quantify the relationship between K and performance outcomes, we calculated the correlation values for each

level of performance (Figure 2-2d). As suggested by the consistent band at three items correct, the correlation between number of 3-correct trials and K was non-significant (r = .02, p = .89, 95% CI [-.27, .31]). On the other hand, the number of correct objects in categories above (r = .54, p < .001, 95% CI [.66, .88]) and below (r = -.52, p < .001, 95% CI [-.71, -.26]) this mode strongly predicted K.

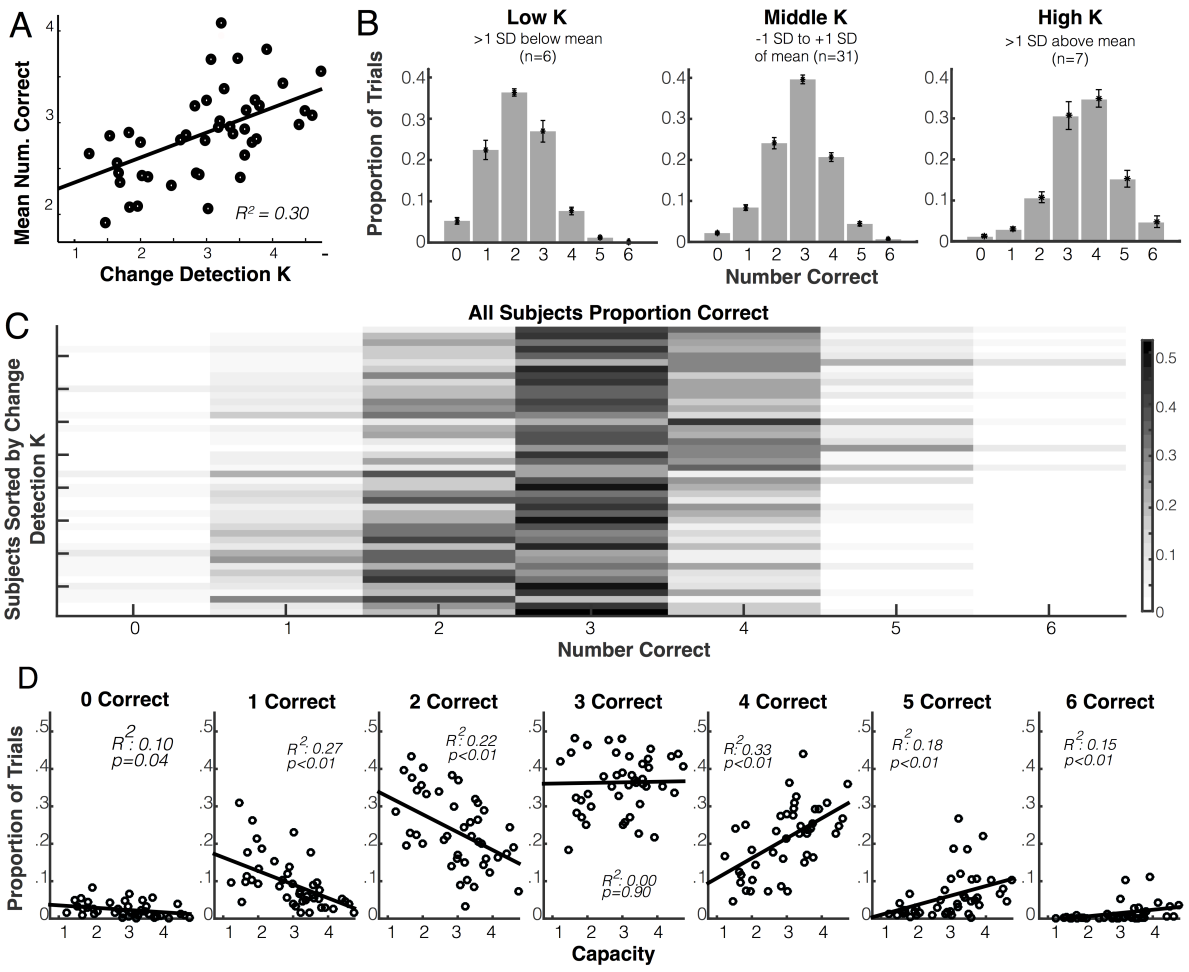

Figure 2-2 Whole-report performance distributions for Experiment 2-1b
(a) Correlation between mean whole-report performance and change detection capacity in Experiment 2-1b. (b) Performance distributions for subjects split into extreme groups by their change detection score. Error bars represent the SEM. (c) Performance distributions for all subjects in Experiment 2-1b. Each column represents the performance outcome (number of items correct for a given trial) and each row represents a subject (sorted by capacity). Differences between subjects are best characterized as a subtle upward or downward shift of the performance distribution (with a central tendency at 3 for most subjects). (d) Performance outcomes correlated with change detection estimates of capacity for all levels of performance except for 3 correct.

The greater number of trials in Experiment 2-1b allowed us to use simulation approaches to test potential models of individual differences in performance. In particular, we were interested in whether performance failures are better characterized as an all or nothing engagement in the task (lapses) or as varying degrees of engagement in the task (attentional control). For the

purposes of modeling, the size of our maximum resource pool is described in items, but this is only because of the nature of our behavioral assay. Our task necessarily involves a discrete outcome (a discrete number of items correct out of six), and all of our modeling efforts rely upon such estimates of trial-by-trial performance. However, this Experiment 2-cannot speak to any of the current debates about the nature of the limit on working memory. We make no claims that discrete slots models are preferred over continuous resources models on the basis of these data. Instead, we are interested in tracking variations in a subject's typical performance level (the deployment of their resources, whatever the underlying structure of the resources). Indeed, it has been proposed that both discrete and continuous models of working memory resources could plausibly implement trial-by-trial variability in the availability of resources (van den Berg et al., 2014).

Lapse performance and guessing inflation. Before testing lapse and attentional control models, we first characterized the performance outcomes for guessing among 6 items and the effects of guessing inflation. Guessing without replacement for set-size six yielded 0 or 1 correct 85% of the time for nine possible colors (Figure 2-3a). This means that on the remaining 15 to 17% of guessing trials, a subject may have reported two or more items correct, even when they truly had zero items in mind. Given our simulation results, we used 0 or 1 correct as a conservative definition of a performance failures for all analyses. We also ran a simulation to account for the effects of guessing inflation given knowledge about three items and guessing without replacement. The guessing inflation distribution had a strong peak at three but also included a large number of trials where subjects got 4 or 5 items correct by chance (45%; Figure 2-3b). Thus, guessing inflation can account for a large percentage of above-3 trials for a subject

with a true maximum capacity of 3, and it is important to control for this effect in any simulation model.

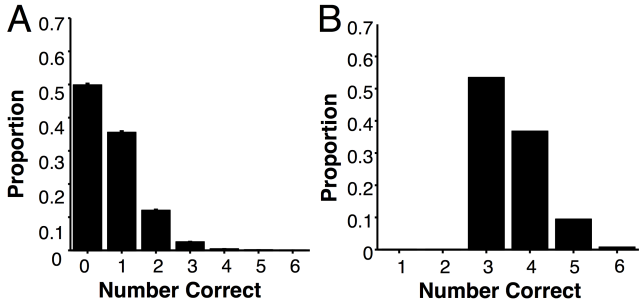

Figure 2-3 Monte Carlo simulation of lapse performance and guessing inflation
(a) Results from a simulation of guessing without replacement from 9 colors over 6 objects. (b) Results from a simulation of guessing inflation if subjects got 3 items correct and then guessed without replacement from the remaining colors for 3 objects.

Testing lapse and attentional control models. The lapse model specified that lapse events occur as a total loss of attentional engagement, while the attentional control model specified that lapse events occur as a variable loss of attentional engagement. Using a maximum capacity parameter of three items correct, we could successfully recreate the observed mean whole-report performance with both the lapse model ($R^2 = .95, p < .01$, 95% CI [0.90, 0.97]) and the graded attentional control model ($R^2 = .82, p < .01$, 95% CI [0.70, 0.90]. However, only the attentional control model reliably fit the observed distribution of responses (Mean RMSE = 0.14, SEM = .01). The failure of the complete lapse model (Mean RMSE = 0.29, SEM = .01) was due to an overestimation of the proportion of trials in which individuals correctly reported zero or three, and an underestimation of the proportion of trials in which individuals reported two items, thus producing a bimodal distribution of expected responses. This difference in model fit was significantly different, $t(43) = -8.5, p = 4.5 \times 10^{-11}$, 95% CI [-.11, -.17]. Additionally, we found that the attentional control model was better than the complete lapse model for the group of low-K subjects, $t(5) = -14.1, p = 1.64 \times 10^{-5}$, 95% CI [-.33, -.24], and middle-K subjects, $t(30) = -$

8.34, p = 1.3 x 10^{-9}, 95% CI [-.17, -.10], though neither model was reliably better for the high-K subjects, $t(6)$ = -.51, p = .31, 95% CI [-.07, .03], (Figure 2-4).

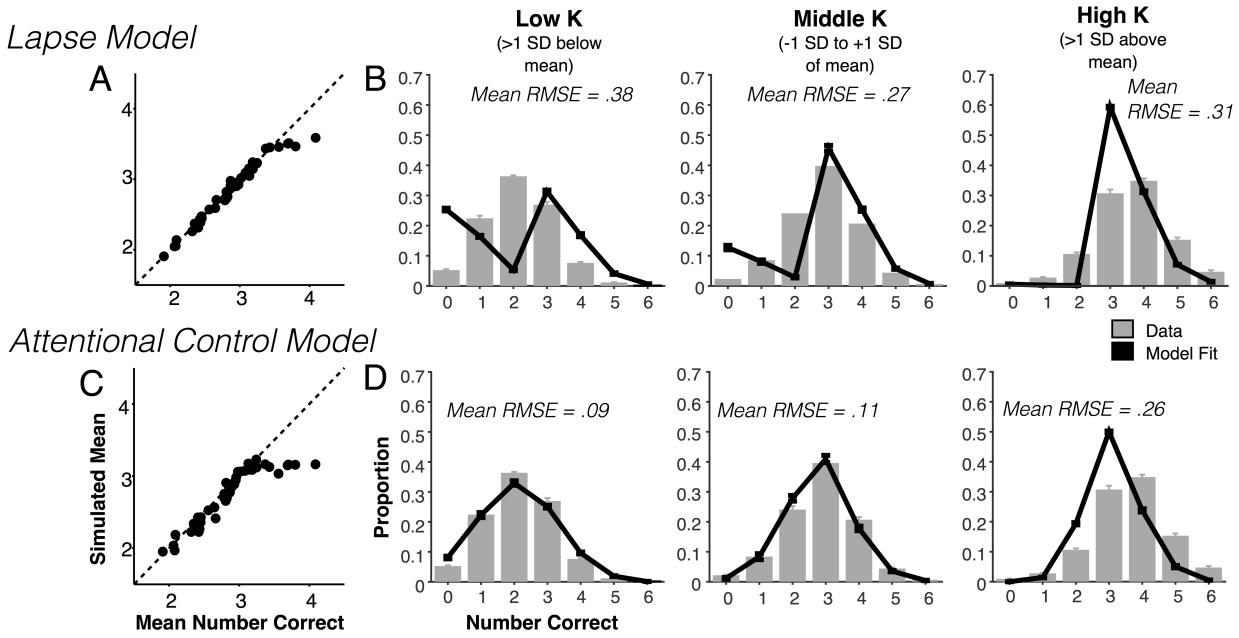

Figure 2-4 Monte Carlo simulation results for lapse and attentional control models of performance fluctuations
(a) The simulated mean number correct from the lapse model as a function of the actual mean number correct. (b) Data (*gray bars)* and lapse model fits (*black lines)* from the extreme groups split of subjects. (c) The simulated mean number correct from the attentional control model as a function of the actual mean number correct. (d) Data (*gray bars*) and attentional control model fits (*black lines*) shown over the extreme groups split of subjects.

Next, we tested whether changing the maximum capacity parameter in the attentional control model would result in improved fits. In particular, we noticed that the high-K group was fit poorly by a maximum capacity of three items (for either model). Additionally, we thought that subjects in the low-K group, with an overall downward shifted distribution, may be better fit by a maximum capacity of two items. We found that changing the maximum capacity parameter from three to two items did not improve fits for the low-K group (ΔRMSE = -.04, SEM = .04, p = .30) and significantly decreased fits for the other two groups (ΔRMSE = -.21, SEM = .02, p = 6 x 10^{-10}, and ΔRMSE = -.23, SEM = .03, p = .001, respectively). Next, we tested if increasing the

maximum capacity parameter would increase fits for the high-K group. This model reproduced means for subjects, $R^2 = 0.96$, $p < .01$, %95 CI [0.92, 0.98]. Increasing the maximum capacity parameter from three to four improved fits for the high-K group (ΔRMSE = +.19, SEM = .04, $p = .005$) but significantly decreased fits for the low-K (ΔRMSE = -.09, SEM = .02, $p = .005$) and middle-K (ΔRMSE = -.03, SEM = .01, $p = .04$) groups. Finally, we tested a model in which there was no capacity maximum (maximum capacity parameter is six items). We found that this model accurately reproduced mean performance values, $R^2 = 0.91$, $p<.01$, 95% CI [0.83, 0.95], but resulted in poor fits that were significantly worse than the proposed limited capacity model (ΔRMSE = -.16, SEM = .02, $t(43) = -8.66$, $p = 3 \times 10^{-11}$, 95% CI [-.19, -.12]).

Performance fluctuations occur consistently over time

One alternative explanation for the increased prevalence of performance failures for low K subjects is that they took much longer to learn the task and had an inflated level of performance failures in early blocks. Similarly, the relationship between performance failures and capacity could also be explained if low K subjects "give up" at the end of the experiment. To test these time-based explanations of performance, we examined the occurrence of performance failures over time.

In Figure 2-5a, we illustrate performance for all subjects and all trials across time. Each row represents a subject, and rows are sorted by overall whole-report performance. As such, low-performing subjects are on the bottom of the graph and high-performing subjects are on the top. Black tick marks represent extreme performance failures (0 or 1 correct), gray tick marks represent below typical modal performance (2 correct) and white tick marks represent full engagement trials (3 or more correct). Red vertical lines represent block breaks. Because there was only condition in Experiment 2-2 (set-size six), there is a unique opportunity to look at

variations in working memory performance, on a single trial basis, that are not due to condition or set-size difficulty. Instead, fluctuations in performance represent fluctuations in task engagement. As can be seen by the heterogeneous appearance of black and white tick marks, performance failures were scattered throughout the experiment for both high and low K subjects.

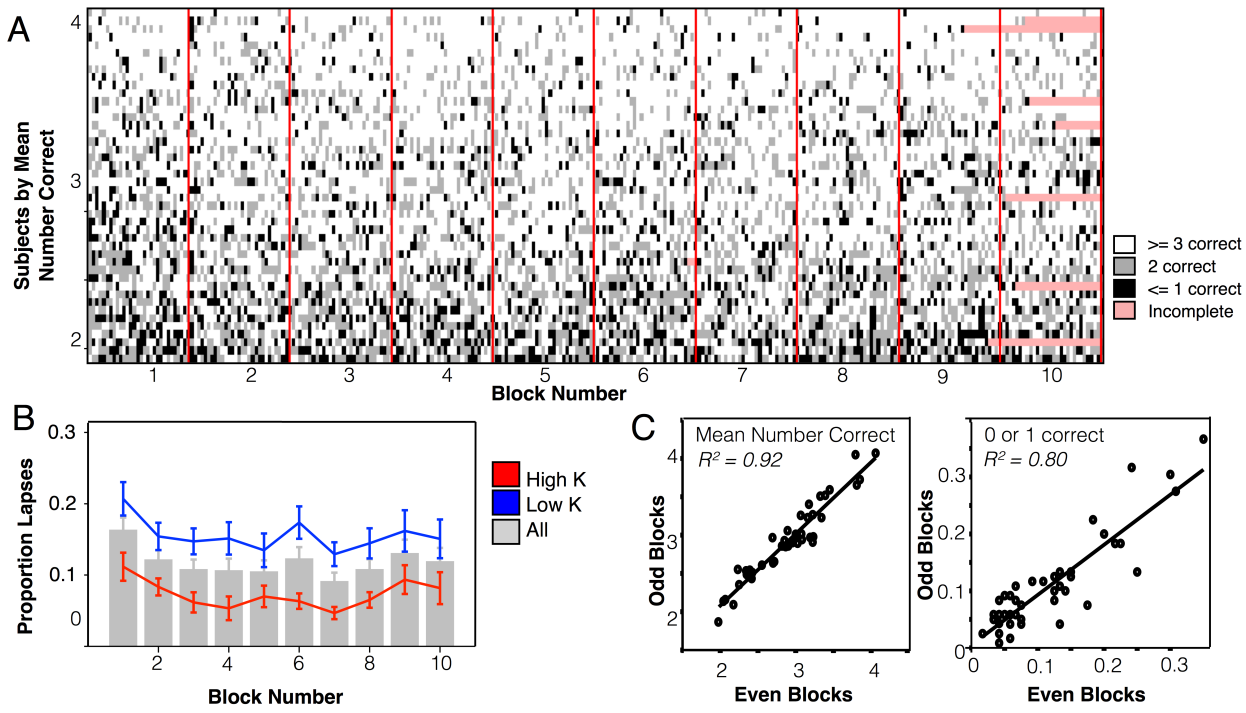

Figure 2-5 Performance fluctuations over time in Experiment 2-1b
(a) Data from all subjects and all trials is shown over time. Each subject is a row (sorted by whole-report performance) and the horizontal axis corresponds to trials over time. Tick mark color corresponds to the performance outcome; red lines delineate block breaks. (b) A summary of the prevalence of performance failures (0 or 1 items correct) over blocks. Gray bars show the mean performance level; red and blue lines illustrate a median split of subjects. (c) Reliability of mean number correct (left) and prevalence of performance failures (right) for even versus odd blocks.

To quantify this trend, we examined the frequency of performance failures over time (Figure 2-5b). We performed a median split based on change detection K and ran a two-way repeated measures ANOVA with run (9 blocks, within-subjects) and group (two-groups, between-subjects) as the main factors. We plotted average accuracy for all subjects and blocks

but we restricted the ANOVA to blocks 1 – 9 because some subjects did not reach block 10 (those with pink tick marks in Figure 2-5a). We found a significant main effect of group, $F(1, 42) = 132.5$, $p < .001$, and of block, $F(8, 336) = 4.96$, $p = < .001$, but no interaction of block and group, $F(8, 336) = .677$, $p = .71$. The lack of interaction indicates that that while there were block effects for both groups of subjects the difference between high- and low-capacity subjects was consistent over time. Post-hoc pairwise comparisons revealed that only the first block was significantly different from any of the other blocks. Lapses were significantly higher in the first block than in blocks 2 through 5 ($p < .015$) or in blocks 7 through 8 ($p < .02$). Notably, subjects did not complete a set of practice trials before beginning the experiment, so a learning effect would be expected. In sum, an increased lapse rate in the first block explains the difference in lapses over time, but a differential ability to learn the task does not explain the consistent difference in lapses between subjects.

As an additional check on the reliability of whole-report performance within a session, we performed a between-blocks reliability analysis. The reliability of average performance and performance failures is shown in Figure 2-5c as the correlation between performance on even blocks and odd blocks. Since not all subjects finished 10 blocks, we restricted the analysis to blocks 2 – 9 and included all subjects. We used Cronbach's alpha to quantify reliability. Performance failures were highly reliable for both mean performance (Cronbach's $\alpha = .97$, blocks 2 - 9) and for performance failures (Cronbach's $\alpha = .93$, blocks 2 - 9). In sum, we found that differences in the preponderance of performance failures are not due to learning differences between high- and low-capacity subjects. Furthermore, whole-report estimates of working memory performance are highly reliable throughout the session.

Performance fluctuations are not due to artifacts or sensory encoding differences

In Experiment 2-2 we recorded EOG and EEG while subjects completed the whole-report task. First, we wanted to examine the role of simple task non-compliance on the rate of poor performance trials (fewer than 3 items correct). We instructed subjects to keep their eyes on a fixation cross and not to close their eyes during the presentation of the memory array; if subjects did not follow these instructions (e.g., moving their eyes away from the screen, blinking during the presentation array), then they may show degraded performance. To test this possibility, we measured the occurrence of poor performance trials before and after excluding trials containing ocular artifacts. We found that the overall ratio of low-performance trials was not significantly changed after removing ocular artifacts, $t(22) = 1.2$, $p = .24$ (Figure 2-6a), and the relationship between poor performance trials and change detection K was preserved, $R^2 = .23$, $p = .022$. Furthermore, the relationship between the percentage of artifact-rejected trials and K was non-significant, $R^2 < .01$, p = .85. Thus, for the vast majority of lapse trials, the subject's eyes were indeed open and pointed toward the screen. The percentage of trials that subjects are negligent of eye movement instructions does not predict their overall performance level.

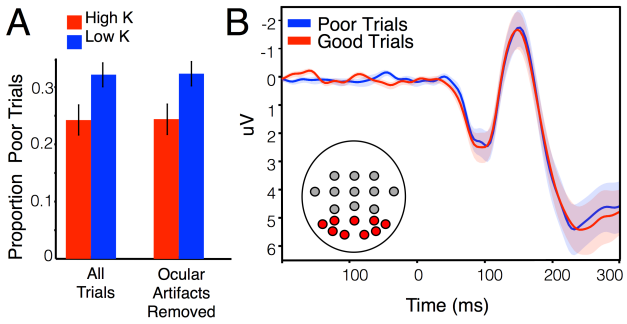

Figure 2-6 Performance fluctuations do not relate to task compliance or sensory encoding (a) The prevalence of poor trials (<= 2 correct) before and after removing ocular artifacts. (b) The P1/N1 visual-evoked response as a function of performance outcome (good versus poor trials).

Next, we wanted to examine whether poor performance trials were associated with decreased sensory processing (Weissman, Roberts, Visscher, & Woldorff, 2006). We measured the mean amplitudes of the visual-evoked P1/N1 ERP components and found that there were no significant differences in amplitude between poor trials (<3 correct) and good trials (>3 correct) for the P1(70 - 120 ms), $t(22) = .59$, $p = .56$, or the N1(130 - 170 ms), $t(22) = .49$, $p = .63$ (Figure 2-6b). Together, we found no evidence that early sensory processing of external stimuli was impaired during poor performance trials.

Performance fluctuations are related to frontal theta and posterior alpha power

Finally, we tested whether hypothesized neural correlates of attentional engagement and working memory storage predicted whole-report performance. In particular, we focused on spectral power in the theta and alpha frequency bands. Frontal theta power has been shown to relate to measures of executive control (Cavanagh & Frank, 2014), working memory load (Deiber et al., 2007; Jensen & Tesche, 2002), successful retrieval (Hsieh & Ranganath, 2014), and manipulation of information in WM (Itthipuripat, Wessel, & Aron, 2013). Decreased alpha power has been shown to relate to attention and semantic memory performance (Klimesch, 1999; Klimesch, Doppelmayr, Schimke, & Ripper, 1997), and with task difficulty and WM load (Gevins, 2000; Gevins, Smith, McEvoy, & Yu, 1997; Stipacek, Grabner, Neuper, Fink, & Neubauer, 2003). Here, we wanted to examine whether, given the same difficult task load, we could predict trial-to-trial fluctuations in subjects' success using markers shown to be related to overall task difficulty.

First, we wanted to test the simple hypothesis that average frontal theta or posterior alpha power (calculated over the entire trial period) correlates with overall working memory ability. While some studies have found evidence for a relationship between individual differences in

spectral power and cognitive ability (Gevins, 2000; Zakrzewska & Brzezicka, 2014), such relationships are more often unreported in the literature. In our study, we found no relationship of mean posterior alpha power with either change detection (r = .07, p = .88) or whole-report performance (r = -.24, p = .26). We found that frontal theta power correlated positively with change detection capacity (r = .45, p = .03), consistent with previous measures of theta power and working memory. However, theta power did not correlate with whole-report performance (r = .28, p = .19), despite the strong relationship between change detection and working memory performance for this sample (r = .61, p < .01). In sum, we conclude that overall frontal theta power may relate to working memory performance in some settings, but either (1) it does not consistently relate to all working memory tasks or (2) the power in this sample is insufficient to show the relationship consistently across tasks.

Despite inconsistent findings for a strong between-subjects relationship of WM ability and theta power, we are in a good position to examine within-subject, trial-by-trial predictors of performance. In the current experiment, subjects completed many trials of the same condition; as such, we can examine oscillatory predictors of trial-by-trial performance. In Figure 2-7, each subplot represents the time-frequency plot for a single electrode during Experiment 2-2. Each heat map was created by subtracting the time-frequency plot for poor trials (<=2 correct) from good trials (>=4 correct). A difference in theta was especially prominent at frontal channels, while a difference in alpha power was prominent at the posterior channels. Collapsing across frontal electrodes (Figure 2-8a), we found that decreased frontal theta power (4-7Hz) predicted poor-performance trials (Figure 2-8b) starting about midway through the pre-trial period and sustaining throughout the retention interval. To summarize significance for key trial events, we examined theta power for early and late pre-trial periods, the encoding period, and the retention

interval (Figure 2-8b). Time-point zero corresponds to stimulus onset. We found that theta did not predict performance in the first half of the pre-trial period (-1400 to -700 ms), $t(22) = .845$, $p = .80$, but began to predict performance in the second half of the pre-trial period, $t(22) = 2.05$, $p = .026$. Theta power continued to predict working memory performance throughout the encoding period (0 to 250 ms), $t(22) = 3.06$, $p < .01$, and during the retention interval (250 ms to 1550 ms), $t(22) = 2.80$, $p < .01$.[1]

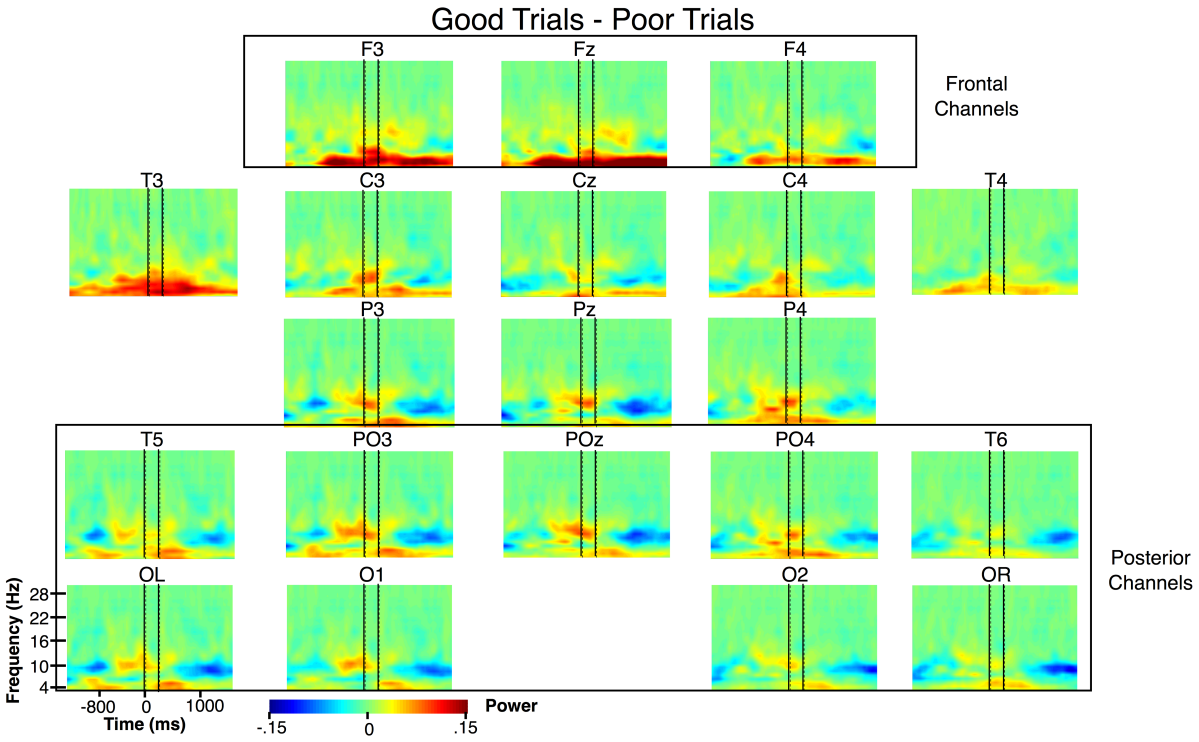

Figure 2-7 Spectrogram for good trials minus poor trials for each electrode
Each spectrogram represents spectral power at all frequencies from 4 to 30 Hz for each of the electrode sites measured during Experiment 2-2.

[1] P-values are not corrected for multiple comparisons; the Bonferroni-corrected threshold for four comparisons is p = .013. The pre-trial effect (p = .026) does not survive this more conservative thresholding. If we choose only a single time window (-700 to 0 ms), we can check for this effect at all electrodes of interest (see Figure 2-7): Fz (p = .026), F3 (p = .037), F4 (p = .1), T3 (p = .03).

Collapsing across posterior channels (Figure 2-8c), we found that decreased alpha power (8 - 12Hz) was associated with higher performance, starting near the end of the retention interval (around 800 ms). We again determined alpha's ability to predict trial performance during key trial periods (Figure 2-8d). Unlike frontal theta, posterior alpha power did not predict performance in the pre-trial period (-1400 to 0 ms), $t(22) = .62$, $p = .73$, or in the encoding period (0 to 250 ms), $t(22) = 1.16$, $p = .13$. However, decreased alpha power started to predict better performance during the retention interval (250 ms to 1550 ms), $t(22) = 2.50$, $p = .01$. This effect was driven by the second half of the retention interval. Alpha power was no different for good and poor trials in the first half of the retention interval (250 ms to 950 ms), $t(22) = 1.2$, $p = .12$, but was significantly modulated in the second half of the retention interval, $t(22) = 2.57$, $p < .01$, perhaps indicating that subjects were less likely to drop items from memory toward the end of successful trials. This finding is consistent with previous work showing that greater alpha-band desynchronization is associated with increasing cognitive load.

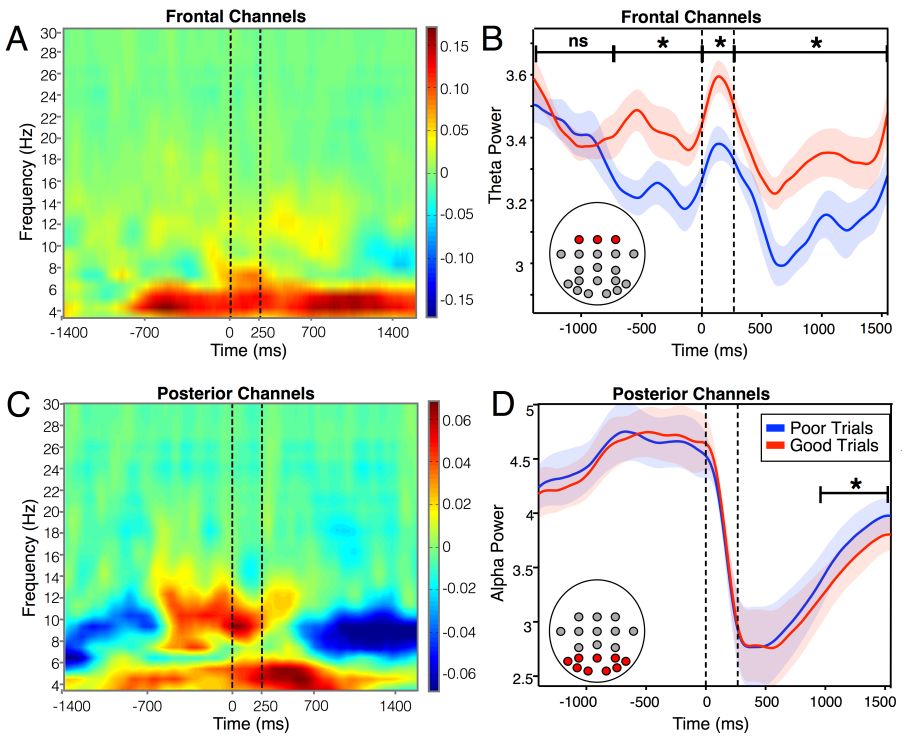

Figure 2-8 Theta and alpha power as a function of whole-report performance
(a) Spectral power for good trials minus poor trials at frontal electrode sites. (b) Mean theta power (4 to 7 Hz) as a function of trial performance. (c) Spectral power for good trials minus poor trials at posterior electrode sites. (d) Mean alpha power (8 to 12 Hz) as a function of trial performance.

Discussion

We investigated two plausible models for how within-subject variance in working memory performance within a session gives rise to individual differences in WM capacity. To do this, we developed and validated a novel whole-report task that provides a trial-by-trial estimate of working memory successes and failures. Examining several criteria, we found evidence against a coarse lapse model and substantial positive evidence in favor of an attentional control model. First, we found that failures rates increased as the task-load increased. This result is inconsistent with lapse models that assume that such failures should be equivalent irrespective of task demands (Rouder et al., 2008, 2011), but is consistent with an attentional control account (Fukuda, Woodman, et al., 2015). Second, we observed that while low capacity individuals had

more complete performance failures than high capacity individuals, this higher failure rate was better explained as a downward shift of performance distributions for the low capacity individuals. We found no evidence for a bimodal distribution of performance, as would be predicted by a lapse model. This clear distinction between the two hypothesized performance distributions was confirmed by using simulations to test the lapse and attentional control models. We found that though both models could simulate mean performance levels, only the attentional control model produced the distribution of outcomes observed in the data. Finally, our neural data indicated that performance failures are associated with changes in oscillatory signatures of attentional control: decreased frontal theta power and increased posterior alpha power.

In addition to providing evidence for the attentional control model, our data allowed us to test several potential mechanisms that potentially underlie performance failures. One plausible explanation of individual differences in the rate of performance failures is the effect of time within the session; individual differences could simply be due to how quickly subjects learn the task (e.g., many failures at beginning) or become fatigued (e.g., many failures at the end). This hypothesis predicts that individual differences are disproportionately explained by performance at the very beginning or end of the experiment. Contrary to this hypothesis we found that performance failures occurred consistently throughout the experiment and that the differences between high and low capacity individuals were stable throughout the entire experiment. A second explanation for performance failures is simple task non-compliance. Though we instruct subjects to keep their eyes open and focused on the central fixation dot, subjects blinking or moving their eyes away from fixation could result in poor performance. However, we found no relationship between the rate of EOG artifacts and performance failures, and that these failure rates are preserved even after excluding trials with ocular artifacts. A third reason for poor

performance is that there is insufficient sensory encoding of the memory array items. However, we found no difference in the visual-evoked response (P1 and N1) to the memory array items between poor and good performance trials. Together, these results suggest that working memory performance failures are not simply due to slow learning, fatigue, ocular artifacts, or poor sensory encoding.

In contrast to the above results, our neural measures provided positive evidence that performance failures are related to well-known oscillatory markers of attentional control mechanisms: frontal theta and posterior alpha. We found that mean frontal theta power was higher on successful trials than on failure trials and that this difference began a few hundred milliseconds before memory encoding and persisted throughout the retention period. There is a substantial literature relating frontal theta power to attentional control and memory success and the current work is broadly consistent with these findings. In addition, the finding that theta power can distinguish between working memory successes and failures before the trial has begun suggests that it reflects preparatory mechanisms of attentional control that need to be engaged in advance to adequately perform these tasks (Leber, Turk-Browne, & Chun, 2008). Likewise, towards the end of the retention interval, we observed that increased posterior alpha power predicted performance failures. This may reflect an inability to sustain the alpha desynchronization that is necessary for ongoing memory storage. Together, our findings fit well within the literature showing that increased theta and decreased alpha at encoding predict successful memory performance (Klimesch, 1999; Klimesch et al., 1997; Stipacek et al., 2003). Our findings are also consistent with hypotheses about (1) the strong relationship between working memory and attention (Chun, Golomb, & Turk-Browne, 2011; Engle & Kane, 2003; Unsworth et al., 2014), (2) the trial-to-trial variability of attention (Esterman, Noonan,

Rosenberg, & DeGutis, 2013; Esterman, Rosenberg, & Noonan, 2014) and (3) the importance of pre-frontal networks in sustaining attentional control (Giesbrecht, Woldorff, Song, & Mangun, 2003; Liesefeld, Liesefeld, & Zimmer, 2014).

Individual differences in visual working memory capacity are robust, stable, and predictive of fluid intelligence and have been proposed to be due to variations in attentional control (Engle, Kane, et al., 1999; Fukuda, Vogel, et al., 2010; Unsworth et al., 2014). However, a compelling alternative model proposes that these differences are instead due to how frequently the individual is completely disengaged from the task at hand. Our current results reject such a coarse lapse model, and suggest that *graded* fluctuations in attentional control from trial to trial within a session drive the individual differences in capacity that are observed in traditional aggregate measures of performance. Failed attentional control on a trial would be expected to produce a wide swath of processing errors such as insufficient individuation, poor resolution, item position swapping, and retrieval failures. The present work suggests that the ability to prevent such failures by consistently engaging attentional control mechanisms during challenging tasks is a central component of an individual's cognitive ability.

CHAPTER 3. CLEAR EVIDENCE FOR ITEM LIMITS IN VISUAL WORKING MEMORY

Introduction

Working memory (WM) is an online memory system where information is maintained in the service of ongoing cognitive tasks. Although there is a broad consensus that WM resources are sharply limited, there has been sustained debate about the precise nature of these limits. On the one hand, discrete-resource models argue that only a handful of items can be maintained at one time, such that some items fail to be stored when the number of memoranda exceeds the observer's capacity (Awh et al., 2007; Cowan, 2001; Rouder et al., 2008; Zhang & Luck, 2008). On the other hand, continuous resource models argue that WM storage depends on a central pool of resources that can be divided across an unlimited number of items (Bays & Husain, 2008; van den Berg et al., 2012; Wilken & Ma, 2004).

Of course, it has long been known that memory performance declines as the number of memoranda increases in a WM task. For example, Luck and Vogel (1997) varied the number of simple colors in a change detection task that required subjects to detect whether one of the memoranda had changed between two sequential presentations of a sample display. They found that while performance was near ceiling for set sizes up to three items, accuracy declined quickly as set size increased beyond that point. This empirical pattern is well described by a model in which subjects store 3-4 items in memory and then fail to store additional items. However, the same data can be accounted for by a continuous resource model that presumes storage of all items, but with declining precision as the number memoranda increases (Wilken & Ma, 2004). According to continuous resource accounts, increased errors with larger set sizes are caused by

insufficient mnemonic precision rather than by storage failures (but for a critique of this account see Nosofsky & Donkin, 2016b). Thus, a crux issue in this literature has been to distinguish whether performance declines with displays containing more than a handful of items are due to storage failures or sharp reductions in mnemonic precision.

In this context, Zhang and Luck (2008) offered a major step forward with an analytic approach that provides separate estimates of the probability of storage and the quality of the stored representations. They employed a continuous recall WM task in which subjects were cued to recall the precise color of an item from a display with varying numbers of memoranda. Their key insight was that if subjects failed to store a subset of the items, there should be two qualitatively distinct types of responses within a single distribution of response errors. If subjects had stored the probed item in memory, responses should be centered on the correct color, with a declining frequency of responses as the distance from the correct answer increased. But if subjects had failed to store the probed item, then responses should be random with respect to the correct answer, producing a uniform distribution of answers across the entire space of possible colors. Indeed, their data revealed that the aggregate response error distribution was well described as a weighted average of target-related and guessing responses. Thus, Zhang and Luck (2008) provided some of the first positive evidence that working memory performance reflects a combination of target-related and guessing responses.

Subsequent work, however, has argued that the empirical pattern reported by Zhang and Luck (2008) can be explained by continuous resource models that presume storage of all items in every display (van den Berg et al., 2014, 2012). A key feature of these models has been the assumption that precision in visual WM may vary substantially; thus, while some items may be represented precisely, other representations in memory may contain little information about the

target item. Using this assumption, van den Berg et al. (2012) showed that they could account for the full distribution of errors – including apparent guessing – and that their model outperformed the one proposed by Zhang and Luck (2008). Indeed, converging evidence from numerous studies has left little doubt that precision varies across items in these tasks (e.g., Fougnie, Suchow, & Alvarez, 2012). That said, the question of whether precision is variable is logically separate from the question of whether observers ever fail to store items in these procedures. To examine the specific reasons why one model might achieve a superior fit over another, it is necessary to explore how distinct modeling decisions influence the outcome of the competition. Embracing this perspective, van den Berg, Awh and Ma (2014) carried out a factorial comparison of WM models in which the presence of items limits and the variability of precision were independently assessed. Although this analysis provided clear evidence that mnemonic precision varies across items and trials, the data were not decisive regarding the issue of whether working memory is subject to an item limit. There was a numerical advantage for models that endorsed item limits, but it was not large enough to draw strong conclusions. Thus, the critical question of whether item limits in visual working memory elicit guessing behavior remains unresolved.

Here, we report data that offer stronger traction regarding this fundamental question about the nature of limits in working memory. Much previous work has focused on explaining variance within aggregate response-error distributions (i.e., the shape of the response distribution and how it changes across set sizes). Here, we chose a different route. Rather than developing a new model that might explain a small amount of additional variance in "traditional" partial report datasets, we developed a new experimental paradigm in which subjects recalled—in any order that they wished—the precise color (Experiment 3-1a) or orientation (Experiment 3-1b) of every

item in the display. This procedure has the key benefit of measuring the quality of all simultaneously remembered items, and it yields the clear prediction that if there are no item limits, then there should be measurable information across all responses. To anticipate the results, this whole report procedure provided rich information about the quality of all items within a given trial as well as subjects' metaknowledge of variations in quality. Observers consistently reported the most precisely remembered items first, yielding monotonic declines in information about the recalled item with each successive response. Critically, for the plurality of subjects, the final three responses made were best modeled by the parameter-free uniform distribution that indicates guessing. In additional analyses and experiments, we showed that subjective guess ratings tracked mixture model guessing parameter (Experiments 3-1 & 3-2), that output interference could not explain our estimates of capacity (Experiment 3-2), and that making subjective guess ratings did not drive our evidence for guessing (Experiment 3-3). Finally, we used simulations to question a key claim of the variable precision model – that representations used by this model all contain measurable information. Previously, others have suggested that the variable precision model may mimic guess responses with ultra-low precision representations (Nosofsky & Donkin, 2016b; Sewell, Lilburn, & Smith, 2014). Here, we advanced these claims by showing that variable precision models that eschew guessing posit a high prevalence of memories that are indistinguishable from guesses. Moreover, the frequency of these putative representations precisely tracked the estimated rate of guessing in models that acknowledge item limits.

In sum, there has been a longstanding debate over whether there is any limit in the number of items that can be stored in working memory. Our findings provide compelling evidence that working memory is indeed subject to item limits, disconfirming a range of prior

models that deny guessing entirely or posit an item limit that varies from trial to trial without any hard limit in the total number of items that can be stored (e.g., Sims et al., 2012; van den Berg et al., 2014). Instead, our results point toward a model where each individual has a capacity ceiling (e.g., 3 items), but they frequently under-achieve their maximum capacity, likely due to fluctuations in attentional control (Adam, Mance, Fukuda, & Vogel, 2015).

Experiment 3-1

Materials & Methods

Experiment 3-1a: Color memoranda

Subjects. 22 subjects from the University of Oregon completed Experiment 3-1a for payment ($8 per hour) or class credit. All participants had normal color vision and normal or corrected-to-normal visual acuity, and all gave informed consent according to procedures approved by the University of Oregon institutional review board.

Stimuli. Stimuli were generated in MATLAB (The MathWorks, Inc., Natick, MA, www.mathworks.com) using the Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997). Stimuli were presented on a 17-inch flat cathode ray tube computer screen (60 Hz refresh rate) on a Dell Optiplex GX520 computer running Windows XP and viewed from distance of approximately 57 cm. A chin rest was not used, so all visual angle calculations are approximate. A gray background (RGB = 128 128 128) and a white fixation square subtending .25 by .25 degrees of visual angle appeared in all displays. In Experiment 3-1a, subjects were asked to remember the precise color of squares in the memory array, each subtending 1.7 by 1.7 degrees. Colors for memory stimuli were chosen randomly from a set of 360 colors taken from a CIE $L^*a^*b^*$ color space centered at $L = 54$, $a = 18$ and $b = -8$. Note, colors were generated in CIE L*a*b* space, but they were likely rendered with additional variability; monitors were not

calibrated to render true-equiluminant colors. Others have compared calibrated versus uncalibrated monitors and found consistent results (Bae, Olkkonen, Allred, & Flombaum, 2015; Bae, Olkkonen, Allred, Wilson, & Flombaum, 2014). Uncalibrated monitors may exaggerate the amount of variability in precision across different colors in the color wheel.

Spatial positions for colored stimuli were equidistant from each other on the circumference of circle with a radius of 3.75 degrees around the fixation point. At test, a placeholder array of dark gray squares (RGB = 120 120 120) and a color wheel (radius = 11.9 degrees of visual angle) appeared, and the mouse cursor was set to the fixation point. The color wheel rotated on each trial so that subjects could not use spatial locations of colors to plan responses. When a placeholder item was selected for response, its color changed to light gray (RGB = 145 145 145). During response selection, the selected square changed colors to match the color that matched the current position of the mouse cursor; after response selection the selected square returned to dark gray.

Procedures. The session for Experiment 3-1a lasted approximately 1.5 hours and participants completed 5 blocks of 99 trials (99 trials per set size). On each trial, a memory array with one, two, three, four or six colored squares appeared briefly (150 ms) followed by a blank retention interval (1000, 1200, or 1300 ms). Retention interval length was jittered to add variability and were collapsed for analyses. There were not enough trials to separately fit model parameters to each retention interval duration. After the retention interval, the test display appeared, containing the color wheel and placeholder squares. The response order was determined freely by the subject. Subjects chose the first response item by clicking on one of the dark gray placeholders; the dark gray placeholder turned light gray, indicating that this square was chosen for response. The mouse cursor was set back to the fixation point to avoid any

response bias based on spatial proximity of a chosen square to a section of the color wheel. Then, the subject selected a color on the color wheel that best matched their memory of the square. During each response, participants were instructed to indicate their confidence with a mouse click. Participants were instructed to use one mouse button to make their response if they felt they had "any information about the item in mind," and to use the other mouse button to indicate when they felt they "had no information about the item in mind." After responding to the first item, the mouse was set back to the fixation point. Subjects repeated the item selection and color selection procedure until they had responded to all of the items. After finishing all responses, the placeholder squares disappeared and the next trial began after a blank inter-trial interval (1300 ms).

Experiment 3-1b: Orientation memoranda

Subjects. 23 subjects from the University of Chicago participated in Experiment 3-1b for payment ($10 per hour). Three subjects participated in the study but were not included in the final sample because they left the session early (2 subjects) or stayed for the full session but failed to complete all trials (1 subject); this left a total of 20 subjects for analysis. All participants had normal or corrected-to-normal visual acuity, and all gave informed consent according to procedures approved by the University of Chicago institutional review board.

Stimuli. Stimuli were presented on a 24-inch LCD computer screen (BenQ XL2430T; 120 Hz refresh rate) on a Dell Optiplex 9020 computer running Windows 7 and viewed from distance of approximately 70 cm. A chin rest was not used, so all visual angle calculations are approximate. In Experiment 3-1b, subjects were asked to remember the precise orientation of a line embedded in a circle (see inset of Figure 3-1). The stimuli in each memory array had a radius of approximately .9 degrees of visual angle, and their orientations were randomly chosen

from 360 degrees of orientation space. Stimuli were presented on a gray background (RGB = 85 85 85), and a white fixation circle appeared in all displays (diameter = .14 degrees). Dark gray placeholder circles (RGB = 45 45 45) were used during the test period. Spatial positions of the discs were randomly placed within a box 6.6 degrees of visual angle to the left and right of fixation and 6.1 degrees above and below fixation, with the stipulation that there must be a minimum distance of 1.25 items between items' centers.

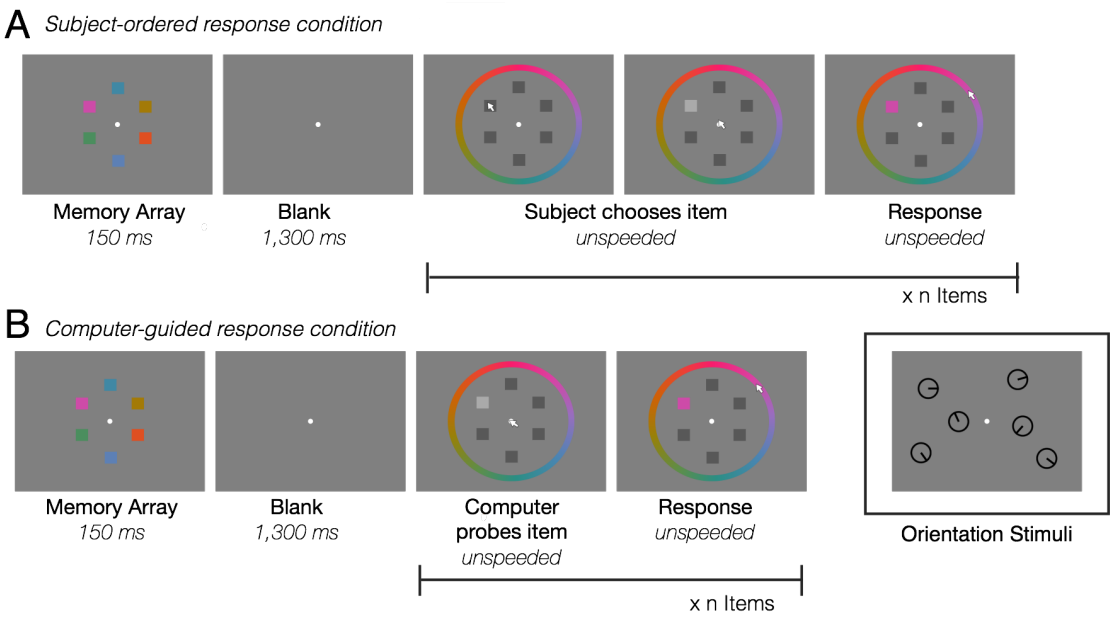

Figure 3-1 Task design
Panel A depicts the order of events in Experiment 3-1. Panel B depicts the order of events in Experiment 3-2. Color stimuli were used in Experiment 3-1a and -2a, the orientation stimuli (inset) were used in Experiment 3-1b and -2b.

Procedures. Procedures in Experiment 3-1b (orientation) were similar to Experiment 3-1a. Each experimental session lasted approximately 2.5 hours and subjects completed 20 blocks of 50 trials (200 trials per set size). On each trial in Experiment 3-1b, a memory array with one, two, three, four or six black orientation stimuli appeared briefly (200 ms) followed by a blank retention interval (1000 ms). After the retention interval, the test display appeared, containing

dark gray placeholder circles. The response order was determined freely by the subject. Subjects responded to each item by first clicking on the location of the item they wished to report. After selecting the item, the gray placeholder circle was replaced by a black test circle. Participants then used the mouse to click the edge of the circle at the location of the remembered orientation. During each response, participants were instructed to indicate their confidence with a mouse click. Participants were instructed to use one mouse button to make their response if they felt they had "any information about the item in mind," and to use the other mouse button to indicate when they felt they "had no information about the item in mind." After responding to all items, the next trial began with the blank inter-trial interval (1000 ms).

Fitting Response Error Distributions

Model-free circular statistics. To quantify change in mnemonic quality without committing to contentious model assumptions, we used a circular statistics measure to quantify mnemonic performance. Circular statistics were calculated using "CircStat", a circular statistics toolbox for MATLAB (Berens, 2009); for more information on statistics in circular space, see Zar (2010). Response error distributions are centered around 0 degrees of error in a circular normal space (e.g., -180 degrees is the same as 180 degrees of error). The direction and variability of data-points in a given response error distribution can be described by the mean ("circ_mean.m"; Zar (2010) pp. 612) and the mean resultant vector length (MRVL; r, "circ_r.m"; Zar (2010) pp. 615) of the distribution. The circular mean indicates the average direction of data-points (e.g., the central tendency), whereas MRVL indicates the variability of data-points. MRVL varies from 0 (indicating a complete absence of information about the target) to 1 (indicating perfect information about the target).

Model fitting. In addition to using circular statistics, for some analyses we fit a mixture model to response error distributions (Zhang & Luck, 2008). Although there is debate regarding the mixture model's assumption that subjects sometimes guess, this analytic approach allowed us to compare subjects' self-reports of guessing to the frequency of guess responses postulated by these mixture models. Thus, response errors for each response at each set size were fit for each subject with a mixture model using a maximum likelihood estimation procedure in the MemToolbox package (Suchow, Brady, Fougnie, & Alvarez, 2013, www.memtoolbox.org). The mixture model fits response errors with a mixture of two distributions, a Von Mises distribution (circular normal) and a uniform distribution. The contribution of the uniform distribution to the response error distribution is described by the guessing parameter, *g*, and the dispersion of the Von Mises component is described by the precision parameter, *sd*.[1] The guessing parameter ranges from 0 to 1, and quantifies the "proportion of guesses" in the distribution, whereas the precision parameter is given in degrees (higher values indicate poorer precision). Using MemToolbox, we could also compare BIC values for the mixture model to an all-guessing model (uniform distribution).

Results & Discussion

In Experiment 3-1 we tested for the presence of guessing in response errors. A model of working memory that includes item limits and guessing predicts uniformly distributed responses for items that subjects cannot recall. That is, some responses are based on positive knowledge of the target, and some responses are completely random. To preview our results, we found that a uniform distribution (with zero free parameters) was the best-fitting distribution for a substantial

[1] The dispersion of the Von Mises probability density function used in the model is specified with κ (concentration), this is later converted to *sd* for interpretation.

portion of responses and that subjects consistently reported that these responses were guesses. Data from Experiment 3-1 and all following experiments are available on our Open Science Framework project page, https://osf.io/kjpnk/.

Change in quality across set sizes and responses

To examine the effects of set size, we collapsed across all responses to create the response error distribution for each set size. Consistent with the previous literature, we observed a systematic decline in MRVL across set sizes (Figure 3-2) in Experiment 3-1a (Color), $F(2.45, 51.38) = 431.9$, $p < .001$, $\eta_p^2 = .95$, and Experiment 3-1b (Orientation), $F(1.45, 27.62) = 675.58$, $p < .001$, $\eta_p^2 = .97$[2]. As the memory load increased, the distribution of response errors became increasingly diffuse, indicating that on average less information was stored about each memorandum. Importantly, overall MRVL values and the slope of their decline across set sizes was similar to that observed in past studies using single probe procedures. Below, a more detailed analysis of Experiment 3-2 findings will provide further evidence for this observation. Thus, requiring the report of all items did not induce qualitative changes in performance at the aggregate level. Nevertheless, as the following results will show, the whole report procedure provided some important new insights about the distribution of mnemonic performance across the items within a trial.

[2] Greenhouse-Geisser corrected values are reported wherever the assumption of sphericity is violated.

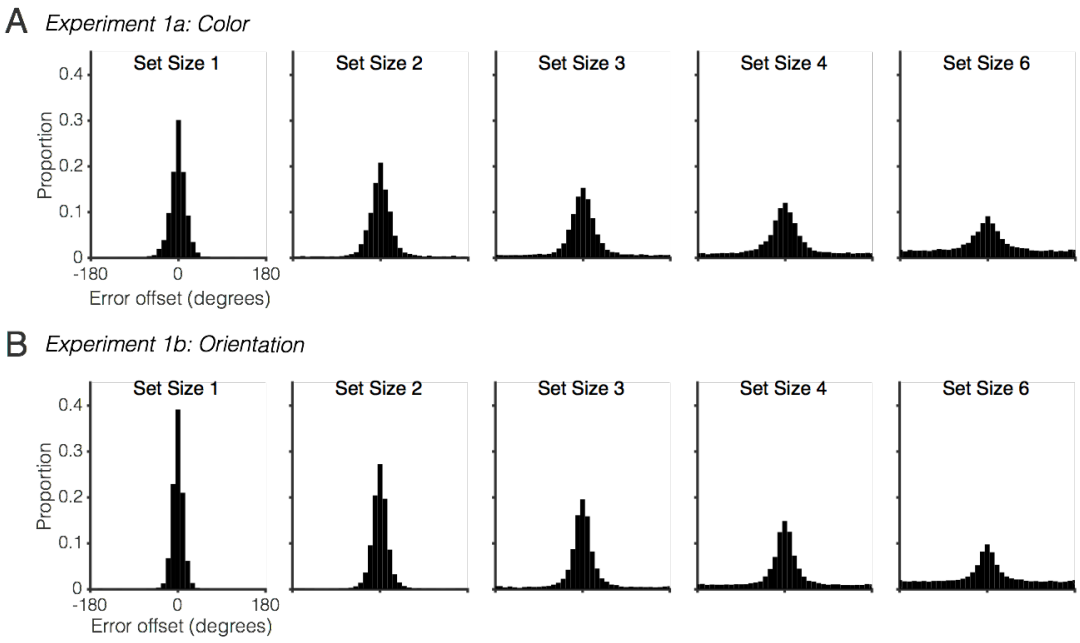

Figure 3-2 Aggregate data for Experiment 3-1
Collapsing across all responses within a set size, we see a typical decline in precision with increasing memory load.

Next, we examined the effect of response order within each set size. Recall that subjects were free to recall each item in whatever order they chose. As Figures 3-3 and 3-4 show, there was a strong tendency for subjects to report the best remembered items first. There was a sharp drop in MRVL from early to late responses within a trial. A repeated-measures ANOVA for each set size with a within-subject factor for response order showed a clear main effect of response order for all set sizes in Experiment 3-1a (Table 1) and Experiment 3-1b (Table 2). Planned contrasts comparing each earlier response to the last response revealed that performance declined monotonically with response order, except for between the fifth and sixth responses in the set size six condition where MRVL values were hovering just above the floor of 0 (p = .96 for Experiment 3-1a, p = .13 for Experiment 3-1b). MRVL decreased by on average .18 per response in Experiment 3-1a and .19 per response in Experiment 3-1b. From the first to the last response in set size 6, estimates of mean resultant vector length decreased by .74 in Experiment 3-1a and

.84 in Experiment 3-1b (maximum possible difference of 1.0), reaching minimum values of between 0.05 and 0.10. To summarize, when subjects were allowed to report all items in any order that they chose, we observed a consistent drop in target information with each additional response. This suggests that subjects reported the best remembered items first, and that they had strong meta-knowledge of which items were remembered the best. Below, we will show that subjects' explicit reports of guessing were also quite accurate in tracking their mnemonic performance.

Table 3-1 Change in Mean Resultant Vector Length across responses in Experiment 3-1a

Set Size	*df1*	*df2*	*F*	*p*	η_p^2
2	1	21	42.4	<.001	.67
3	2	42	129.5	<.001	.86
4	2.2	45.4	223.8	<.001	.91
6	2.8	58.4	295.2	<.001	.93

Table 3-2 Change in Mean Resultant Vector Length across responses in Experiment 3-1b

Set Size	*df1*	*df2*	*F*	*p*	η_p^2
2	1	19	66.1	<.001	.78
3	1.2	21.9	116.1	<.001	.86
4	1.5	28.3	217.3	<.001	.92
6	2.1	40.2	389.5	<.001	.95

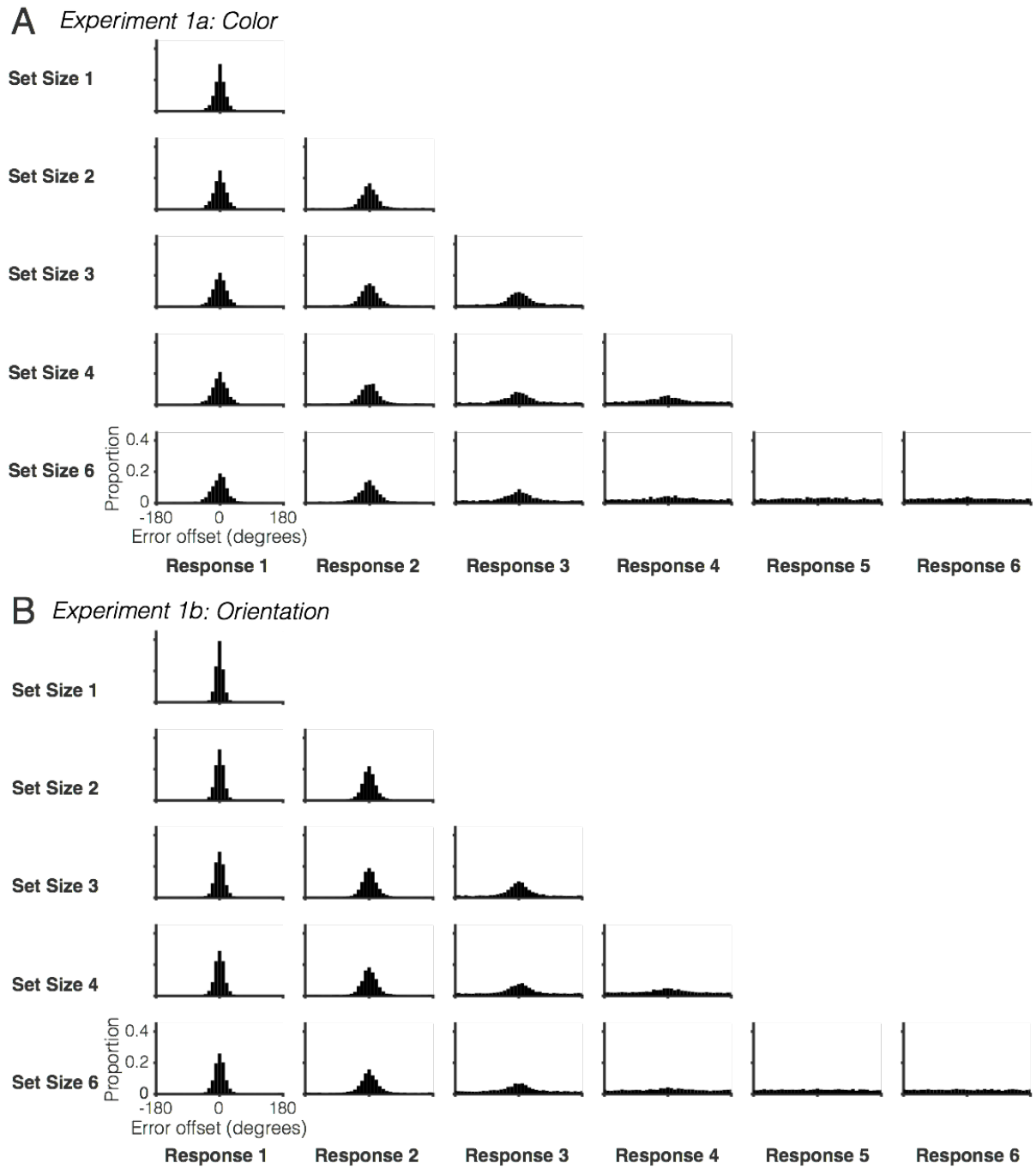

Figure 3-3 Quality of subject-ordered representations covaries with response order
All set sizes and responses are shown for (A) Experiment 3-1a and (B) Experiment 3-1b.

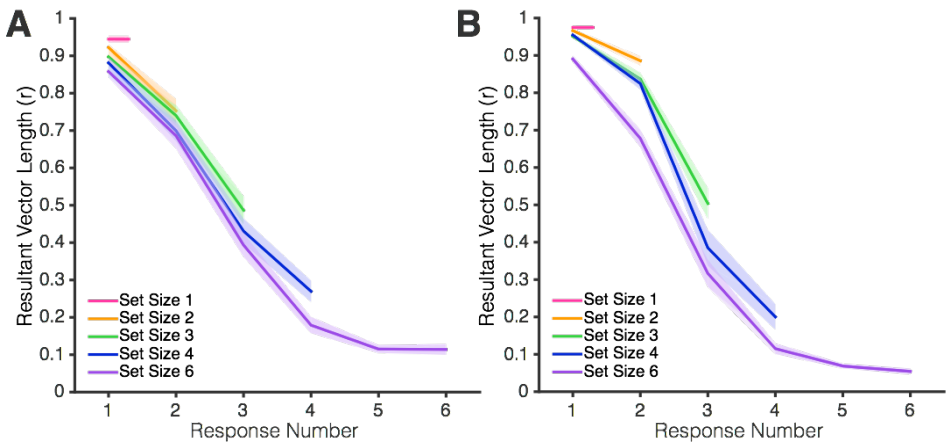

Figure 3-4 Mean Resultant Vector Length for subject-ordered responses
MRVL across responses in Experiment 3-1a (A) and -1b (B). Shaded error bars represent 1 standard error of the mean (SEM).

Evidence for guessing in the set size 6 condition

Thus far, the findings from the whole report procedure have mirrored those found in past studies using single-item probes. The shape of the aggregate error distribution, as well as the decline in MRVL values with increasing set size, fell in line with past studies. Going beyond this, testing all items in each trial provided new insight into the full range of memory performance across each item in the sample array. Notably, there appeared to be uniform error distributions – the distribution predicted when subjects fail to store the relevant item – for the fifth and sixth items recalled in the set size 6 condition (see Figure 3-3). This may be a critical observation. While past work has shown that it is difficult to adjudicate between models that endorse guessing and those that propose high prevalence of low fidelity memories, the whole report procedure appears to provide clear evidence of guessing behavior in the set size 6 condition.

To objectively test the hypothesis that subjects guessed during later responses, we used a BIC approach to compare the fit of a uniform distribution with the fit of a mixture model[3] (the simplest implementation of some guessing plus some information) using MemToolbox (Suchow et al., 2013). Any reliable central tendency in the error distribution should yield a lower BIC value for such a mixture model than for the uniform model. If, however, guessing alone is adequate to explain performance, then the BIC value should be lower for the uniform distribution. For each participant, we operationally defined a capacity limit by counting the number of empirically-defined uniform distributions during Set Size 6 trials. Our logic for this operational definition of capacity is as follows: participants tended to report items in order of decreasing quality, and we would expect that a participant who stored 3 items would first report these items before making any guess responses. Thus, they would have 3 non-uniform responses and 3 uniform responses. Participants who maximally stored different numbers of items would be expected to have different numbers of uniform distributions. Note, this analysis relies on the assumption that participants had robust metaknowledge that enabled them to report items in declining order of quality. Thus, if an individual had poor meta-awareness of stored items (i.e., they sometimes reported their best items toward the end of the trial), then this operational

[3] We thought it unlikely that models with more free parameters than a mixture model would beat a zero-parameter uniform distribution, but we nevertheless ran a second version of the model competition in which we included 5 models available in the MemToolbox: (1) Uniform (2) Standard Mixture Model (3) Variable Precision (VP) Model, with Gaussian higher-order distribution of precision values (4) VP, with Gamma higher-order distribution (5) VP, with Gamma higher-order distribution plus a guessing parameter. Critically, MemToolbox implementations of these models allow model fitting to individual distributions without specifying set-size; this was important because we had no strong a priori assumptions about what "set size" each response distribution should be equivalent to. There was no difference in the results for either Experiment 3-1a or -1b. The uniform model won for the same individual distributions as when just comparing between the uniform and the mixture models.

definition of capacity would over-estimate that individual's capacity limit. This analysis revealed that all subjects had between one and four responses best described by a uniform distribution, and that these were the last items reported in the trial (Figure 3-5). In Experiment 3-1a (Color), the average number of uniform responses was 2.64 (SD = .73), and in Experiment 3-1b (Orientation) the average number was 2.80 (SD = .77). Supplementary analyses revealed that the uniformity of later responses cannot be explained by a sudden increase in the tendency to report the value of the wrong item[4]. However, some guess responses may be due to retrieval failures rather than to lack of storage (Harlow & Donaldson, 2013; Harlow & Yonelinas, 2016). We also found that circular statistics approaches to test for uniformity yield similar conclusions[5]. In sum, we obtained clear positive evidence that the final responses in the set size 6 condition were guesses.

[4] We ran a repeated-measures ANOVA on the Set Size 6 swap rates with Response Number as a within-subjects factor. Planned contrasts compared each earlier response (1-5) to the final response (6). In Experiment 3-1a, there was no main effect of response order on swaps, $F(3.5, 73.4) = 1.2$, $p = .327$. That is, swaps were no more likely to occur for the later, uniform responses than they were to occur for early responses. In Experiment 3-1b, there was a significant main effect of response order on swaps, $F(3.2, 60.3) = 4.41$, $p = .006$. However, planned contrasts revealed that only responses 1 and 2 were different from response 6 ($p < .02$), whereas responses 3 through 5 were indistinguishable from the final response ($p > .10$). In sum, the uniformity observed for responses 5 and 6 cannot be explained by a sudden increase in swap errors.

[5] As an alternative to the BIC model comparison between uniform and mixture model distributions, we used a modified Rayleigh Test for uniformity. In Experiment 3-1, this analysis revealed an average of 2.36 uniform distributions per subject in Experiment 3-1a (SD = .79, ranging from1 to 4) and 2.85 uniform distributions per subject in Experiment 3-1b (SD = .59, ranging from 2 to 4).

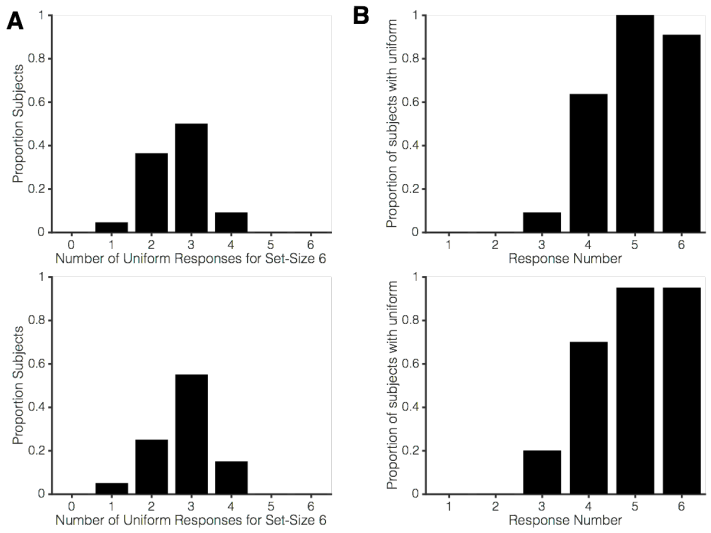

Figure 3-5 Number of uniform responses for Set Size 6 in Experiments 3-1a and 3-1b
(A) Histogram of subjects' total number of Set Size 6 responses that were best fit by a uniform distribution. Top: Color Condition; Bottom: Orientation Condition. (B) Number of subjects' responses best fit by uniform distributions as a function of response number. Top: Color Condition; Bottom: Orientation Condition.

Our estimates of individual differences in capacity closely tracked the range found in the literature. A majority of participants had an estimated capacity between 3 and 4 items, and there was some variability on either side of this (Figure 3-5). Thus while every subject in this study showed evidence of an item limit, there was no direct evidence for a common storage limit across observers, in line with past work that has documented individual variations in neural markers of WM storage (e.g., Todd & Marois, 2004, 2005; Unsworth, Fukuda, Awh, & Vogel, 2015; Vogel & Machizawa, 2004) and behavioral measures of performance (e.g., Engle, Kane, et al., 1999; Vogel & Awh, 2008; Z. Xu et al., 2017). Likewise, previous modeling efforts have acknowledged the need to account for individual differences in performance; models allow parameters to vary across individual subjects (e.g., Bays, Catalao, & Husain, 2009; van den Berg et al., 2014; Zhang & Luck, 2008).

Strong correspondence between subjective reports of guessing and the guessing parameter in a mixture model.

Previous work has demonstrated that subjective confidence strongly predicts mnemonic precision (Rademaker, Tredway, & Tong, 2012) and correlates strongly with fluctuations in trial-by-trial performance (Adam & Vogel, 2017; Cowan et al., 2016). Our finding that subjects consistently reported the best-remembered items first also suggests that subjects have strong meta-knowledge regarding the contents of working memory. To provide an objective test of this interpretation, we examined whether or not subjects' self-reports of guessing fell in line with the probability of guessing estimated with a standard mixture model (Zhang & Luck, 2008). A tight correspondence between subjects' claims of guessing and mixture model estimates of guessing would demonstrate that subjects have accurate meta-knowledge and bolster the face validity of the guessing parameter employed in mixture models.

To examine the correspondence between subjective and objective estimates of guess rates, we fit a separate mixture-model to response errors for each response within each set size (16 total model estimates per subject). We also calculated the percentage of subjects' responses that were reported guesses (guess button used) for each of these 16 conditions. Then, we correlated the *g* parameter with the percentage of reported guessing. If there is perfect correspondence between the guessing parameter and subjective reports of guessing, a slope of 1.0 and intercept of 0.0 would be expected for the regression line. The resulting relationship between the model and behavioral guessing was strikingly similar to this idealized prediction (Figure 3-6). The average within-subject correlation coefficient was $r = .94$ (SD = .05, all p-values < .001) in Experiment 3-1a and $r = .93$ (SD = .06, all p-values < .001) in Experiment 3-1b, indicating a tight relationship between the model's estimates of guessing and subjects' own

reports of guessing. Participants were on average slightly over-confident (as indicated by a slope slightly less than one when the model's guessing parameter was plotted on the x-axis). In Experiment 3-1a, participants had an average slope of .80 (*SD* = .29) and intercept of .03 (*SD* = .08). In Experiment 3-1b, participants had an average slope of .83 (SD = .19) and intercept of -.05 (SD = .02). In sum, the striking correspondence between mixture model estimates of guessing and subjective reports of guessing suggests that subjects had excellent metaknowledge. The frequency with which subjects endorsed guessing precisely predicted the height of the uniform distribution estimated with a standard mixture model.

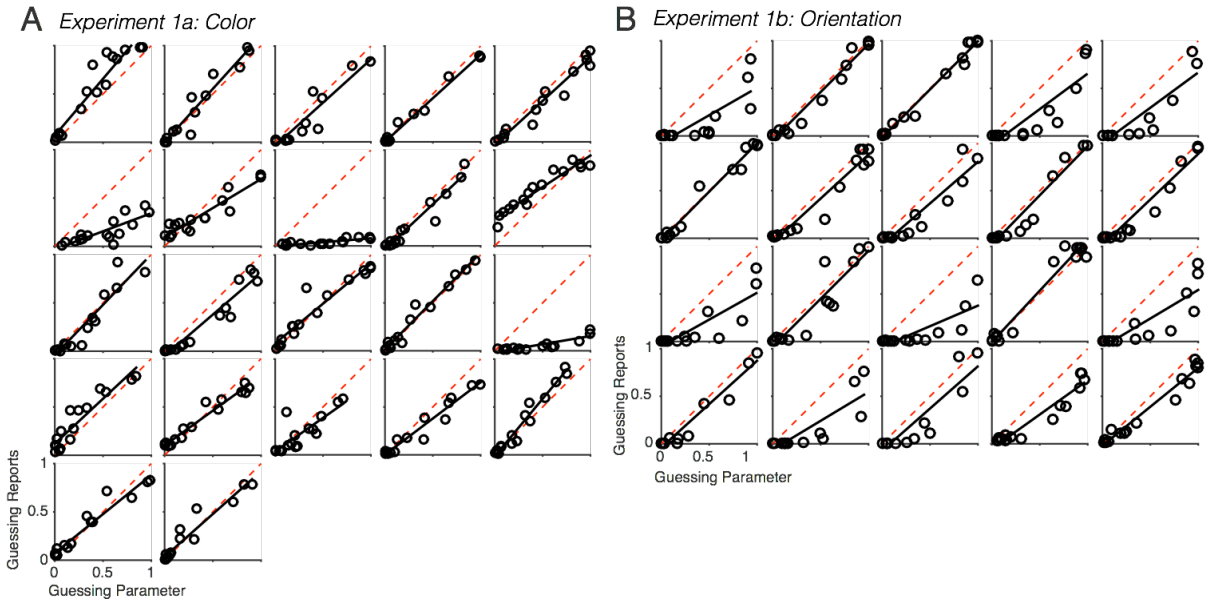

Figure 3-6 The relationship between behavioral guessing and modeled guessing
Each panel shows an individual's correlation between reported guessing and the mixture model *g* parameter for each of 16 conditions (every response made for Set Sizes 1-4 and 6). The red dotted line represents perfect correspondence between behavioral guessing and modeled guessing (slope = 1, intercept = 0).

Output interference as a potential source of guessing behavior

The data from Experiment 3-1 provide compelling evidence that that a substantial portion of subjects' responses were best characterized by a uniform distribution associated with

guessing. Because subjects tended to respond first with the items that they remembered the best, these uniform distributions were observed in the last items that were reported within the whole report procedure. An important alternative explanation, however, is that the decline in memory performance across responses may have been due to output interference. Specifically, we considered whether merely reporting the initial items could have elicited the drop in performance that we saw for the last items reported within each trial. Indeed, output interference has been demonstrated in past studies of working memory (Cowan, Saults, Elliott, & Moreno, 2002). Thus, Experiment 3-2 was designed to measure the strength of output interference in our whole report procedure. Subjects in Experiment 3-2 had to respond to the items in a randomized order specified by the computer. If the uniform responses that we observed in Experiment 3-1 were due to output interference, then we should observe a similar drop in performance across responses in Experiment 3-2. Furthermore, a randomized response-order design allowed us to decouple confidence ratings from response order. In Experiment 3-1, confident responses never occurred at the end of the trial. By randomly selecting the order of report in Experiment 3-2, we had the opportunity to observe whether subjects also guessed when they were making the earliest responses in a trial.

Experiment 3-2

Materials & Methods

Experiment 3-2a: Color memoranda

Subjects. 17 subjects from the University of Oregon completed Experiment 3-2a. All participants had normal color vision and normal or corrected-to-normal visual acuity, and they were compensated with payment ($8/hour) or course credit. All participants gave informed

consent according to procedures approved by the University of Oregon institutional review board.

Procedures. Stimuli and procedures in Experiment 3-2a were identical to Experiment 3-1a except for the order in which responses were collected. On each trial, the response order was determined randomly by the computer on each trial, which will be referred to as a "random response" order. In Experiment 3-2a (color), one of the remembered items turned light gray, indicating that the participant should report the color at that location. The participant used the mouse to click on the color in the color wheel that best matched the memory for the probed square. The response process repeated until subjects had responded to all of the items in the display. Participants were again instructed to use the two mouse buttons to indicate their confidence in each response.

Experiment 3-2b: Orientation memoranda

Subjects. 21 subjects from the University of Chicago completed Experiment 3-2b. One participant began participation, but left the session early. After analyzing the data, one additional participant was excluded for poor performance (> 30% guessing rate for set size 1). This left a total of 19 subjects for data analysis. All participants had normal color vision and normal or corrected-to-normal visual acuity. Participated were compensated with payment ($10/hour) and all gave informed consent according to procedures approved by the University of Chicago institutional review board.

Procedures. Trial events in Experiment 3-2b were identical to Experiment 3-1b except for the order in which responses were collected. At test, the cursor was set on top of one of the remembered items. The participant used the mouse to rotate the probed item to the remembered orientation, and clicked to set the response. After a response was collected, the test item was

replaced by a gray placeholder circle, and a new untested item was probed. The response process repeated until subjects had responded to all of the items in the display. Participants were again instructed to use the two mouse buttons to indicate their confidence in each response.

Results & Discussion

Changes in response quality across set sizes and responses.

The whole report procedure was meant to tap into the same cognitive limits that have been observed in past studies using single-item probes. Thus, an important goal was to determine whether this procedure produced the same kinds of error distributions that have been observed in studies employing single-item probes. In the random response-order condition, participants were probed on all of the items in a randomized fashion. Thus, the first probed response in this condition was similar to a typical partial-report procedure, where only one item is randomly probed and output interference cannot influence the results. We compared overall performance in the free-response order experiments (i.e., combining all responses into one set size level distribution) to performance for the first randomly probed item in the random response-order experiments. Then, we quantified the quality for each set size distribution using MRVL. For color stimuli (Experiments 3-1a and -2a), an ANOVA with within-subjects factor Set Size and between-subjects factor Experiment revealed that there was no main effect of Experiment on MRVL, $F(1,37) = 0.12$, $p = .72$, $\eta_p^2 = .003$. Likewise, there was no main effect of Experiment for orientation stimuli (Experiments 3-1b and -2b), $F(1,37) = 0.17$, $p = .68$, $\eta_p^2 = .005$. These null results suggest that the free-response whole report procedure elicited similar aggregate error distributions as observed in past procedures that have probed only a single item.

To more directly measure the effect of output interference within the random response-order experiments, we examined how response quality changed across responses made within

each set size. A decline in MRVL values across responses in Experiment 3-2 would provide positive evidence for output interference. Indeed, the results revealed some decline in performance across responses. Critically, however, the slope of this decline was far shallower than when subjects chose their own response order in Experiment 3-1 (Figure 3-7 and Figure 3-8). To quantify the decline in mnemonic quality across responses, we again calculated MRVL values for each response within each set size, and we ran a repeated-measures ANOVA for each set size with the within-subjects factor Response. There was a significant main effect of response for all set sizes in Experiment 3-2a (Table 3). In Experiment 3-2b, there was no significant main effect of response for Set Size 2 ($p = .96$) but there was a significant difference for all other set sizes (Table 4). However, while these findings provide some evidence of output interference, there were striking differences from the pattern that was observed in Experiment 3-1. In Experiment 3-1, we observed a monotonic decline in memory quality across all responses, such that aggregate error distributions were completely uniform for the fifth and sixth responses. By contrast, in Experiment 3-2 only the first response was consistently better than the final response. Rather than an accumulation of output interference across each successive response, this empirical pattern suggests an advantage for the first response or two. In Experiment 3-2a, planned contrasts comparing each earlier response to the last response revealed that performance was better for only the first response for Set Sizes 3 and 4 and for the first and second responses for Set Size 6. In Experiment 3-2b, contrasts revealed that only the first response was significantly better than the final response for all set sizes.

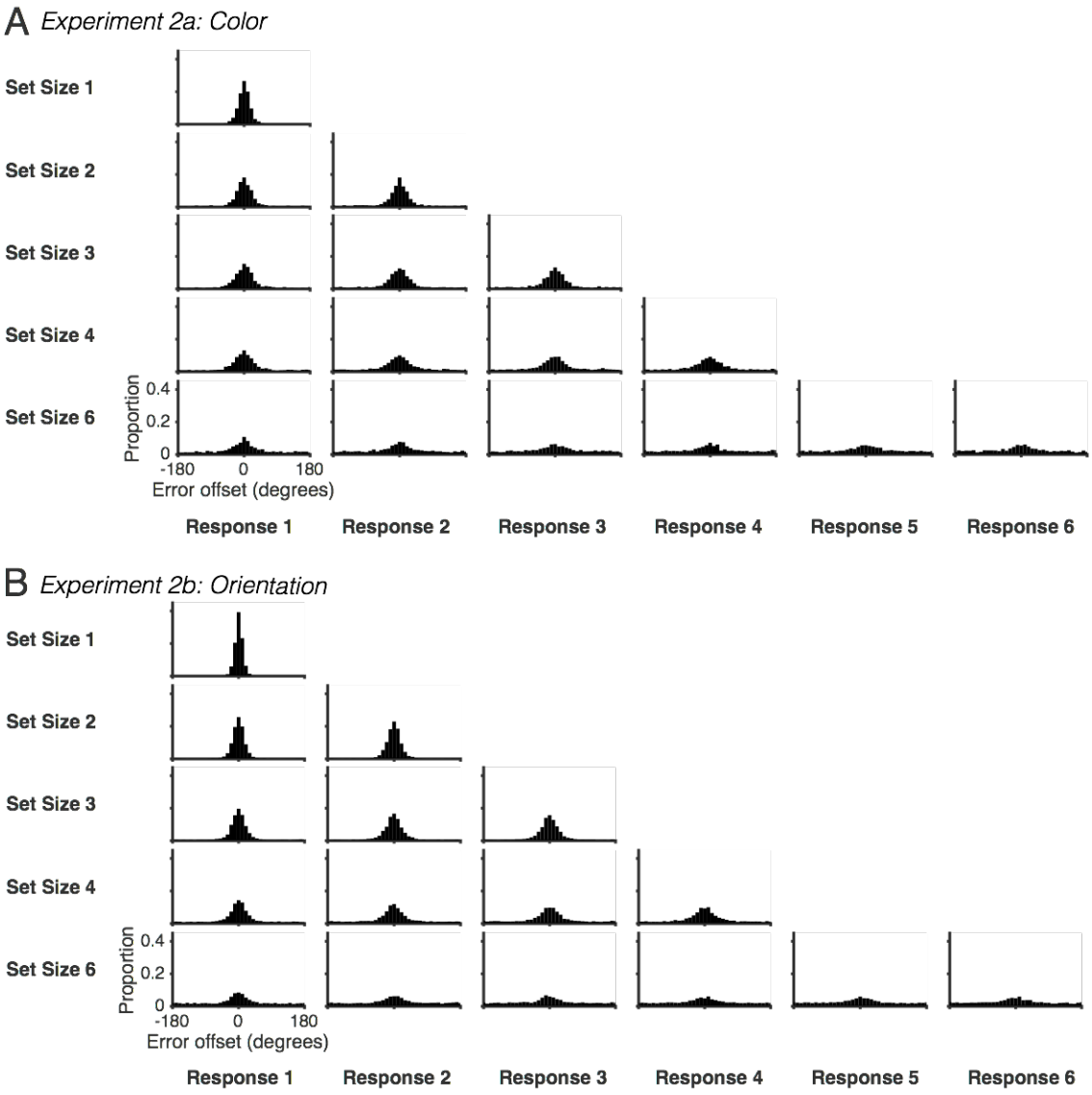

Figure 3-7 Computer-ordered responses are relatively unaffected by response order
All set sizes and responses are shown for (A) Experiment 3-2a and (B) Experiment 3-2b.

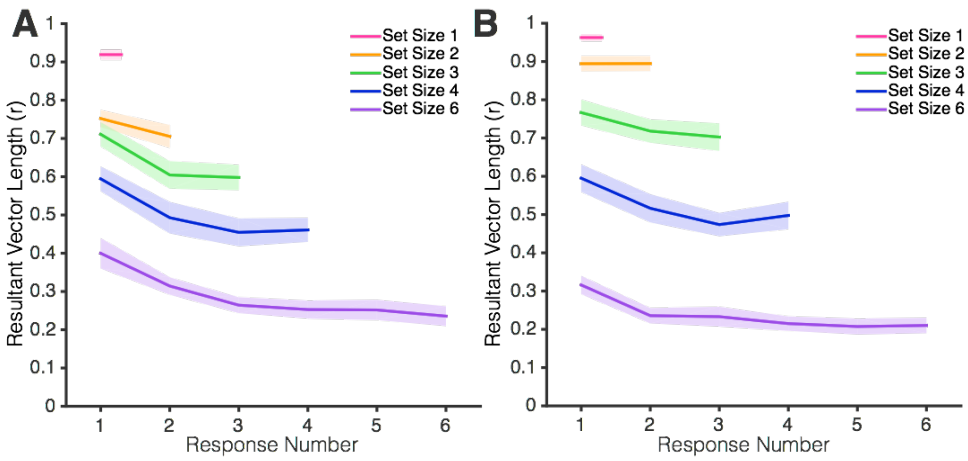

Figure 3-8 Mean Resultant Vector Length for computer-ordered responses
MRVL across responses in Experiment 3-2a (A) and -2b (B). Shaded error bars represent 1 standard error of the mean (SEM).

To summarize, although both Experiments 3-2a and -2b produced evidence for a modest amount of output interference, the slope of this decline in the computer-guided condition was more than six times shallower than when subjects chose the response order themselves in Experiment 3-1 (Figure 3-8). While resultant vector length decreased by 0.18-.19 per response in the self-ordered experiments (Experiments 3-1a and -1b), it decreased by only .04 and .02 per response in Experiments 3-2a and -2b. That is, output interference could at most explain around 16% of the decline observed in Experiment 3-1. Thus, the effect of output interference alone cannot explain the dramatic decrease in performance during the self-ordered response procedure used in Experiment 3-1. Instead, we conclude that subjects in Experiment 3-1 used metaknowledge to report the best stored items first. Thus, the final responses in the subject-ordered conditions contained no detectable information about the target.

Table 3-3 Change in Mean Resultant Vector Length across responses in Experiment 3-2a

Set Size	*df1*	*df2*	*F*	*p*	η_p^2
2	1	16	11.75	.003	.42
3	2	32	25.3	<.001	.61
4	3	48	18.4	<.001	.53
6	5	80	14.9	<.001	.48

Table 3-4 Change in Mean Resultant Vector Length across responses in Experiment 3-2b

Set Size	*df1*	*df2*	*F*	*p*	η_p^2
2	1	18	.003	.96	<.001
3	2	36	16.1	<.001	.47
4	3	54	25.0	<.001	.58
6	5	90	13.3	<.001	.43

Subjective ratings of guessing again predict uniform-distributed responses.

Experiment 3-2 replicated the finding that participants' subjective ratings of guessing strongly corresponded with mixture-model estimates of a guessing parameter. We fit a mixture model to each response within each set size (i.e., Set Size 2 first response, Set Size 2 second response, etc.), and we quantified the percentage of responses that the subjects reported guessing with a mouse button click. In line with the earlier results, we found a tight relationship between participants' subjective reports of guessing and the mixture model's estimation of a guess state. One subject in Experiment 3-2a showed no relationship between confidence ratings and mixture model parameters, because they reported that every item was not a guess (they did not use both buttons). This subject was not included in metaknowledge analyses. On average, the strength of the within-subject correlation was r = .91 (SD = .03, all p-values < .001) in Experiment 3-2a and r = .94 (SD = .06, all p-values <.001) in Experiment 3-2b. Once again, subjects were slightly over-confident. When plotting the model's guessing parameter on the x-axis, the average slope was 0.89 (SD = .16) for Experiment 3-2a and 0.70 (SD = .25) for Experiment 3-2b; the average intercept was -.01 (SD = .13) for Experiment 3-2a and -.06 (SD = .04) for Experiment 3-2b.

If the guessing we observed in Experiment 3-1 was due to output interference, then purely uniform error distributions should remain concentrated in the final responses of the trial even when the order of report was randomized in Experiment 3-2. By contrast, if the drop in performance across responses in Experiment 3-1 resulted from a bias to report the best remembered items first when subjects controlled response order, then guesses should be evenly distributed across responses in Experiment 3-2 when response order was randomized. This predicts that there should be some trials in which the best stored items happened to be probed last while items that could not be stored were probed first. Such a pattern of results could not be explained by an output interference account. Thus, we used subjects' confidence ratings to identify two types of trials from Experiment 3-2: (1) Trials where the three items probed late in the trial were guess responses and (2) Trials where the three items probed *early* in the trial were guess responses. We took these trials across subjects and binned them together, then tested whether a uniform distribution best fit each response in the trial. Consistent with a mnemonic variability account of Experiment 3-1, we found that uniform error distributions occurred early -- but not late -- in the trial when participants reported guessing for the early responses (Figure 3-9). Likewise, when participants indicated that they were guessing during the final three responses, purely uniform distributions were observed for the last three – but not the first three – responses. Thus, Experiment 3-2 showed that guessing prevalence was decoupled from response order, arguing against an output interference account of the observed uniform distributions. This analysis also gives insight into subjects' metaknowledge accuracy; subjective confidence ratings nicely tracked the location of guess responses (early versus late). However, these ratings were imperfect; "confident" responses still contained a sizeable uniform component, indicating that

participants sometimes had less information in mind than they reported, converging with earlier evidence (Adam & Vogel, 2017).

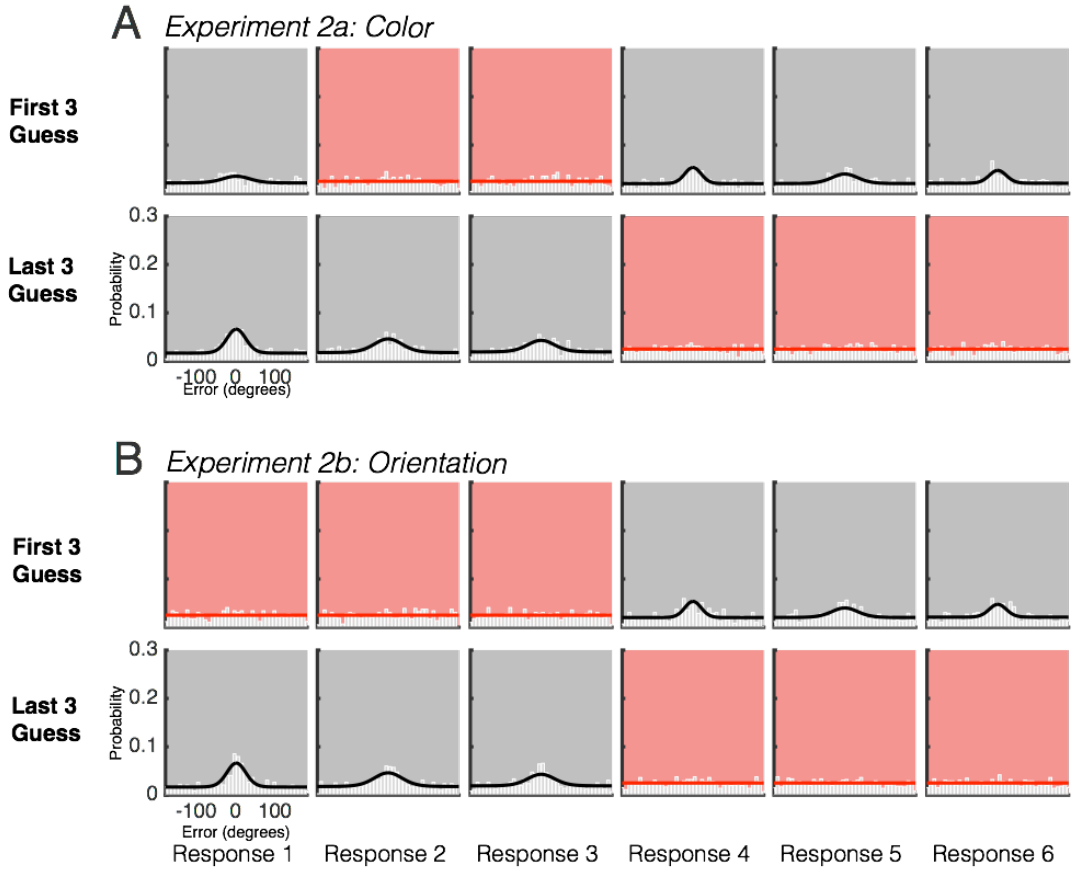

Figure 3-9 Response distributions for Set Size 6 trials in Experiment 3-2, split by when participants reported guessing
Trials are binned by whether participants reported the first three responses as guesses (top row) or the final three responses as guesses (bottom row) in (A) Experiment 3-2a and (B) Experiment 3-2b.

The role of instructions in guessing

Experiments 3-1 and 3-2 both provided compelling evidence that participants do not remember all items with a set size of 6 items, and that the observed guesses were not a result of output interference. Nevertheless, we also considered whether the specific instructions subjects received in our study may have influenced their tendency to guess. In both experiments, we

asked participants to make a dichotomous "some" or "no" information confidence judgment. Gathering confidence ratings was extremely useful for validating the relationship between subjective guessing states and mixture model estimates of guess rates. However, we also wanted to assess whether instructions that emphasized the possibility of guessing may have artificially encouraged guessing behaviors. To test this possibility, we ran a similar whole report procedure in which we eliminated the meta-knowledge assessment and participants were instructed to never randomly guess.

Experiment 3-3

Materials & Methods

Ten participants from the University of Chicago community participated in Experiment 3-3 for payment ($10/hour). Stimuli and procedures were nearly identical to Experiment 3-1b; participants were asked to remember the orientation of 1, 2, 3, 4 or 6 items and to report all items in any order they chose. However, participants did not make confidence ratings using the two mouse buttons. During the instruction period, the experimenter instructed the participants that they should remember all of the items. Participants were instructed: "Even if you feel you have no information in mind, do your best when making your responses. Even the information that you have in mind is extremely imprecise, it will still lead you in the right direction."

Results & Discussion

Even though participants were not instructed to dichotomously report guess states, we still observed responses that were best described by a uniform distribution (Figure 3-10). Using a BIC approach to compare the fit of a uniform distribution with the fit of a mixture model, we found a mean of 2.6 uniform distributions (SD = .97, range from 1 to 4). Non-parametric tests of uniformity agreed with this number. To conclude, the key finding that some working memory

responses are best fit by a uniform distribution was not a result of instructions that allowed for the possibility of guessing.

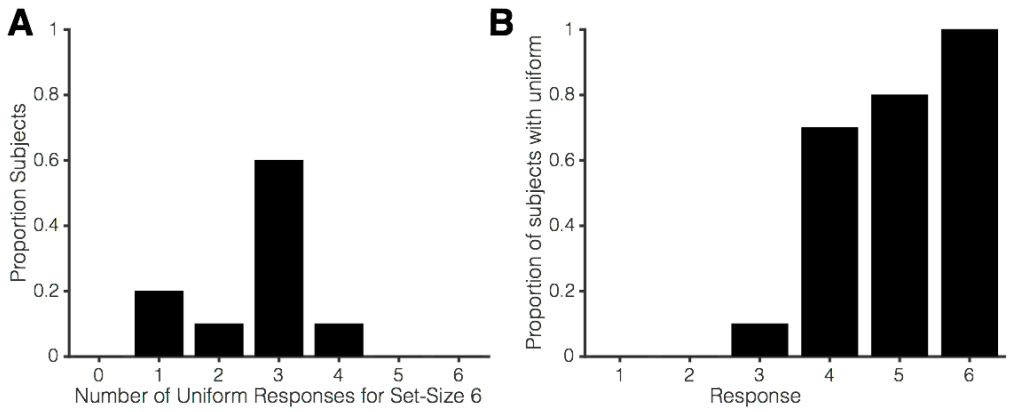

Figure 3-10 Number of uniform responses for Set Size 6 in Experiment 3-3
(A) Histogram of subjects' total number of Set Size 6 responses that were best fit by a uniform distribution. (B) Number of subjects' responses best fit by uniform distributions as a function of response number.

Simulation results: Variable precision models mimic guessing behaviors by positing very low precision memories

Across several experiments, we have shown that guessing accounts for a large proportion of subjects' responses when working memory load is high (e.g., ~50% of Set Size 6 responses). This result is apparently in conflict with past findings that a variable precision model that denies guessing (hereafter called "VP-no guessing") is a strong competitor for models that presume a high prevalence of storage failures (van den Berg et al., 2014). How does the VP-no guessing model achieve close fits of these aggregate error distributions? We hypothesized that the VP-no guessing model may succeed by postulating memories that are so imprecise that they cannot be distinguished from random guesses. Indeed, others have noted that this is a potentially troubling feature of VP models (Nosofsky & Donkin, 2016b; Sewell et al., 2014). To test this hypothesis, we implemented a VP-no guessing model to fit the aggregate error distribution from Experiment

3-1a. This provided a clear view of the range of precision values used by this model to account for performance with six items. Next, we assessed what percentage of these memories could be distinguished from random guesses with varying amounts of noise-free data. To anticipate the findings, the VP-no guessing model posits memories that are undetectable within realistic experimental procedures, and it does so at a rate that tracks the guessing parameter within a standard mixture model.

Using code made available from van den Berg et al. (2014, code accessed at http://www.ronaldvandenberg.org/code.html), we fit the VP-no guessing model to individuals' aggregate data (i.e., combining all responses for a given set size) from Experiment 3-1a. The variable precision model proposes that precision for each item in the memory array is von Mises-distributed with concentration κ. The precision of the von Mises for each item in the memory array is randomly pulled from a range of possible precision values. The range of possible precision values is determined by a gamma-distributed higher-order distribution, and the shape of this gamma distribution changes with different numbers of target items. This particular implementation of the variable precision model fits parameters while simultaneously considering performance across all set sizes. For Set Size 1 memory arrays, the gamma distribution contains mostly high precision von Mises distributions; for larger memory arrays, the gamma distribution contains a larger proportion of low precision von Mises distributions (Figure 3-11).

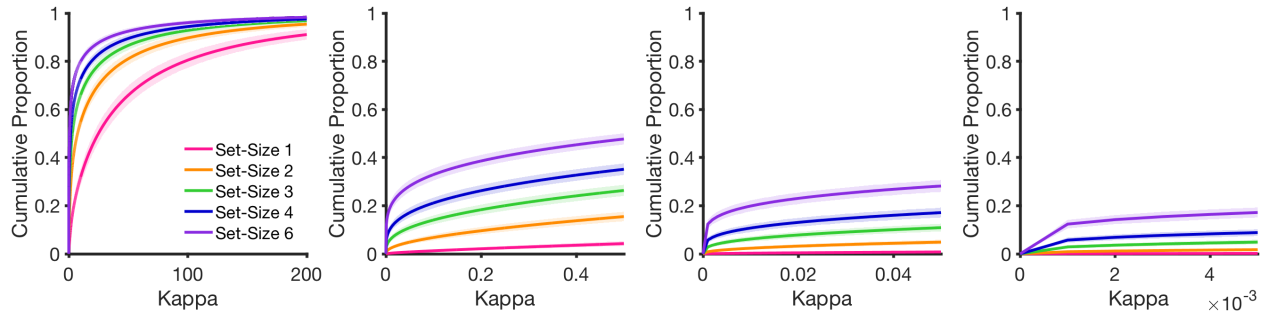

Figure 3-11 Cumulative distribution functions for the variable precision model across set sizes
X-axis represents the concentration parameter of the von Mises distributions pulled from the higher order gamma distribution. Y-axis represents the cumulative proportion of trials in which a given concentration (κ) or less is pulled. From left to right, the scale of the x-axis is zoomed in to better illustrate the proportion of very low precision representations that make up each higher-order distribution. Shaded error bars represent 1 Standard Error.

For this analysis, we focused on the higher-order distribution of precision values for the Set Size 6 condition. We found that the VP-no guessing model posits a high prevalence of representations with exceedingly low precision. To further visualize this point, we computed decile cut-offs for each participant, and then took the median value for each decile cut-off across participants. Figure 3-12 shows the probability density function for each decile of von Mises distributions posited by the VP-no guess model. As can be seen, the von Mises PDF appears, by eye, to be perfectly flat for a large proportion of Set Size 6 trials (20 -30%). Critically, this PDF visualization is hypothetical in that it assumes infinite numbers of samples from a von Mises PDF, and with infinite numbers of trials we can easily distinguish a diffuse von Mises distribution (e.g., 40th Percentile, K = .318) from a uniform distribution. However, if we were to experimentally sample "trials" from the hypothetical PDF, our ability to distinguish this distribution from uniform would depend on the number of samples. Below, we show that variable precision models that deny guessing must postulate a very high prevalence of memory representations that cannot be detected with a feasible behavioral study.

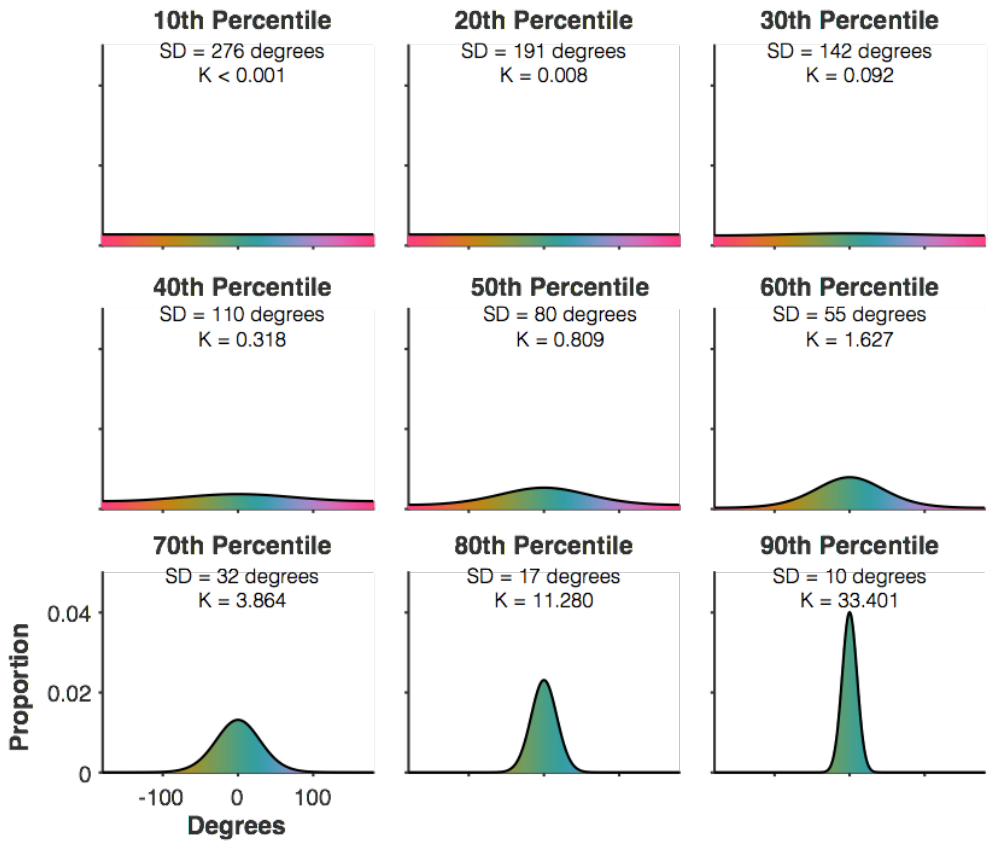

Figure 3-12 Illustration of von Mises distributions used by the variable precision model to account for Set Size 6 performance
Precision values of the von Mises distributions are given as both concentration (K) and standard deviation (SD) in degrees.

To illustrate how poor the VP-no guessing model's hypothesized representations are, we ran simulations in which we used varying amounts to data to discriminate between a von Mises distribution of precision κ and a uniform distribution. We did so for a range of precision values and a range of trial numbers. We chose 10 log-spaced bins between 10 and 1,000,000 samples ("trials"), and we ran 200 iterations of randomly sampling trials from a von Mises distribution with various concentrations (κ). For each iteration, we compared the fit of the von Mises to the fit of a uniform distribution using BIC comparison in MemToolbox (Suchow et al., 2013), and we took the difference score in BIC fits for the uniform and the von Mises distributions. We

defined our ability to discriminate from uniform as the precision value at which the 95% Confidence Intervals for a given number of trials remained in favor of the uniform. Discriminable precision values are shown as a function of the number of samples in Figure 3-13. With only 10 samples from a von Mises PDF, we could discriminate between uniform and a von Mises with $\kappa = 0.65$ (SD = 88 degrees). With 1,000,000 samples from a von Mises PDF, we could discriminate between uniform and a von Mises with $\kappa = 0.007$ (SD = 193 degrees). That is, a concentration of less than $\kappa < .007$ is so diffuse that the central tendency is *undetectable* with 1,000,000 samples of noise-free data sampled from a true von Mises distribution. Of course, because these simulations presume noise-free data, these simulations overestimate the true detectability of these putative representations in a real experiment.

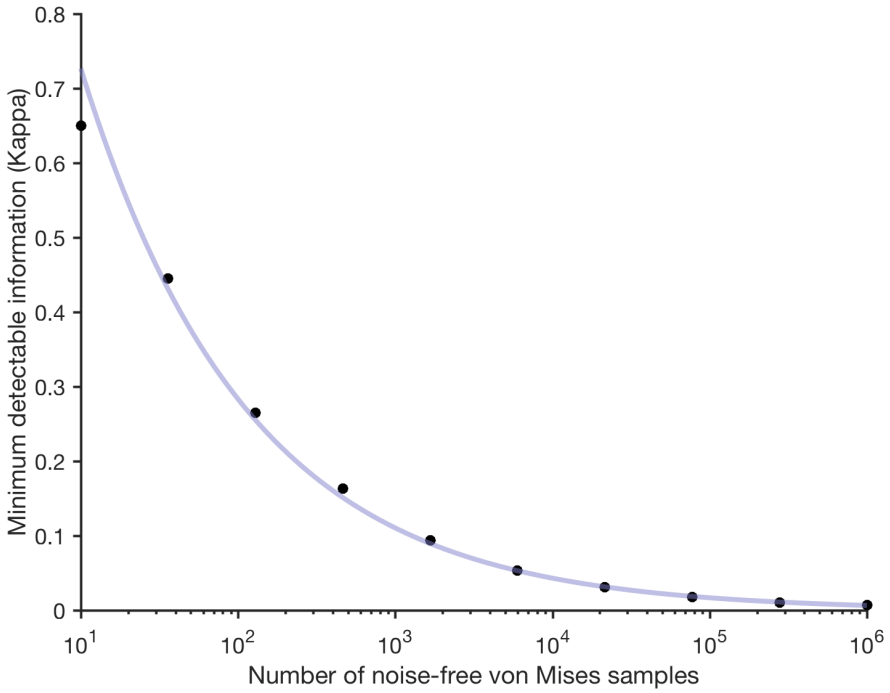

Figure 3-13 The minimum detectable amount of information as a function of the number of samples
This figure depicts the minimum detectable amount of information (Kappa) needed to distinguish a von Mises distribution from a uniform distribution, as a function of the number of noise-free von Mises-distributed samples. The fitted line is the linear fit of the log-transform of the sample number and the log-transform of the precision threshold.

Finally, we determined the prevalence of effectively undetectable memories for each subject (defined as those requiring at least one million trials to discriminate from uniform). On average, 22.0% (SD = 11.2%) of VP-no guessing distributions had a precision less than or equal to κ =.007, and this percentage ranged from 7.0% to 46.3% of a subjects' higher-order distribution values. Further, the proportion of these representations was tightly coupled with mixture model estimates of the guess rate for Set Size 6 (r = .81, p <.001), underscoring the possibility that the VP-no guessing model is mimicking random guesses. Thus, although the VP-no guessing models can achieve close fits to aggregate error distributions, their success depends

on the assumption that a substantial number of these memories would be literally undetectable with a feasible behavioral paradigm.

In cases of model mimicry, it is important to question which model is the mimic. Could it be that models endorsing guessing are mimicking low precision representations? We offer three arguments to the contrary. First and foremost, the whole report data from Experiments 3-1 and 3-2 revealed a high prevalence of error distributions that were best modeled by a uniform distribution with zero free parameters. Thus, while it has been difficult to adjudicate between models that endorse and deny guessing with the aggregate error distributions generated by partial-report studies, (van den Berg et al., 2014), whole report data provide clear positive evidence that the modal subject guesses about half of the time for Set Size 6. Second, participants report that they were guessing, and these subjective reports were excellent predictors of simple mixture-model estimates of guessing rates (Experiments 3-1 and 3-2). We contend that subjective judgments of whether an item has been stored are decidedly relevant for characterizing the contents of a memory system that is thought to hold information "in mind" (however, see general discussion for limitations to this claim). Finally, we think it's reasonable to question the plausibility of "memories" that are as imprecise as those posited by variable precision models that deny guessing. For example, the VP-no guessing model requires that the worst memory out of six simultaneously-presented items would have a standard deviation of around 230 degrees. Based on our earlier simulation of trial numbers and precision values, this means that an ideal observer would require over 900 million trials to produce above-chance performance in detecting the largest possible difference in color or orientation (e.g., red vs. green, or horizontal versus vertical). At the least, there may be a consensus that such homeopathic amounts of information would be of little use for purposeful cognitive tasks.

General Discussion

Understanding the nature of capacity limits in working memory has been a longstanding goal in memory research. The capacity debate has been dominated by a theoretical dichotomy of "number" versus "precision." Discrete resource models have argued that capacity is limited by the number of items that can be concurrently stored, and that subjects to resort to guessing when more than a handful of memoranda are presented (Fukuda, Awh, & Vogel, 2010; Zhang & Luck, 2008). By contrast, continuous resource models argue that mnemonic resources can be distributed amongst an unlimited number of items (Bays & Husain, 2008; van den Berg et al., 2012; Wilken & Ma, 2004). In addition, there is a growing consensus that mnemonic precision varies across items and across trials (e.g., Bae et al., 2014; Fougnie et al., 2012). In the present data, for example, item-by-item varations in precision were plainly demonstrated. However, variable mnemonic precision is compatible with both discrete resource models that propose an item limit and continuous resource models that allow for the storage of unlimited numbers of items. In sum, despite a proliferation of work a key aspect of the number/precision theoretical dichotomy has remained unresolved: Do participants guess? If so, do we need capacity limits to explain guessing behavior? Even some of the most sophisticated modeling efforts have reached a stalemate (van den Berg et al., 2014).

In this context, our findings provide compelling evidence that the modal subject guesses about half of the time with a memory load of six items. This confirmation of the guessing construct is critical for the broader idea that working memory performance differs both in terms of the number and the precision of the representations that are stored. Further, the variations in the number of items that individuals could store (2 – 5 items) aligned closely with past estimates of items limits in visual working memory (Cowan, 2001; Fukuda, Vogel, et al., 2010; Luck &

Vogel, 1997; Zhang & Luck, 2008). Finally, Experiments 3-2a and -2b showed that output interference could not explain the powerful decline in mnemonic performance across responses that resulted in uniform distributions. While there was some evidence of output interference when all items were probed in a random order, it could not account for the dramatic declines in performance that were observed when subjects chose the order of their response. Thus, while past studies have shown that aggregate error distributions can be equally well fit by competing models that endorse and deny guessing behaviors, our whole report procedure provided unambiguous evidence for a high prevalence of guessing responses.

The finding that individual differences in capacity yielded similar findings to previous estimates rules out two key classes of models. First, this provides compelling evidence against "no guessing" models in which participants are able to store all items in the array. To account for individual differences in performance, "no guessing" models posit that participants with poor working memory performance have less precise memories, but store some information about all presented items. Our findings clearly rule out such a model; we did not find any participants for whom a "no guessing" model was best. Second, these results constrain models of trial-by-trial variability in performance. Some models have proposed that effective capacity (i.e., the number of stored items) varies from trial to trial, and it could do so in several ways. Models that include guessing while denying item limits propose that the number of stored items varies dynamically from trial to trial but is not limited by an upper bound. For instance, trial-by-trial performance may be Poisson- or uniform-distributed (van den Berg et al., 2014). On the other hand, item-limited models propose that performance varies from trial to trial, but only in one direction – below the capacity limit for each individual (Adam et al., 2015). Critically, the first class of models (dynamic but capacity-unlimited) also yield the prediction that the majority of subjects

should show non-uniform distributions at all 6 response positions. Instead, we find support for the latter model (dynamic but capacity-limited); there is still some degree of guessing for early responses, indicating that subject frequently under-perform their maximum capacity. However, participants do not over-perform past a hypothetical "mean limit", as shown by the pure uniform error distributions we observed for late response in the set size six condition.

Interestingly, objective estimates of guessing dovetailed with the subjects' own reports of whether they had any information about the item in question. The frequency with which subjects endorsed guessing precisely tracked the guess rate as estimated by a standard mixture model (Zhang & Luck, 2008). Note, the alignment of the guessing parameter and subjective guess reports alone cannot distinguish between a guessing account and a variable precision account of working memory limits. For example the variable precision model might posit that participants will choose the "guess" label whenever mnemonic precision does not pass a certain threshold. However, the precise alignment of subjective ratings with the mixture model's guess parameter greatly bolsters the face validity of the mixture model's guessing parameter.

There is a growing consensus that working memory responses have variable precision. In particular, many recent papers have found that a large amount of variability in working memory precision may arise from stimulus-specific differences, such as color categories, orientation categories, or verbal labels (Bae et al., 2015, 2014; Donkin, Nosofsky, Gold, & Shiffrin, 2015; Hardman, Vergauwe, & Ricker, 2017; Pratte, Park, Rademaker, & Tong, 2017). However, there is growing doubt as to whether variability in precision is due to variations in the allocation of a mnemonic resource *per se*, as opposed to lower level differences in the quality of encoding or the imposition of categorical structure for different types of stimuli. For instance, recent work by Pratte and colleagues (2017) measured performance in a task that required the storage of

orientations in working memory and replicated earlier work showing variability in precision across concurrently stored items in working memory. Rather than presuming that this variability reflects variation in the allocation of mnemonic resources *per se* (van den Berg et al., 2012), Pratte et al. showed that much of this variability was explained by higher precision for for orientations near the horizontal and vertical meridians. This observation falls in line with the "oblique effect" that has been documented in past studies of visual perception (Appelle, 1972). Strikingly, when the oblique effect was incorporated into competing discrete and continuous resource models, discrete models that endorsed guessing were the clear winner of the model competition. In other words, once stimulus-driven sources of variable precision were acknowledged, the best account of the data posited a high prevalence of guessing responses.

Clear evidence for guessing in a working memory task has important implications for our taxonomy of the processes that determine memory performance. In past work, putative item limits in working memory have been argued to predict variations in fluid intelligence, scholastic achievement, and attentional control (Cowan et al., 2005; Engle, Tuholski, et al., 1999; Fukuda, Vogel, et al., 2010; Unsworth et al., 2014). A simple interpretation is that a common mental resource determines the number of items that can be stored in working memory and one's ability to handle a variety of cognitive challenges. According to continuous resource models, however, it is not possible to measure individual differences in the number of items that can be stored in working memory, because all observers can store all items regardless of set size. By this account, the apparent variations in the number of items that can be maintained are an illusion created by limitations in memory quality. Thus, continuous resource models explicitly argue that individual differences in memory performance will be explained by a single factor that determines memory *quality*.

Evidence from Awh et al. (2007) challenged the idea that memory quality is the determining factor of working memory limits. They measured performance in a change detection task while manipulating the size of the changes that occurred in the test display. When changes were very large, they reasoned that subjects should be able to detect the change whenever the probed item had been stored, because precision should not be a limiting factor. When changes were relatively small, however, they reasoned that successful change detection would be limited more by memory quality, because more precise memories would be needed to detect a relatively small mismatch between the sample and test. By contrast, a continuous resource model asserts that precisely the same mnemonic resource determines performance with small and large changes, because there is no limit to the number of items that can be stored. Disconfirming this prediction, Awh et al. (2007) found that performance with big and small changes was completely uncorrelated, despite having positive evidence that both scores were reliable. This finding has since been extended by looking at the pattern of errors to big-change and small-change trials across multiple response probabilities (Nosofsky & Donkin, 2016a); participants frequently endorse that large changes are "the same" on big-change trials, even though they are capable of discriminating much smaller changes. Thus, number and precision may represent distinct facets of working memory ability.

Memory precision does not seem to be the limiting factor for detecting changes within displays, and there is also little evidence that precision predicts individual differences in working memory performance. Fukuda, Vogel, et al. (2010) carried out a latent variable analysis of a variety of change detection tasks that required the detection of either very large or small changes. This analysis revealed distinct factors for the detection of large and small changes, with no reliable cross loading between these factors. This result provides a robust confirmation of the

earlier finding, suggesting that number and resolution may indeed be dissociable aspects of memory ability. Moreover, Fukuda, Vogel, et al. (2010) found that while the number factor was a robust predictor of fluid intelligence, there was zero evidence for such a link between precision and fluid intelligence. Likewise, capacity is reduced but precision is spared in people with schizophrenia (Gold et al., 2010). Thus, a two-factor model that distinguishes between the number of items stored and the precision of those mnemonic representations is needed to account for performance with large and small changes, and these two factors have unique relationships with fluid intelligence.

In conclusion, we present clear evidence for a high prevalence of guessing responses in a visual working memory task. When subjects were allowed to choose the order of report in a whole report memory task, we observed a monotonic decline in memory performance with each successive response, and the modal observer produced uniform error distributions – the hallmark of guessing – for three of the items in a six item display. Control experiments ruled out the hypothesis that output interference generated this monotonic decline in performance across responses; modest evidence of output interference was observed, but it accounted for only a modest proportion of the decline across responses. Instead, we conclude that subjects used accurate metaknowledge to report the best remembered items first. In turn, this yielded robust evidence of guessing behaviors when the last responses were examined, and supports the idea that working memory is subject to clear item limits.

CHAPTER 4. DECODING THE LIMITS OF SIMULTANEOUS STORAGE IN VISUAL WORKING MEMORY

Introduction

There is broad consensus in the literature that working memory (WM) is a limited resource, but the precise nature of working memory's limitations is still under debate. While competing models agree that there are limits on the total amount of available WM resources, they differ in how these resources are distributed. Capacity-limited models (i.e., "slots") propose that resources are deployed in an item-limited fashion; only a few items can be remembered and additional items receive no resources (Luck & Vogel, 1997; Zhang & Luck, 2008). Conversely, information-limited (i.e., "continuous resources") accounts propose that all items in the display receive some amount of mnemonic resources, and each item's representation may be very imprecise (Bays & Husain, 2008; van den Berg et al., 2012; Wilken & Ma, 2004).

Prior studies of the neural correlates of working memory have revealed neural signals that scale with load and, critically, reach an asymptote at hypothesized working memory limits. This asymptote in load-dependent neural signals has been interpreted as a corroboration of item-limited models; once the item limit is reached, adding more items to the memory set increases load-dependent signals no further. In fMRI, univariate BOLD activity in the intraparietal sulcus (IPS) increases as more to-be-remembered items are added to a memory array (Todd & Marois, 2004; Y. Xu & Chun, 2006). Likewise, in EEG, a lateralized difference wave to memory items, the contralateral delay activity (CDA), similarly reaches an asymptote at behavioral capacity limits (Vogel & Machizawa, 2004). Further, individual differences in load-dependent changes to these hypothesized memory signals track individual differences in behavioral performance (Todd

& Marois, 2005; Vogel & Machizawa, 2004; Vogel et al., 2005), supporting their role in the successful maintenance of a limited amount of information.

Univariate signals of working memory are consistent with capacity-limited models of visual working memory, but the interpretation of these signals is limited in several ways. First, these signals reveal neural activity related to the processing of an entire memory array (e.g., average activity when 6 items are presented versus 3 items), but they do not provide information about item-specific processing (e.g., were only 3 out of 6 items stored, or was an equivalent amount of resources spread thinly among all 6 items?). Further, the description of these signals as asymptotic has been challenged by a competing model which proposes that this signal may be equally well-described as continuously increasing (Bays, 2018). Finally, multivariate methods have revealed that memory-specific information can be decoded in regions where no load-dependent changes to the univariate signal are observed (Harrison & Tong, 2009; Riggall & Postle, 2012; Serences, Ester, Vogel, & Awh, 2009; Serences & Saproo, 2012). Thus, there is ongoing debate as to whether parietal load-dependent signals encode working memory representations or if, instead, they represent peripheral attentional demands related to the task (Magen, Emmanouil, McMains, Kastner, & Treisman, 2009; Mitchell & Cusack, 2008, but also see: Sheremata, Somers, & Shomstein, 2018).

The application of multivariate techniques to neural data has allowed researchers to decode information about an item held in visual working memory from voxelwise or electrode-wise patterns of activity in fMRI or EEG. Multivariate classification accuracy decreases as a function of memory load (1, 2, or 3 items), consistent with the degradation of memory traces due to competitive interactions when multiple items are stored (Buschman et al., 2011; Emrich, Riggall, LaRocque, & Postle, 2013). More recently, inverted encoding models have been applied

to more precisely infer the quality of mnemonic representations from population-level "information channels" tuned to a feature of interest (Brouwer & Heeger, 2009; Sprague, Saproo, & Serences, 2015; Sprague & Serences, 2013). The interpretation of inverted encoding models is specific to the feature of interest (e.g., orientation or spatial location) and provides fine-grained information about the quality of feature-specific representation encoded by patterns of activity across voxels or electrodes. In fMRI, the application of inverted encoding models has revealed stimulus-specific representations in occipital and parietal cortices that degrade from 1 to 2 items (Sprague, Ester, & Serences, 2014). In EEG, the position of an item maintained in working memory can be decoded from the pattern of alpha-power (8-12 Hz) across the scalp (Foster, Sutterer, Serences, Vogel, & Awh, 2016), and this signal also encodes a decrease in the fidelity of memory from 1 to 2 items (Sutterer, Foster, Adam, Vogel, & Awh, in prep).

While set size manipulations of 1 to 3 items (within capacity) has been useful for delineating the neuronal populations involved in the active storage of information (Buschman et al., 2011; Emrich et al., 2013; Sprague et al., 2014), they do not inform crucial debates about capacity limits. To disambiguate competing models of active visual working memory storage, we need methods to simultaneously decode all items from supra-capacity (>4 items) arrays. Here, we present experiments that expand inverted encoding model techniques to examine the representation of item-specific information from supra-capacity arrays (6 items). To do so, we had participants remembered an array of color-space pairings (i.e., report which color belonged at each location), which requires participants to maintain both color and space information. We then use an inverted encoding model to decode remembered locations from the topographic distribution of alpha-band power (Foster et al., 2016). This task explicitly requires subjects to

maintain both color and location, but location is robustly represented even when color is remembered but location is task-irrelevant (Foster, Bsales, Jaffe, & Awh, 2017).

Here, we show direct evidence that items within supra-capacity arrays do not receive equal mnemonic resources. In fact, for many items in supra-capacity arrays we were unable to decode any information. These data rule out equal-precision models of working memory capacity (Bays et al., 2009; Bays & Husain, 2008), and instead support capacity-limited models of visual working memory (Adam, Vogel, & Awh, 2017; van den Berg et al., 2014; Zhang & Luck, 2008).

Overview of experiments

The central aim of the experiments presented in this chapter is to expand multivariate tools in order to decode the identity of multiple items from within a supracapacity array. In Experiment 4-1, we applied an inverted encoding model to human EEG data while participants performed a visual working memory task in which they remembered 1, 2, or 3 colored items. We replicated previous work showing that the fidelity of broad-scale population codes decreased with increasing memory load (from 1 to 3 items). In Experiment 4-2, we expanded this method to test what happened when participants are presented with a supracapacity array (6 items) – can all locations be decoded equally well? Participants again performed a working memory task in which they remembered 1, 3, or 6 colored items. By applying an inverted encoding model to each item from the array as a function of hypothesized mnemonic quality (i.e., as a function of response order), we were able to demonstrate that mnemonic resources were unevenly distributed among items. The location of prioritized memory items could be robustly decoded from the topography of alpha-band power, whereas the location of supra-capacity items (guesses) could not be decoded from this active memory signal.

Materials & Methods

Participants

Participants were recruited from the University of Chicago and surrounding community. All participants provided written, informed consent and procedures were approved by the University of Chicago IRB. Participants were between the ages of 18 and 35 and reported normal or corrected-to-normal visual acuity and normal color vision. Participants were excluded from analysis if fewer than 300 trials remained in the training dataset (set size 1) or if fewer than 150 trials remained in any of the test sets (set size 2 – 6).

A total of 37 people participated in Experiment 4-1 (gender = 22 male, 15 female; mean age = 23.9 years, SD = 4.1). After artifact rejection, 3 subjects were excluded for excessive artifacts. An additional 5 subjects were excluded for the following reasons: the participant was discovered to be color-blind (n = 2); problems with the recording equipment (n = 2); the participant chose to leave the experiment early (n = 1). The final sample size for Experiment 4-1 was 29 (17 male, 12 female; mean age = 23.6 years, SD = 3.5).

A total of 39 people participated in Experiment 4-2 (gender = 19 male, 18 female, 2 other; mean age = 23.4 years, SD = 3.7). A total of 9 subjects were excluded for excess artifacts. An additional 2 subjects chose to leave the experiment early. The final sample size for Experiment 4-2 was 28 (15 male, 12 female, 1 other; mean age = 22.4 years, SD = 2.3).

Stimuli

Stimuli were rendered in MATLAB (The MathWorks, Natick MA) using the Psychophysics toolbox (Brainard, 1997; Pelli, 1997). Participants were seated approximately 78 cm from a 24-inch LCD monitor (XL2430 monitor, BenQ, Taipei, Taiwan; refresh rate = 120 Hz; resolution = 1024 h X 1920 w pixels), with chins resting on a padded chin rest. Colored

squares (side = 0.98°) were presented on a dark gray background (RGB = 85 85 85) at one of 8 positions (equally spaced on an imaginary circle with radius = 2.55°). Subjects were instructed to maintain fixation on a small black dot (radius = 0.12°) at the center of the screen. Colors were drawn from a pool of 9 easily discriminable colors (RGB: red = 255 0 0; orange = 255 128 0; yellow = 255 255 0; green = 0 255 0; blue = 0 0 255; cyan = 0 255 255; magenta = 255 0 255; white = 255 255 255; black = 1 1 1). All colors within a trial were unique. Spatial pre-cues were 1 or several small gray lines (0.06 x 0.12°; RGB = 150 150 150) placed at the center. In Experiment 4-2, light gray placeholder squares (RGB = 151 151 151) were used to hold visual stimulation roughly constant across set size conditions.

Procedures

Experiment 4-1. On each trial, participants were cued to attend 1, 2, or 3 locations with a small spatial cue presented at fixation (Figure 4-1). After the cue (300 ms), a memory array appeared for 150 ms. The memory array always contained 3 uniquely colored items. Subjects were instructed to remember only the items presented at the locations indicated by the spatial pre-cue. Starting at the memory array onset, this design holds visual stimulation constant across all three set sizes. A blank delay period followed the memory array (1,150 ms) followed by an untimed response period. During the response period, participants were presented with colored response grids (3 x 3 grids containing all 9 possible colors) only at the locations that the participants were cued to remember. Participants responded by clicking the color in the grid that corresponded to the color presented at that location. Participants had to make a response to every cued location before moving onto the next trial. After making all responses and making one additional mouse click, the next trial began with a blank inter-trial interval of 500 ms.

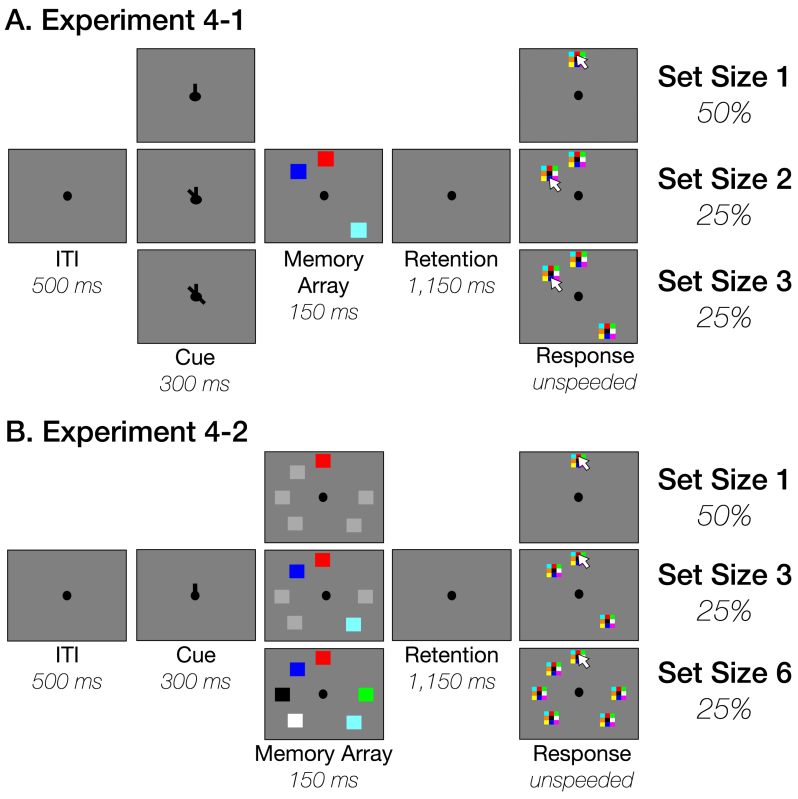

Figure 4-1 Illustration of tasks for Experiment 4-1 and 4-2
Illustration of trial events for Experiment 4-1. Three memory items were always presented (memory array), but subjects were cued to remember either 1, 2, or 3 of the memory items. At test, subjects reported all cued items. (B) Illustration of trial events for Experiment 4-2. Subjects were cued to prioritize 1 item. Then, 6 squares appeared. Either 1, 3, or 6 squares were colored, and participants were instructed to attempt to remember all colored. At test, participants reported all items from the memory array.

Participants performed 20 blocks of 64 trials. After each block of trials, participants received a minimum 30 second break (with the option for taking a longer break whenever needed) and saw performance feedback for the block (mean number of items correctly reported for each set size).

Experiment 4-2. Trial procedures were very similar to Experiment 4-1, with some adaptations for the inclusion of a supra-capacity set size (6 items). Rather than being cued to all locations, participants were pre-cued to a single location. Participants were instructed that they should prioritize the cued location, and also attempt to remember all other colored items in the

array. This procedure ensured that participants prioritized the 8 location bins in approximately equal numbers. Experiment 4-1 revealed strong spatial biases in the first response, which resulted in highly unbalanced test sets for first- versus last- reported items.

To hold visual stimulation roughly constant across all conditions, 6 out of 8 positions contained a square stimulus on every trial. For set size 1 trials (50%), 1 of the 6 squares was colored, 5 squares were gray, and the pre-cue pointed to the to-be-remembered location with 100% validity. For set size 3 trials (25%), 3 of 6 squares were colored and 3 were gray. For set size 6 trials (25%), all 6 squares were uniquely colored.

Participants performed trials in blocks of 64 trials, and they performed as many blocks as possible during the 4-hour session (with a maximum of 24 blocks). The number of blocks completed varied slightly depending on equipment set-up time and breaks (Mean = 22.7 blocks, SD = 1.87, Min = 19, Max = 24). Breaks and feedback were the same as in Experiment 4-1.

EEG acquisition

Continuous EEG data were recorded from 30 active Ag/AgCl electrodes mounted in an elastic cap (Brain Products actiCHamp, Munich, Germany), positioned according to the International 10-20 system (Positions: FP1, FP2, F7, F8, F3, F4, Fz, FC5, FC6, FC1, FC2, C3, C4, Cz, CP5, CP6, CP1, CP2, P7, P8, P3, P4, Pz, T7, T8, FT9, FT10, O1, O2, Oz). A ground electrode was positioned at FPz, and 2 additional active electrodes were affixed with stickers to the left and right mastoids. An additional 5 electrodes were used to examine the electrooculogram (EOG) signal: 1 bipolar pair above and below the right eye, 1 bipolar pair ~1 cm lateral to the external canthus of each eye, 1 ground electrode on the left cheek). Data were referenced online to the right mastoid and re-referenced offline to the algebraic average of the left and right mastoids. Data were acquired with an online filter (low cut-off = .01 Hz, high cut-

off = 80 Hz, slope from low- to high-cutoff = 12 dB/octave) and digitized at 500 Hz using Brain Vision Recorder (Brain Products, Munich, Germany) running on a PC with Windows 7. Impedance values were kept below ~10 kΩ.

Eye tracking

Gaze position was monitored using a desk-mounted EyeLink 1000 Plus infrared eye-tracking camera (SR Research, Ontario, Canada). Gaze position and pupil size were sampled at 1000 Hz. Calibration was performed at the beginning of every experimental block. Eye tracking data were usable for 26 of 29 participants in Experiment 4-1 and 27 of 28 participants in Experiment 4-2.

Artifact Rejection

Trial epochs were extracted from continuous data, and trial epochs were rejected if any artifacts were detected (eye movements, blinks, blocking, muscle noise, signal drift). Trial epochs were flagged based on several automatic criteria. After automatic detection, the data were visually inspected to confirm. If eye tracking data were available, these data were the primary means of rejecting eye movements and blinks. If not of sufficient quality, we instead relied on the EOG signal.

Eye movements. We used a sliding window step-function to check for eye movements in the HEOG and the eye-tracking gaze coordinates. For eye-tracking data, we applied a sliding window step function to the gaze position (window size = 80 ms, step size = 10 ms, threshold = 0.6º). If the difference in average gaze position from the first half of the window to the second half of the window exceeded a threshold of 0.6 degrees (measured as the linear distance between [x1, y1] and [x2, y2]), this epoch was flagged as an eye movement. For HEOG rejection, we also

used a split-half sliding window approach (window size = 150 ms, step size = 10 ms, threshold = 20 μV).

Blinks. In the eye tracking signal, trials were flagged if absolute gaze values exceeded the dimensions of the screen or were missing (i.e., pupil not detected). In the VEOG signal, we used a sliding window step function to check for blinks in the VEOG (window size = 150 ms, step size = 10 ms, threshold = 30 μV).

Drift, muscle artifacts, and blocking. We checked for drift (e.g., skin potentials) by comparing the absolute change in voltage from the first quarter of the trial to the last quarter of the trial. If the change in voltage exceeded 100 μV, the trial was rejected for drift. In addition to slow drift, we checked for sudden step-like changes in voltage with a sliding window (window size = 250 ms, step size = 20 ms, threshold = 100 μV). We excluded trials for muscle artifacts if any electrode had peak-to-peak amplitude greater than 200 μV within a 16 ms time window. We excluded trials for blocking if any electrode had at least 30 time-points in any given 200-ms time window that were within 1μV of each other.

Event-related potentials

To demonstrate that displays were visually balanced, we analyzed the visually-evoked response to the cue onset and to the memory array onset from an average of occipital and parieto-occipital electrode sites (O1, Oz, O2, PO3, PO4, PO7, PO8). Trials were baselined to a 300 ms epoch preceding the cue onset and then averaged. Time-windows for measuring P1 and N1 amplitude were defined relative to the group average waveform (collapsed across all conditions within each experiment). We used a 40 ms window (20 ms on either side of the component peak) for each of these two components. Trial times are listed relative to the memory

array onset, with the cue starting at -300 ms. This method yielded time windows that were highly consistent across experiments 4-1 / 4-2 for three of the four components (Cue P1: -154 / -146 ms; Cue N1: -110 ms / -110 ms; Memory Array N1: 182 ms / 184 ms). The Memory Array P1 could not be adequately estimated in Experiment 4-1 because it overlapped substantially with the ERP activity related to cue processing (The cue-to-stimulus interval was not jittered). As such, we analyzed three visually evoked components to assess visual balance of displays, using common time-windows of -170 to -130 ms for the Cue P1, -130 ms to -90 ms for the Cue N1, and 160 to 200 ms for the Memory Array N1.

Time-frequency analysis

We calculated power in the alpha-band (8 -12 Hz) by first band-pass filtering raw, baselined EEG data ("eegfilt.m", Delorme & Makeig, 2004) then applying a Hilbert transform (MATLAB Signal Processing Toolbox). Instantaneous power was calculated as the squared real magnitude of the complex analytic signal for each trial. In analyses of overall alpha-band suppression, we averaged alpha power across trials and then calculated percent change in alpha power (baselined from 400 to 100 ms before memory array onset). We measured this load-dependent alpha power suppression from an average of occipital and parieto-occipital electrodes (O1, Oz, O2, PO3, PO4, PO7, PO8) during the entire delay period (300 to 1300 ms).

Inverted Encoding Model

Following prior work (Foster et al., 2016), we applied an inverted encoding model to reconstruct spatially selective channel-tuning functions (CTFs) from the multivariate distribution of oscillatory power across electrodes. We assumed that the power at each electrode reflects the weighted sum of eight spatially selective channels, each tuned for a different angular location (Brouwer & Heeger, 2009; Foster et al., 2016; Sprague et al., 2015; Sprague & Serences, 2013).

These information channels are assumed to reflect the activity of underlying neuronal populations tuned to each location channel.

We modeled the response profile of each spatial channel across angular locations as a half sinusoid raised to the seventh power:

$$R = \sin(0.5\theta)^7,$$

where θ is angular location (0–359°), and *R* is the response of the spatial channel in arbitrary units. This response profile was circularly shifted for each channel such that the peak response of each spatial channel was centered over one of the eight location bins. These 8 location bins each spanned 45° (centered on 0° to 315°).

We partitioned our data into independent sets of training data and test data (see section, "Assignment of trials to training and test sets"), and applied an IEM routine at each time-point. This routine proceeded in two stages (training and test). In the training stage, training data B_1 were used to estimate weights that approximate the relative contribution of the eight spatial channels to the observed response measured at each electrode. Let B_1 (*m* electrodes × n_1 observations) be the power at each electrode for each measurement in the training set, C_1 (*k* channels × n_1 measurements) be the predicted response of each spatial channel (determined by the basis functions) for each measurement, and *W* (*m* electrodes × *k* channels) be a weight matrix that characterizes a linear mapping from "channel space" to "electrode space". The relationship between B_1, C_1, and *W* can be described by a general linear model of the form:

$$B_1 = WC_1$$

We obtained the weight matrix through least-squares estimation:

$$\widehat{W} = B_1 C_1^{T} \left(C_1 C_1^{T}\right)^{-1}$$

In the test stage, we inverted the model to transform the observed test data B_2 (m electrodes × n_2 observations) into estimated channel responses, C_2 (k channels × n_2 measurements), using the estimated weight matrix, $\widehat{W}$, that we obtained in the training phase:

$$\widehat{C_2} = \left(\widehat{W}^T\widehat{W}\right)^{-1}\widehat{W}^T B_2$$

Each estimated channel response function was then circularly shifted to a common center by aligning the estimated channel responses to the channel tuned for the remembered location (yielding a single CTF averaged across the eight location bins).

Because the exact contribution of each spatial channel to each electrode (i.e., the channel weights, *W*) varies across participants, we applied the IEM routine separately for each participant. This approach allowed us to disregard differences in the how location-selective activity is mapped to scalp-distributed patterns of power across participants, and instead focus on the profile of activity in the common stimulus or "information" space (Foster et al., 2016; Foster, Sutterer, Serences, Vogel, & Awh, 2017; Sprague et al., 2015).

Assignment of trials to training and test sets

Artifact-free Set Size 1 trials were partitioned equally into three independent sets to be used as training and test data for the IEM procedure (see Inverted Encoding Model). To analyze Set Size 1 trials, 2/3 of Set Size 1 data served as the training set, and 1/3 served as the test set. The same training set (2/3 of Set Size 1 data) was used to estimate the models for multi-item arrays. For analyses based on set size, we randomly chose 1 remembered item from each multi-item trial. For analyses based on response order, we tested multi-item trials as a function of item type (e.g., with the same training set, label test trials based on first reported item location, then label trials based on the second through *n* reported items, then label locations of "foil items" and empty locations). Figure 4-2 illustrates a schematic of training and testing sets for within-array

analyses. In Experiment 4-1, we down-sampled the data so that the training set contained an equal number of trials from each location bin (with each of the three set size 1 blocks containing an equal number of trials). Because of the behavioral bias in location chosen as a function of response number, test sets were allowed to have different numbers of trials across locations. Even with unbalanced test sets, 6 subjects could not be used because of extremely biased behavior, leaving 23 subjects in the response-based analyses in Experiment 4-1. In Experiment 4-2, we down-sampled the data so that each location bin in 1 training block had an equal number of trials as each location bin in the test set. For each training block and test set, we averaged power across trials for each location bin then applied the IEM routine. The resulting CTFs were averaged across each test set.

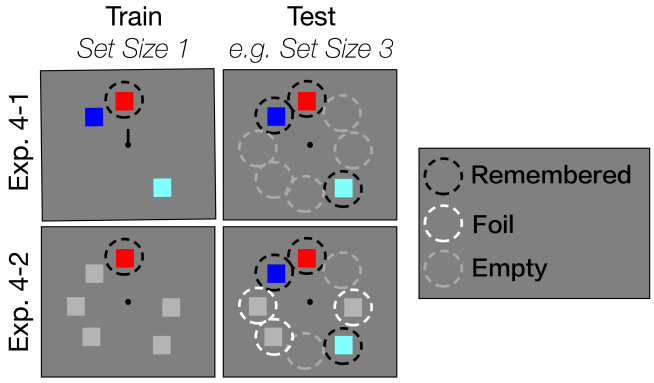

Figure 4-2 Schematic of training and testing for Experiments 4-1 and 4-2.
In both experiments, the inverted encoding model (IEM) was trained on 2/3 of set size 1 data. The IEM model was then tested on trials from multi-item memory arrays. Using this trained model, trials were then separately tested to examine position-specific alpha-band activity related to the positions of memory items, foil items, or empty locations.

Results

Working memory performance

Participants remembered around 2.5 – 3 items on average, consistent with similar visual working memory tasks (Adam et al., 2015; Adam & Vogel, 2017). Average performance was

quantified as the mean number of correctly reported items for each set size. Mean number correct (Figure 4-3) increased across set sizes in both Experiment 4-1, $F(1.02, 28.62) = 715.01$, $p < .001$, $\eta_p^2 = .96$, and Experiment 4-2, $F(1.17, 31.45) = 405.1$, $p < .001$, $\eta_p^2 = .94$. Accuracy as a function of response number closely matched previous results (Adam & Vogel, 2017), with accuracy falling as a function of response number (Table 4-1, Figure 4-4), reaching near chance levels by responses 5 and 6 for Set Size 6 arrays.

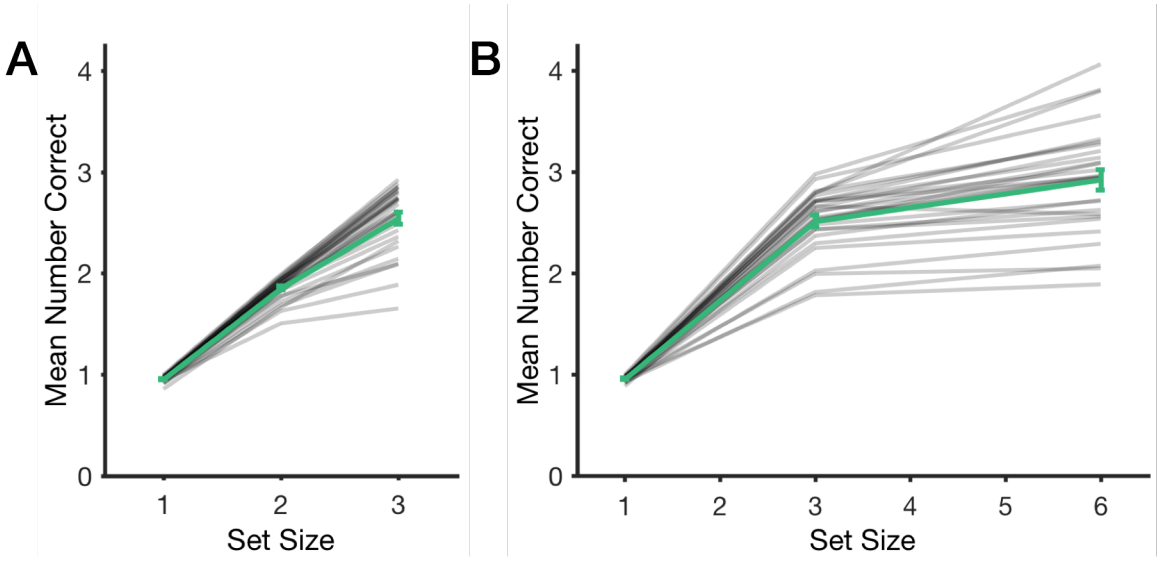

Figure 4-3 Mean number correct as a function of set size.
(A) Results from Experiment 4-1. (B) Results from Experiment 4-2. Error bars represent 1 Standard Error of the Mean (SEM).

Table 4-1 Change in accuracy as a function of response number

Condition	*df1*	*df2*	*F*	*p*	η_p^2
Exp 4-1, Set Size 2	1	28	10.40	.003	.27
Exp 4-1, Set Size 3	1.15	32.28	33.08	<.001	.54
Exp 4-2, Set Size 3	1.10	29.77	42.87	<.001	.61
Exp 4-2, Set Size 6	2.03	54.82	405.85	<.001	.94

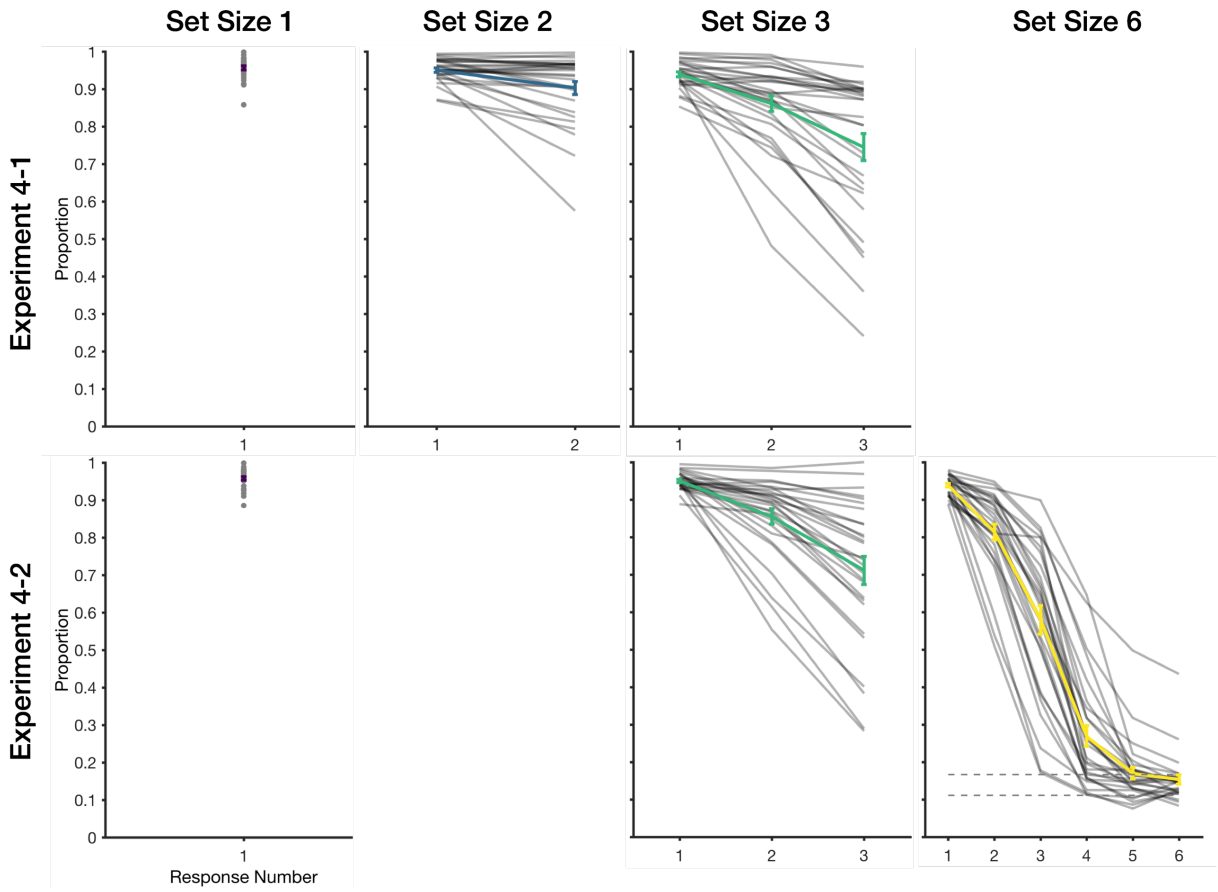

Figure 4-4 Accuracy as a function of response number.
(Top) Results from Experiment 4-1. (Bottom) Results from Experiment 4-2. Thin grey lines represent individual participants. On the Set Size 6 panel, dashed gray lines represent complete chance performance (1/9) and informed guessing (e.g., remember 3 items, guess among 1/6 remaining colors).

Spatial bias in the pattern of responses

In Experiment 4-1, participants were allowed to report the items in any order that they wished. By allowing participants to respond freely, we could take advantage of participants' robust metaknowledge of the quality of items held in mind. As shown in Figure 4-4, participants were most frequently accurate for the early responses. Prior work has shown that this change in performance as a function of response number cannot be explained by output interference alone

(Adam & Vogel, 2017; Adam et al., 2017). Rather, participants' response order closely tracks the hypothesized quality of the mnemonic representation.

In Experiment 4-1, there was a robust bias in the spatial position chosen as a function of response number for both Set Size 2, $F(2.60, 57.29) = 39.56$, $p < .001$, $\eta_p^2 = .64$, and Set Size 3, $F(2.98, 65.48)$, $p < .001$, $\eta_p^2 = .60$ (significant interaction between Response Number and Position in a repeated measures ANOVA for each set size). Participants were biased to choose the upper left locations first (90-180°) and to choose the bottom right locations last (270-360°; Figure 4-5). Because of this strong behavioral bias, some participants had few or no trials in some position bins for each response number (e.g., 0 trials in the 315° position bin for Response 1). For these participants, it was not possible to perform an IEM separately for each response (see "CTF selectivity as a function of response order").

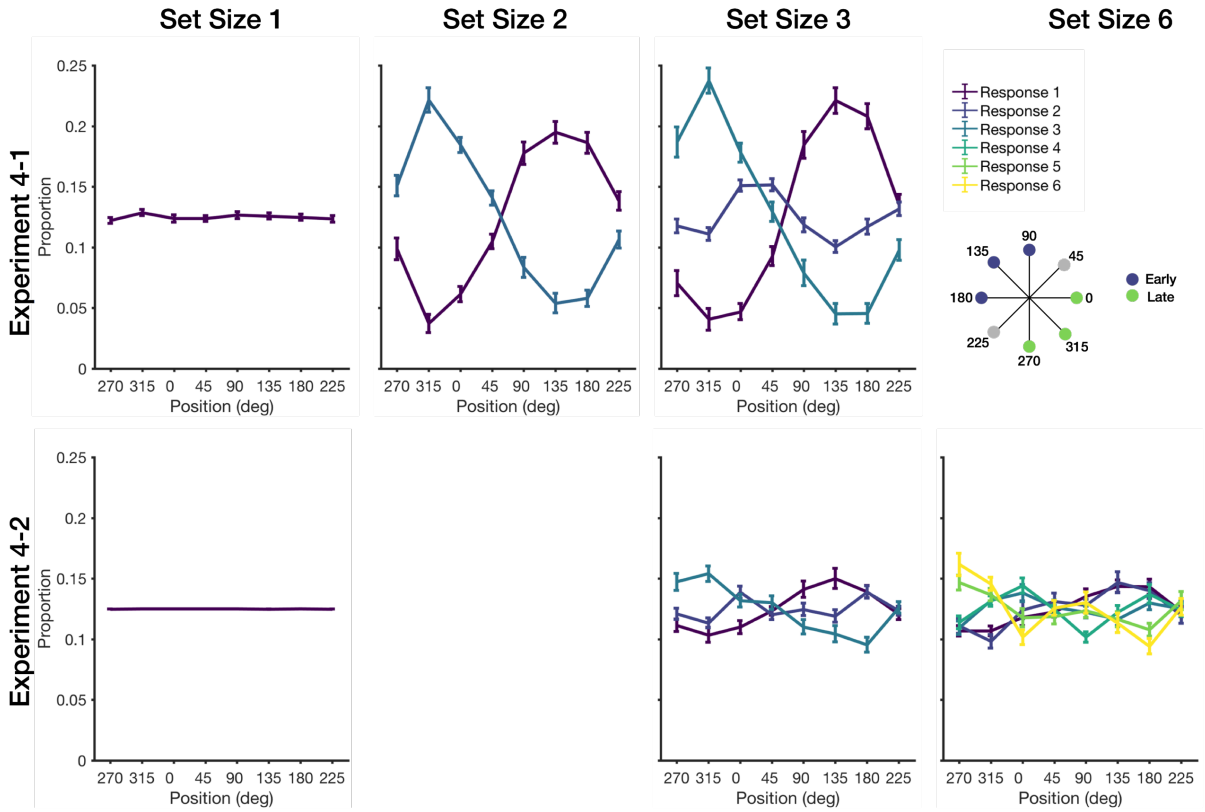

Figure 4-5 Position chosen as a function of response number.
(Top) Behavioral bias in Experiment 4-1. Participants were strongly biased to first report items in the upper left. (Bottom) Behavioral bias in Experiment 4-2. Modifications to the task significantly attenuated the behavioral bias.

In Experiment 4-2, we modified the task procedures to attempt to reduce the spatial bias for early versus late responses. To do so, participants were cued to a single location prior to the onset of the memory array. This cued location was fully counterbalanced within each block of trials. Participants were instructed that they should prioritize the cued item. For Set Size 1 arrays, the cued item was always the only memory item (i.e., the cue was 100% valid). For Set Size 3 and 6 arrays, participants made responses to all items, but were encouraged to prioritize and report the cued item first. Despite the counter-balanced prioritization cue, there was still a small spatial bias for the early versus late responses in both Set Size 3, $F(3.72, 100.33) = 7.74$, $p < .001$, $\eta_p^2 = .22$, and Set Size 6, $F(35, 945) = 5.19$, $p < .001$, $\eta_p^2 = .16$. However, this spatial bias

was significantly attenuated relative to Experiment 4-1. To calculate this, we calculated a "spatial bias score" for the first response of the Set Size 3 condition in both experiments (bias = proportion 90-180º – proportion 270-360º). An ANOVA revealed that spatial bias scores were significantly higher for Experiment 4-1 than for Experiment 4-2, $F(1, 55), = 42.9, p < .001, \eta_p^2 = .44$. Although we successfully reduced behavioral bias, there was no significant difference in Set Size 3 accuracy for Experiment 4-1 versus 4-2, $F(1, 55), = .12, p = .73, \eta_p^2 = .002$. Thus, we changed participants' spatial biases without changing working memory performance.

Load-dependent alpha-power modulation when controlling for visual stimulation

We found typical effects of memory load on oscillatory alpha power (Figure 4-6). Consistent with previous work, we found that alpha power was suppressed for high loads relative to low memory loads in both Experiment 4-1, $F(1.59, 44.55) = 21.52, p < .001, \eta_p^2 = .44$, and 4-2, $F(1.29, 34.70) = 14.77, p < .001, \eta_p^2 = .35$. This result is an important demonstration of the robustness of load-dependent alpha suppression. In most prior work demonstrating changes to alpha power suppression with memory load, the increased load has been confounded with increased visual input (e.g., Fukuda, Kang, & Woodman, 2016; Fukuda, Mance, & Vogel, 2015). This is potentially problematic if imbalances in sensory information have lasting effects on alpha power that propagate through later in the trial.

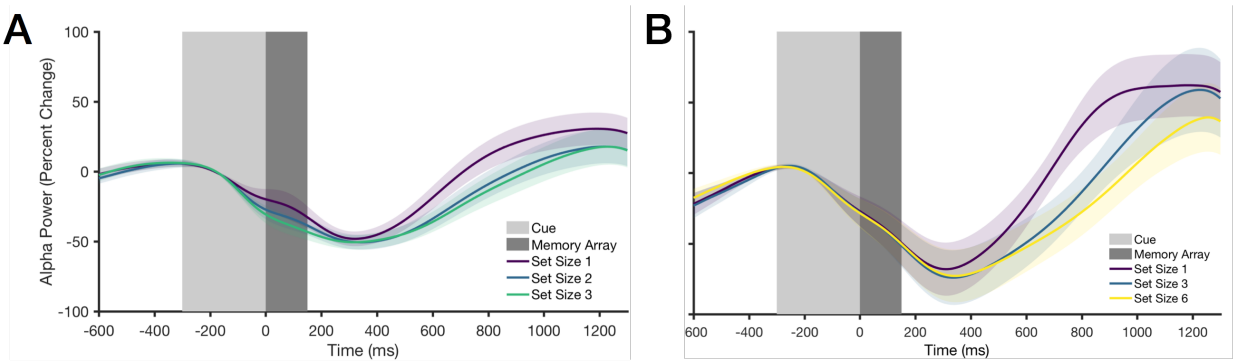

Figure 4-6 Alpha power as a function of memory load.
Alpha power decreased as a function of load in Experiment 4-1 (A) and 4-2 (B). Light grey and dark grey represent the cue and memory array durations, respectively. Shaded error bars represent 1 SEM.

To verify that visual stimulation was controlled for (as we intended), we examined modulation of visually-evoked potentials as a function of set size. Event-related potentials are shown in Figure 4-7. In Experiment 4-1, participants always viewed 3 items occupying 3 out of 8 fixed locations. To manipulate memory load, subjects were cued to remember 1, 2, or 3 of these three items. The cue-related P1 component did not differ by set size ($p = .18$), but the cue-related N1 component increased as a function of set size, $F(1.58, 44.34) = 35.93$, $p < .001$, $\eta_p^2 = .56$. However, as expected, there was no difference in the N1 component evoked by the identical memory array ($p = .58$) as a function of load. In Experiment 4-2, participants always viewed 6 items occupying 6 out of 8 fixed locations. To manipulate memory load, the subjects were instructed to selectively remember colored squares and ignore gray placeholder squares. This design was highly effective at removing any physical differences between memory conditions. As expected, there was no effect of memory load on the cue-related P1, cue-related N1, or the array-related N1 (all p's $> .14$). Thus, any effects that we observe on fidelity of CTFs as a function of set size are not due to an increase in noise due to viewing an increased number of items (e.g., Reddy, Kanwisher, & VanRullen, 2009).

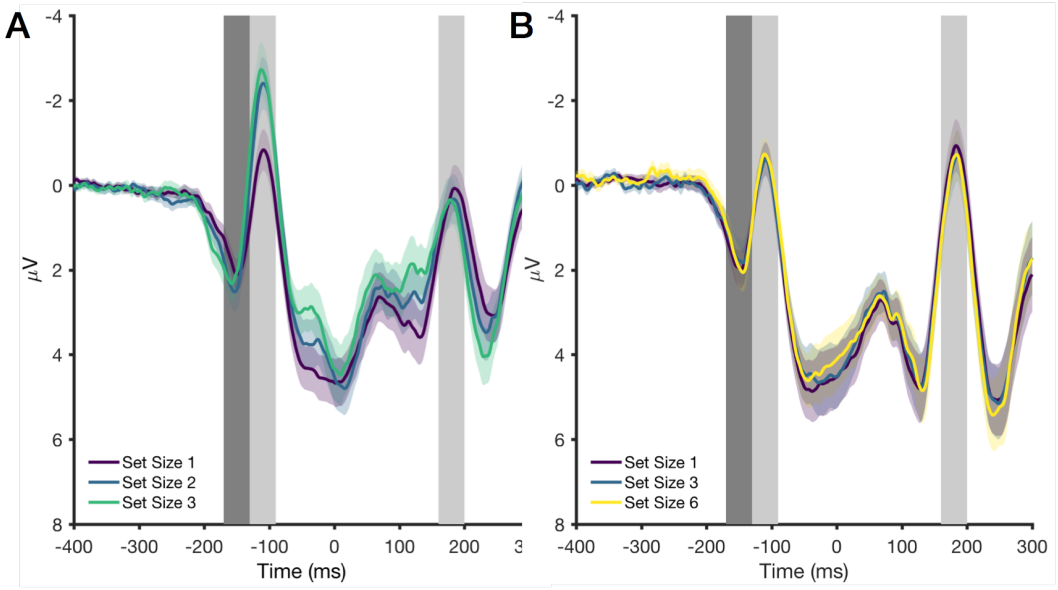

Figure 4-7 Event-related potentials as a function of memory load.
Visual-evoked potentials in Experiment 4-1 (A) and 4-2 (B). Dark grey represents the time-period used to calculated the cue-related P1. Light grey represents the time-period used to calculate the cue- and array-related N1. Times on the x-axis are relative to memory array onset at 0 ms. The spatial pre-cue onset is at – 300 ms. Shaded error bars represent 1 SEM.

CTF selectivity as a function of set size

Consistent with previous work (Buschman et al., 2011; Emrich et al., 2013; Sprague et al., 2014), we were able to decode information about a randomly chosen item for set sizes 1, 2, and 3, and the quality of decoding decreased as a function of set size. Figure 4-8 shows CTF selectivity as a function of set size in Experiments 4-1 and 4-2. The location of an average item from the display could be decoded in a continuous fashion for Set Sizes 1 and 2 (Figure 4-8A), whereas decoding was more sporadic for Set Size 3 and completely absent for set size 6 (Figure 4-8B). In both experiments, CTF selectivity significantly declined as a function of memory load, $F(2,56) = 23.82$, $p < .001$, $\eta_p^2 = .46$ in Experiment 4-1, and $F(1.54, 41.68) = 39.15$, $p < .001$, $\eta_p^2 = .59$ in Experiment 4-2. Post-hoc tests that all set sizes were significantly different from one another in both Experiment 4-1 ($p_{bonferroni} <= .003$) and 4-2 ($p_{bonferroni} < .001$). To quantify the ability

to detect information for each set size separately, we performed one-sample *t*-tests of average delay period decoding, as shown in Table 4-2.

Set sizes 1 through 3 could be decoded robustly, whereas set size 6 did not exceed chance. This analysis suggests, at face value, that we can replicate past results (decoding up to 3 items) but that this approach fails when faced with a supra-capacity array of 6 items. However, in this analysis we examined our ability to decode information about a *randomly chosen* item. Critically, capacity-limited models predict that participants do not store all items from arrays with more than 3-4 items. With an average capacity of 3 items, we would expect that we are feeding in noise to the analysis on around 50% of trials when attempting to decode information about a 6 item display. Such an approach may be misleading by inflating the rate of noise disproportionately for the 6 item displays relative to 1, 2, and 3 item displays. To circumvent this issue, we can instead separately analyze each item from the display as a function of hypothesized mnemonic quality. We predicted that we should be able to decode location information about items that are recalled correctly (e.g., first 3 responses made in the whole-report task), but that we should not be able to decode location information about items that are not stored.

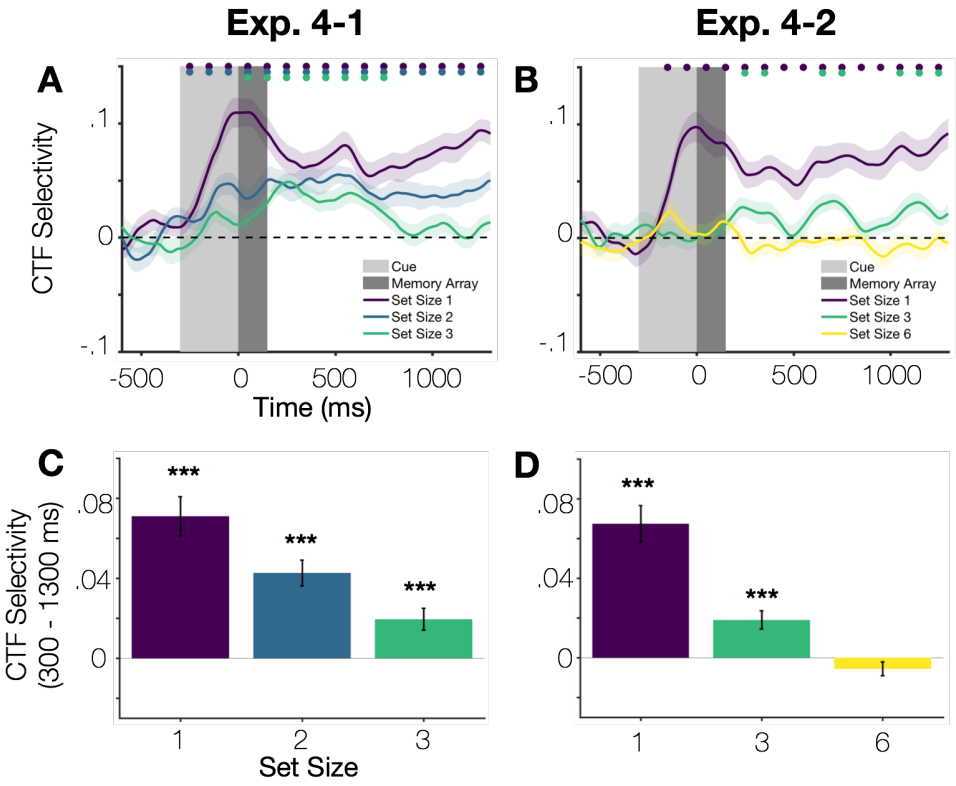

Figure 4-8 CTF Selectivity as a function of memory load.
The top panels illustrate CTF selectivity over time for Experiment 4-1 (A) and 4-2 (B). Dots represent significance above baseline ($p < .05$, 100 ms bins). The bottom panels show average CTF selectivity for the entire delay period (300 – 1300 ms) in Experiment 4-1 (C) and 4-2 (D). Error bars represent 1 SEM, *** represents $p<=.001$.

Table 4-2 One-sample t-test for average delay period CTF selectivity in Experiment 4-1 and 4-2

Condition	*df*	*t*	$p_{uncorrected}$	*Cohen's d*
Exp 4-1, Set Size 1	28	7.30	<.001	1.35
Exp 4-2, Set Size 1	27	7.27	<.001	1.37
Exp 4-1, Set Size 2	28	6.63	<.001	1.23
Exp 4-1, Set Size 3	28	3.55	.001	.66
Exp 4-2, Set Size 3	27	4.16	<.001	.79
Exp 4-2, Set Size 6	27	-1.6	.12	-.30

CTF selectivity as a function of response order

Rather than choosing one random item to examine from each set size, we instead estimated CTF selectivity separately for each position in the array. In previous work employing

whole-report working memory tasks, participants consistently reported their "best" items first (Adam et al., 2015, 2017; Adam & Vogel, 2017). This was replicated in the behavioral findings from Experiments 4-1 and 4-2; subjects consistently had higher accuracy for items reported earlier in the response period compared to later. Thus, we ran the CTF routine separately for each response position. To do so, we would use the same training set (Set Size 1 memory trials) and the same test set (e.g., Set Size 3 trials), but we changed the labels for the test set as a function of item type (e.g., labels for the 1st reported item in Set Size 3 versus labels for the 3rd reported item in Set Size 3). Because of the strong behavioral bias in Experiment 4-1, we could not analyze the response-level data for some participants (i.e., they had 0 trials for Response 1 in the lower right position bin). After excluding these participants, the total sample size for response-level analyses for Experiment 4-1 was 23 subjects.

In addition to each response position, we also analyzed the positions of (1) Foil items and (2) Empty positions. When examining multiple locations from a single array, the Empty position CTF provides an important empirical baseline for hypothesis testing. For example, imagine you always remember a location at 0 degrees and but examine the CTF for the perfectly opposite empty position at 180 degrees. Because you really are remembering 0 degrees, the information channel for 0 degrees will yield the most robust response. But, when you average all location channels and shift them to be centered at 180 degrees, rather than the true remembered location of 0 degrees, this will yield a strongly inverted tuning function (i.e., negative selectivity). Because the spatial position with the remembered item cannot be both remembered and empty, the average of the "empty position" CTF is slightly shifted away from to the remembered item(s).

This item-by-item analysis revealed that the absence of decoding for a random item from the Set Size 6 condition was an artifact of injecting noise into the analysis by assuming that all items are remembered. Because unremembered items that are slightly shifted away from remembered items yield negative selectivity, choosing a random item from supra-capacity arrays yielded an overall CTF selectivity index close to 0. When examining CTFs for each item as a function of its accuracy, we instead found robust CTFs for the prioritized items (i.e., 1st and 2nd responses) for all set sizes (2, 3, and 6). The first and second reported items were reconstructed in a sustained fashion, even for set size 6 trials (Figure 4-9) whereas the later reported items were sporadically or negatively reconstructed. To quantify change in selectivity with item type, we ran a separate repeated measures ANOVA with Item Type entered as a factor (Response 1 – n, Foil, Empty). We used planned simple contrasts to compare each item to an empirical baseline ("Empty," attempting to decode a location where no item was decoded). Results are depicted in Figure 4-10 and statistics are found in Tables 4-3 and 4-4. Again, we found robust decoding of up to 2 remembered locations in all multi-item arrays (2, 3, and 6 items) in both Experiments 4-1 and 4-2. For set size 6, we found decoding of the first 3 response positions but no evidence of the representation of location information for responses 4 through 6. In fact, we found that the 5th reported item yielded significantly negative slopes (likely because in this free-report paradigm, the 5th response position was nearly always spatially opposite from the first response).

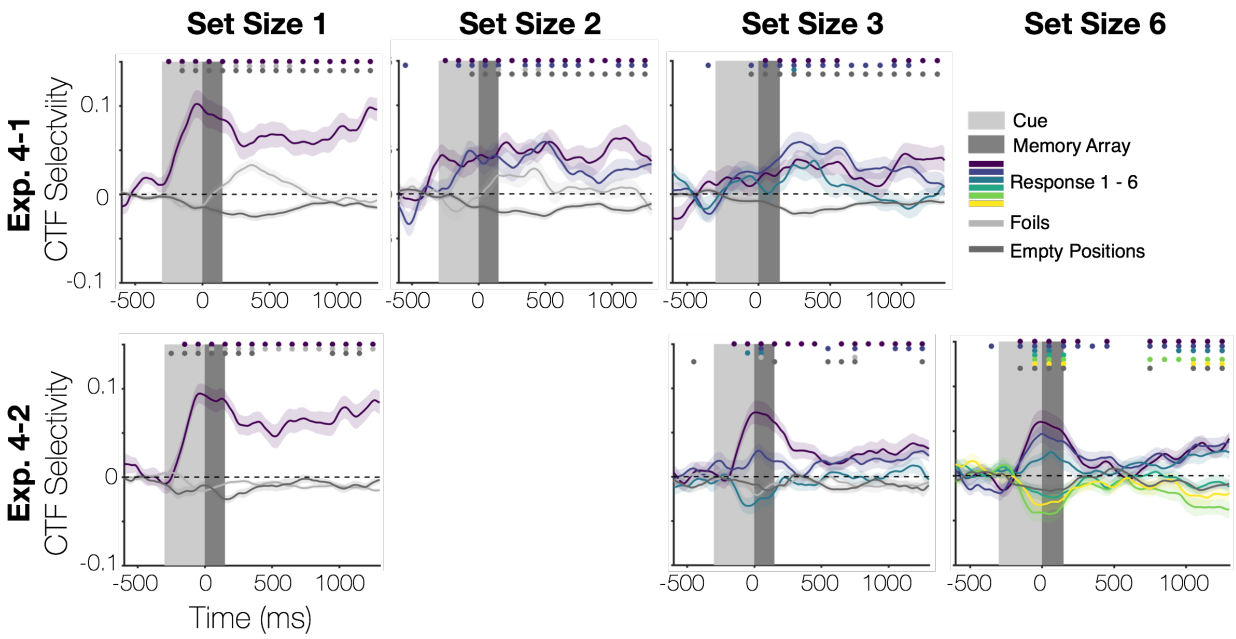

Figure 4-9 CTF selectivity as a function of time and response order.
Average CTF selectivity over time in Experiment 4-1 (Top) and Experiment 4-2 (Bottom). Decoding was robust for the first reported item (most prioritized) and decreased as a function of response order.

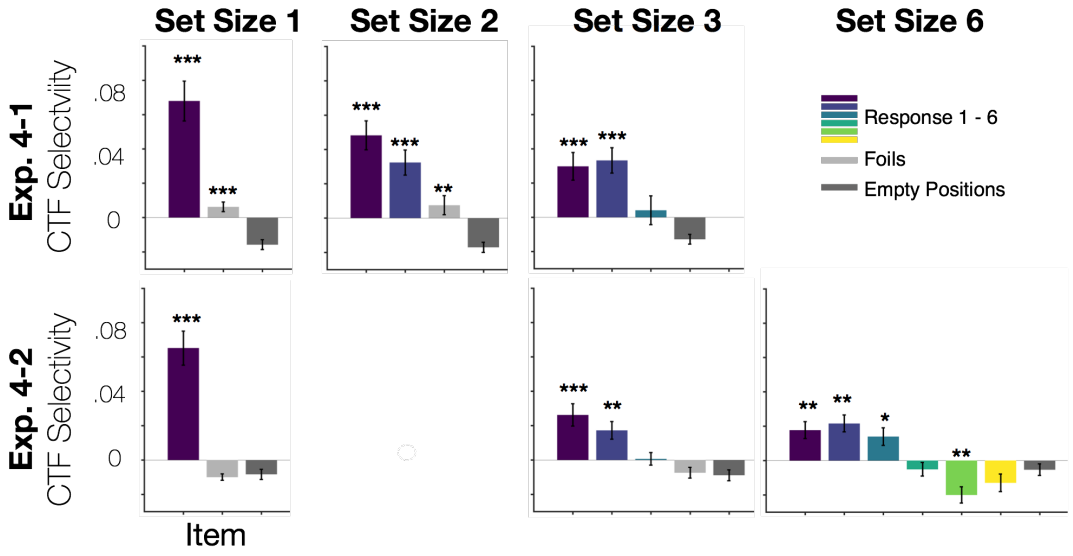

Figure 4-10 Average CTF selectivity during the delay period (300 -1300 ms).
(Top) Average CTF selectivity in Experiment 4-1. (Bottom) Average CTF selectivity in Experiment 4-2. Markers of significance follow simple contrasts comparing each item type to an empirical control (decoding of empty positions in the array, ***p<=.001, ** p<.01, * p<.05). Across experiments, the first and second reported items could be consistently decoded from multi-item arrays.

Table 4-3 Planned simple contrasts for Experiment 4-1

Item Type	*Set Size*	*F*	*p*	η_p^2
Foil	Set Size 1	18.48	<.001	.46
	Set Size 2	10.22	.004	.321
Response 1	Set Size 1	34.01	<.001	.61
	Set Size 2	36.01	<.001	.62
	Set Size 3	18.43	<.001	.46
Response 2	Set Size 2	24.77	<.001	.53
	Set Size 3	21.66	<.001	.50
Response 3	Set Size 3	3.01	.097	.12

Table 4-4 Planned simple contrasts for Experiment 4-2

Item Type	*Set Size*	*F*	*p*	η_p^2
Foil	Set Size 1	.19	.67	.007
	Set Size 3	.11	.74	.004
Response 1	Set Size 1	42.24	<.001	.61
	Set Size 3	18.05	<.001	.40
	Set Size 6	10.16	.001	.27
Response 2	Set Size 3	14.40	.001	.35
	Set Size 6	13.41	.001	.33
Response 3	Set Size 3	3.31	.08	.11
	Set Size 6	5.97	.02	.18
Response 4	Set Size 6	.001	.97	<.001
Response 5	Set Size 6	9.06	.006	.25
Response 6	Set Size 6	2.70	.11	.09

Discussion

In this chapter, we found that only a subset of items are encoded by the topographic distribution of alpha power (8 – 12 Hz), a neural measure that tracks the active maintenance of spatial information in working memory (Foster, Bsales, et al., 2017; Foster et al., 2016; Sutterer et al., in prep). This result is consistent with item-limited models of visual working memory capacity, which propose that observers store a subset of items from large arrays (e.g., 3 out of 6) and fail to store items beyond this capacity. This work demonstrates the feasibility of new methods for decoding the contents of working memory in an item-specific fashion. Prior work

examining load-dependent changes to information content has been limited to interpreting the average representation of items from a multi-item array (Buschman et al., 2011; Emrich et al., 2013; Sprague et al., 2014). Here, we offer a step forward by examining how representational quality of the neural code varies across individual items within an array.

Decoding approaches bolster behavioral evidence for item limits

The present results provide strong traction against equal-precision models of visual working memory capacity (Bays, 2015; Bays et al., 2009; Bays & Husain, 2008). Such models predict that working memory resources are allotted equally among all items. By this account, an increase in neural noise among competing representations leads to a decrease in the fidelity of working memory behavioral performance (Bays, 2015). In stark contrast to this prediction, we found that prioritized items (e.g., response 1 and 2) from a 6 item array are not qualitatively different in fidelity from prioritized items in a 3 item array. Thus, whereas equal-precision models suggest that neural fidelity of *every* item decreases when memory load is increased from 3 to 6 items, the current data suggest that this model's prediction is an artifact of averaging across high-quality items in a prioritized state with noise from items that were not stored. Thus, the present results are consistent with capacity-limited models of working memory (Luck & Vogel, 1997; Zhang & Luck, 2008), especially recent instantiations which propose that while capacity is limited, the allocation of mnemonic resources can vary dynamically across items and trials (Adam et al., 2015; Adam, Robison, & Vogel, 2018; van den Berg et al., 2014).

Assessing the role of additional working memory mechanisms

One important caveat, of course, is that the active mnemonic processes tracked by alpha-band representations do not represent the sum total of all possible working memory signals. An obvious example of this is feature-specificity; participants remembered both color and special

location, but alpha-band representations only track the remembered location of items. In addition, recent models of working memory have proposed that representations may be held in an "activity silent" state by, for example, rapidly shifting synaptic weights (Rose et al., 2016; Stokes, 2015; Wolff, Jochim, Akyürek, & Stokes, 2017, but also see: Christophel, Iamshchinina, Yan, Allefeld, & Haynes, 2018). Analyses of alpha power alone cannot speak to the information content of other active signals or of hypothesized representations that are in a latent state. Yet, measurement of alpha-band representations still provides useful insights into the allocation of working memory resources. Alpha-band representations are thought to play a pivotal role in the maintenance of visual information, even when location is irrelevant (Foster, Bsales, et al., 2017), and recent work has shown that alpha-band representations are behaviorally relevant, predicting changes in the successful deployment of spatial attention (Foster, Sutterer, et al., 2017) and the reinstatement of remembered locations from long-term memory into working memory (Sutterer, Foster, Serences, Vogel, & Awh, 2018). Thus, even if other hypothetical active signals were measurable for all items in the array, the measurable asymmetry of alpha-band representations across items would still suggest that some items are more robustly represented than others.

Future directions

The present results offer a step forward by directly demonstrating that a key marker of working memory maintenance asymmetrically encodes the positions of items from a multi-item array (i.e., some items are decodable from the topographic distribution of alpha power and other items are not). However, the interpretation of the present results is somewhat limited by the poor spatial resolution of EEG. Here, we are unable to determine which cortical areas subserve the active maintenance of spatial information in the alpha-band. Future work applying single-item decoding techniques to new and existing fMRI data will be critical for understanding the

complex interrelationship between attention and active working memory signals, and how the information content of multivariate signals is instantiated across sensory and control regions (Christophel et al., 2018).

CHAPTER 5. GENERAL CONCLUSIONS

Visual working memory is a severely limited mental resource used to maintain and manipulate information in the service of behavior. While researchers broadly agree that there are constraints on the amount of information that may be held actively in mind, there is active debate about the nature of working memory's limits (Adam et al., 2017; Bays, 2018; van den Berg et al., 2014). In this dissertation, I presented experiments testing key features of visual working memory and its limits. In Chapter 2, I questioned the assumption that individual differences in working memory capacity represent differences in the maximum number of items that may be stored. Instead, modeling of trial-by-trial working memory performance revealed that a majority of subjects (~85%) shared a common capacity and differed in the consistency with which they filled this capacity. In Chapter 3, I used a fine-grained, continuous measure of visual working memory to strongly test competing accounts of the structure of visual working memory – do participants guess when capacity limits are exceeded, or is working memory performance better explained by resources that are shared among all items? By measuring the quality of all items held in mind and comparing these responses to hypothesized models of guessing, I demonstrated clear behavioral evidence for capacity limits in visual working memory. Finally, in Chapter 4, I combined a multivariate encoding model with item-specific behavior in order to identify whether and how capacity limits are borne out in active neural signals during the delay period. Below, I briefly discuss open questions related to these findings.

The complex interrelationship between attention and working memory processes

From the experiments presented in Chapter 2, and related work in the literature, it is clear that fluctuations of attention affect working memory performance and subsequent recall from

long-term memory (Adam et al., 2015; Aly & Turk-Browne, 2016; deBettencourt, Norman, & Turk-Browne, 2017). The processes of spatial attention and working memory have been proposed to have a very tight link (Awh & Jonides, 2001; Chun, 2011; Postle, 2006), with some recent work proposing that neural measures of working memory storage actually reflect the focus of attention within working memory (e.g., Berggren & Eimer, 2016). Further work is need to delineate the often fine-grained differences between hypothesized attention and working memory components of processing within working memory tasks. For example, recent work by Sheremata and colleagues (Sheremata et al., 2018) revealed that IPS is more robustly recruited when a task contains both storage and spatial attention demands compared to attention demands alone, corroborating the role of elevated IPS activity in the maintenance of information in WM. Future experiments exploring similar, fine-grained behavioral dissociations will lead to new insights about the complementary contributions of item maintenance and attention-based rehearsal to working memory performance.

Identifying sources of variable precision

Chapter 3 revealed systematic variability among remembered items; the fidelity of behavioral responses declined monotonically as a function of response order, and this decrease in precision could not be explained by simple output interference. Instead, it appears that subjects had rich, precise metaknowledge of the quality of stored items. Previous work has similarly demonstrated that participants have robust metaknowledge of the quality of items held in working memory (Fougnie et al., 2012; Rademaker et al., 2012). However, there is ongoing debate about the extent to which this variability in quality represents true mnemonic precision versus incidental stimulus features (e.g., the oblique effect, Appelle, 1972; Pratte et al., 2017). For example, color categories strongly bias responses during both perception and memory (Bae

et al., 2015, 2014). Recent work by Hardman and colleagues (Hardman et al., 2017) used a model-based approach to try to dissociate such "categorical" responses from truly continuous responses. Intriguingly, they found that the prevalence of categorical information increases as a function of memory load. This leaves open the possibility that only the most prioritized items (or even a single prioritized item) are represented in a truly continuous fashion, whereas other "accessory" memory items are stored via a coarse categorical code. An interesting open question is the relative contribution of stimulus-specific variability and true mnemonic variability to working memory performance, and how these distinct representations are encoded in the brain.

The functional role of working memory signals

An ongoing debate in the working memory literature is *where* working memory representations are coded. The *sensory recruitment hypothesis* suggests that working memory representations are coded by distributed patterns of activity in cortical areas that also support perception (Ester, Serences, & Awh, 2009; Harrison & Tong, 2009; Riggall & Postle, 2012; Serences et al., 2009). An alternative account suggests that prefrontal and parietal control representations are key (Christophel, Hebart, & Haynes, 2012; Courtney, Ungerleider, Keil, & Haxby, 1997; Todd & Marois, 2004), and that sensory areas may be too susceptible to interference from new sensory inputs to support memory (Bettencourt & Xu, 2016). In addition, others have noted that items that are in working memory, but no longer actively attended, are no longer decodable from sensory regions (LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013; Riggall & Postle, 2012).

Rather than representing mutually exclusive possibilities, it could instead be that different neuronal populations uniquely support separable aspects of successful working memory performance. For instance, Christophel et al. (2018) employed a dual retro-cue paradigm to test

where attended and unattended working memory items are represented. They replicated previous results that the unattended memory item could not be decoded from sensory cortex (V1-V4). However, they found that the unattended memory item could be robustly in frontal and parietal regions. This work calls into question the currently dominant interpretation that unattended items are held in an "activity silent" state (Rose et al., 2016; Stokes, 2015; Wolff et al., 2017), instead suggesting that information about the unattended memory item is still actively encoded in broadly-distributed patterns of neural activity.

The recent work by Christophel et al. raises an important question: If unattended items are actively represented in frontal and parietal cortical areas, then what is the purpose of "reviving" seemingly redundant neural codes in sensory areas (Riggall & Postle, 2012; Sprague, Ester, & Serences, 2016)? One possibility is that sensory representations are not redundant. Perhaps they support fine-grained representations ("continuous" memories rather than coarse "categorical" memories in the terminology of Hardman et al., 2017), whereas frontal and parietal representations can only support coarse judgements. Future work employing single-item decoding methods (Chapter 4) in conjunction with fMRI could address this important question. For example, if we hypothesize that sensory codes support fine-grained judgements, then we would predict that only the strongest-prioritized items would be represented in both sensory and fronto-parietal areas (e.g., responses 1 and 2), coarsely-represented items would be represented only in fronto-parietal areas (e.g., responses 3 and 4), and no information would be detectable for supra-capacity items (e.g., responses >4). Experiments focusing on item-specific decoding, like those presented here, would enhance our understanding of the representation of mnemonic variability and capacity limits in population codes.

Conclusions

Despite broad agreement that working memory resources are limited, there has been no consensus about the precise nature of working memory's limitations. Namely, is working memory limited by a capacity (e.g., certain number of items that can be stored) or by the distribution of resources (e.g., all items are stored, but imprecisely)? In part, consensus on this question has not been reached because of a stalemate between highly sophisticated and well-fitting models of a limited set of behavioral measures (e.g., van den Berg et al., 2014). Here, we circumvented this stalemate through a combination of novel behavioral measures, modelling approaches, and application of innovative multivariate techniques to neural data. Combined, the work presented in this dissertation supports a view of working memory that has the following characteristics: (1) A mostly-common capacity limit, whereby participants store information about a few items and guess for items beyond this capacity (2) Variability in precision among remembered items, with higher fidelity for items that are strongly prioritized, and (3) Variability in the allocation of working memory resources across trials, with the consistency of performance predicting a substantial share of individual differences in overall working memory ability.

REFERENCES

Adam, K. C. S., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The Contribution of Attentional Lapses to Individual Differences in Visual Working Memory Capacity. *Journal of Cognitive Neuroscience*, *27*(8), 1601–1616. https://doi.org/10.1162/jocn_a_00811

Adam, K. C. S., Robison, M. K., & Vogel, E. K. (2018). Contralateral Delay Activity Tracks Fluctuations in Working Memory Performance. *Journal of Cognitive Neuroscience*, 1–12. https://doi.org/10.1162/jocn_a_01233

Adam, K. C. S., & Vogel, E. K. (2017). Confident failures: Lapses of working memory reveal a metacognitive blind spot. *Attention, Perception, & Psychophysics*, *79*(5), 1506–1523. https://doi.org/10.3758/s13414-017-1331-8

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

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.

Aly, M., & Turk-Browne, N. B. (2016). Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences*, *113*(4), E420–E429. https://doi.org/10.1073/pnas.1518931113

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. https://doi.org/10.1037/h0033117

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. https://doi.org/10.1111/j.1467-9280.2007.01949.x

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119–126. https://doi.org/10.1016/S1364-6613(00)01593-X

Awh, E., & Vogel, E. K. (2008). The bouncer in the brain. *Nature Neuroscience*, *11*(1), 5–6. https://doi.org/10.1038/nn0108-5

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744–763. https://doi.org/10.1037/xge0000076

Bae, G.-Y., Olkkonen, M., Allred, S. R., Wilson, C., & Flombaum, J. I. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision*, *14*(4). https://doi.org/10.1167/14.4.7

Bays, P. M. (2015). Spikes not slots: noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431–438. https://doi.org/10.1016/j.tics.2015.06.004

Bays, P. M. (2018). Reassessing the Evidence for Capacity Limits in Neural Signals Related to Working Memory. *Cerebral Cortex*, *28*(4), 1432–1438. https://doi.org/10.1093/cercor/bhx351

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

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

Berens, P. (2009). CircStat: A MATLAB Toolbox for Circular Statistics. *Journal of Statistical Software*, *31*(10). https://doi.org/10.18637/jss.v031.i10

Berggren, N., & Eimer, M. (2016). Does Contralateral Delay Activity Reflect Working Memory Storage or the Current Focus of Spatial Attention within Visual Working Memory? *Journal of Cognitive Neuroscience*, *28*(12), 2003–2020. https://doi.org/10.1162/jocn_a_01019

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

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

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and Reconstructing Color from Responses in Human Visual Cortex. *Journal of Neuroscience*, *29*(44), 13992–14003. https://doi.org/10.1523/JNEUROSCI.3577-09.2009

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

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. https://doi.org/10.1016/j.tics.2014.04.012

Christophel, T. B., Hebart, M. N., & Haynes, J.-D. (2012). Decoding the Contents of Visual Short-Term Memory from Human Visual and Parietal Cortex. *Journal of Neuroscience*, *32*(38), 12983–12989. https://doi.org/10.1523/JNEUROSCI.0184-12.2012

Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C., & Haynes, J.-D. (2018). Cortical specialization for attended versus unattended working memory. *Nature Neuroscience*, *21*(4), 494–496. https://doi.org/10.1038/s41593-018-0094-4

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, *49*(6), 1407–1409. https://doi.org/10.1016/j.neuropsychologia.2011.01.029

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1997). Transient and sustained activity in a distributed neural system for human working memory. *Nature*, *386*(6625), 608–611. https://doi.org/10.1038/386608a0

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114-185. https://doi.org/10.1017/S0140525X01003922

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition*, *34*(8), 1754–1768. https://doi.org/10.3758/BF03195936

Cowan, N., Hardman, K., Saults, J. S., Blume, C. L., Clark, K. M., & Sunday, M. A. (2016). Detection of the number of changes in a display in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 169–185. https://doi.org/10.1037/xlm0000163

Cowan, N., Saults, J. S., Elliott, E. M., & Moreno, M. V. (2002). Deconfounding Serial Recall. *Journal of Memory and Language*, *46*(1), 153–177. https://doi.org/10.1006/jmla.2001.2805

Coyle, T. (2003). A review of the worst performance rule: Evidence, theory, and alternative hypotheses. *Intelligence*, *31*(6), 567–587. https://doi.org/10.1016/S0160-2896(03)00054-0

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language*, *25*(1), 1–18. https://doi.org/10.1016/0749-596X(86)90018-5

deBettencourt, M. T., Norman, K. A., & Turk-Browne, N. B. (2017). Forgetting from lapses of sustained attention. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-017-1309-5

Deiber, M.-P., Missonnier, P., Bertrand, O., Gold, G., Fazio-Costa, L., Ibañez, V., & Giannakopoulos, P. (2007). Distinction between Perceptual and Attentional Processing in Working Memory Tasks: A Study of Phase-locked and Induced Oscillatory Brain Dynamics. *Journal of Cognitive Neuroscience*, *19*(1), 158–172. https://doi.org/10.1162/jocn.2007.19.1.158

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

Donkin, C., Kary, A., Tahir, F., & Taylor, R. (2016). Resources masquerading as slots: Flexible allocation of visual working memory. *Cognitive Psychology*, *85*, 30–42. https://doi.org/10.1016/j.cogpsych.2016.01.002

Donkin, C., Nosofsky, R., Gold, J., & Shiffrin, R. (2015). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychonomic Bulletin & Review*, *22*(1), 170–178. https://doi.org/10.3758/s13423-014-0675-5

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the Frontal Lobe: The Organization of Goal-Directed Behavior. *Cognitive Psychology*, *30*(3), 257–303. https://doi.org/10.1006/cogp.1996.0008

Duncan, J., Schramm, M., Thompson, R., & Dumontheil, I. (2012). Task rules, working memory, and fluid intelligence. *Psychonomic Bulletin & Review*, *19*(5), 864–870. https://doi.org/10.3758/s13423-012-0225-y

Elmore, L. C., Magnotti, J. F., Katz, J. S., & Wright, A. A. (2012). Change detection by rhesus monkeys (Macaca mulatta) and pigeons (Columba livia). *Journal of Comparative Psychology*, *126*(3), 203–212. https://doi.org/10.1037/a0026356

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. *Journal of Neuroscience*, *33*(15), 6516–6523. https://doi.org/10.1523/JNEUROSCI.5732-12.2013

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Engle, R. W., & Kane, M. J. (2003). Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 44, pp. 145–199). New York, NY: Elsevier.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual Differences in Working Memory Capacity and What They Tell Us About Controlled Attention, General Fluid Intelligence, and Functions of the Prefrontal Cortex. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (pp. 102–134). Cambridge: Cambridge University Press. Retrieved from http://ebooks.cambridge.org/ref/id/CBO9781139174909A014

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology. General*, *128*(3), 309–331. http://dx.doi.org/10.1037/0096-3445.128.3.309

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

Esterman, M., Noonan, S. K., Rosenberg, M., & DeGutis, J. (2013). In the Zone or Zoning Out? Tracking Behavioral and Neural Fluctuations During Sustained Attention. *Cerebral Cortex*, *23*(11), 2712–2723. https://doi.org/10.1093/cercor/bhs261

Esterman, M., Rosenberg, M. D., & Noonan, S. K. (2014). Intrinsic Fluctuations in Sustained Attention and Distractor Processing. *Journal of Neuroscience*, *34*(5), 1724–1730. https://doi.org/10.1523/JNEUROSCI.2658-13.2014

Foster, J. J., Bsales, E. M., Jaffe, R. J., & Awh, E. (2017). Alpha-Band Activity Reveals Spontaneous Representations of Spatial Position in Visual Working Memory. *Current Biology*, *27*(20), 3216-3223.e6. https://doi.org/10.1016/j.cub.2017.09.031

Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K., & Awh, E. (2016). The topography of alpha-band activity tracks the content of spatial working memory. *Journal of Neurophysiology*, *115*(1), 168–177. https://doi.org/10.1152/jn.00860.2015

Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K., & Awh, E. (2017). Alpha-Band Oscillations Enable Spatially and Temporally Resolved Tracking of Covert Spatial Attention. *Psychological Science*, *28*(7), 929–941. https://doi.org/10.1177/0956797617699167

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229. https://doi.org/10.1038/ncomms2237

Fukuda, K., Awh, E., & Vogel, E. K. (2010). Discrete capacity limits in visual working memory. *Current Opinion in Neurobiology*, *20*(2), 177–182. https://doi.org/10.1016/j.conb.2010.03.005

Fukuda, K., Kang, M.-S., & Woodman, G. F. (2016). Distinct neural mechanisms for spatially lateralized and spatially global visual working memory representations. *Journal of Neurophysiology*, *116*(4), 1715–1727. https://doi.org/10.1152/jn.00991.2015

Fukuda, K., Mance, I., & Vogel, E. K. (2015). Alpha Power Modulation and Event-Related Slow Wave Provide Dissociable Correlates of Visual Working Memory. *Journal of Neuroscience*, *35*(41), 14009–14016. https://doi.org/10.1523/JNEUROSCI.5003-14.2015

Fukuda, K., & Vogel, E. K. (2009). Human Variation in Overriding Attentional Capture. *Journal of Neuroscience*, *29*(27), 8726–8733. https://doi.org/10.1523/JNEUROSCI.2145-09.2009

Fukuda, K., & Vogel, E. K. (2011). Individual Differences in Recovery Time From Attentional Capture. *Psychological Science*, *22*(3), 361–368. https://doi.org/10.1177/0956797611398493

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. https://doi.org/10.3758/17.5.673

Fukuda, K., Woodman, G. F., & Vogel, E. K. (2015). Individual Differences in Visual Working Memory Capacity: Contributions of Attentional Control to Storage. In P. Jolicoeur, C. Lefebvre, & J. Martinez-Trujillo (Eds.), *Mechanisms of Sensory Working Memory: Attention and Performance XXV* (pp. 105–120). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780128013717000090

Gevins, A. (2000). Neurophysiological Measures of Working Memory and Individual Differences in Cognitive Ability and Cognitive Style. *Cerebral Cortex*, *10*(9), 829–839. https://doi.org/10.1093/cercor/10.9.829

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral Cortex (New York, N.Y.: 1991)*, *7*(4), 374–385. https://doi.org/10.1093/cercor/7.4.374

Gibson, B., Wasserman, E., & Luck, S. J. (2011). Qualitative similarities in the visual short-term memory of pigeons and people. *Psychonomic Bulletin & Review*, *18*(5), 979–984. https://doi.org/10.3758/s13423-011-0132-7

Giesbrecht, B., Woldorff, M. G., Song, A. W., & Mangun, G. R. (2003). Neural mechanisms of top-down control during spatial and feature attention. *NeuroImage*, *19*(3), 496–512. https://doi.org/10.1016/S1053-8119(03)00162-9

Gold, J. M., Hahn, B., Zhang, W. W., Robinson, B. M., Kappenman, E. S., Beck, V. M., & Luck, S. J. (2010). Reduced Capacity but Spared Precision and Maintenance of Working

Memory Representations in Schizophrenia. *Archives of General Psychiatry*, *67*(6), 570. https://doi.org/10.1001/archgenpsychiatry.2010.65

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical Working Memory Representations are used in Delayed Estimation of Continuous Colors. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 30–54. https://doi.org/doi:10.1037/xhp0000290

Harlow, I. M., & Donaldson, D. I. (2013). Source accuracy data reveal the thresholded nature of human episodic memory. *Psychonomic Bulletin & Review*, *20*(2), 318–325. https://doi.org/10.3758/s13423-012-0340-9

Harlow, I. M., & Yonelinas, A. P. (2016). Distinguishing between the success and precision of recollection. *Memory*, *24*(1), 114–127. https://doi.org/10.1080/09658211.2014.988162

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

Hsieh, L.-T., & Ranganath, C. (2014). Frontal midline theta oscillations during working memory maintenance and episodic encoding and retrieval. *NeuroImage*, *85 Pt 2*, 721–729. https://doi.org/10.1016/j.neuroimage.2013.08.003

Isbell, E., Fukuda, K., Neville, H. J., & Vogel, E. K. (2015). Visual working memory continues to develop through adolescence. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00696

Itthipuripat, S., Wessel, J. R., & Aron, A. R. (2013). Frontal theta is a signature of successful working memory manipulation. *Experimental Brain Research*, *224*(2), 255–262. https://doi.org/10.1007/s00221-012-3305-3

Jensen, O., & Tesche, C. D. (2002). Frontal theta activity in humans increases with memory load in a working memory task. *The European Journal of Neuroscience*, *15*(8), 1395–1399. https://doi.org/10.1046/j.1460-9568.2002.01975.x

Katus, T., Müller, M. M., & Eimer, M. (2015). Sustained Maintenance of Somatotopic Information in Brain Regions Recruited by Tactile Working Memory. *Journal of Neuroscience*, *35*(4), 1390–1395. https://doi.org/10.1523/JNEUROSCI.3535-14.2015

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, *29*(2–3), 169–195. https://doi.org/10.1016/S0165-0173(98)00056-3

Klimesch, W., Doppelmayr, M., Schimke, H., & Ripper, B. (1997). Theta synchronization and alpha desynchronization in a memory task. *Psychophysiology*, *34*(2), 169–176. https://doi.org/10.1111/j.1469-8986.1997.tb02128.x

LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding Attended Information in Short-term Memory: An EEG Study. *Journal of Cognitive Neuroscience*, *25*(1), 127–142. https://doi.org/10.1162/jocn_a_00305

Leber, A. B., Turk-Browne, N. B., & Chun, M. M. (2008). Neural predictors of moment-to-moment fluctuations in cognitive flexibility. *Proceedings of the National Academy of Sciences*, *105*(36), 13592–13597. https://doi.org/10.1073/pnas.0805423105

Lee, E.-Y., Cowan, N., Vogel, E. K., Rolan, T., Valle-Inclan, F., & Hackley, S. A. (2010). Visual working memory deficits in patients with Parkinson's disease are due to both reduced storage capacity and impaired ability to filter out irrelevant information. *Brain*, *133*(9), 2677–2689. https://doi.org/10.1093/brain/awq197

Liesefeld, A. M., Liesefeld, H. R., & Zimmer, H. D. (2014). Intercommunication Between Prefrontal and Posterior Brain Regions for Protecting Visual Working Memory From Distractor Interference. *Psychological Science*, *25*(2), 325–333. https://doi.org/10.1177/0956797613501170

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

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience & Biobehavioral Reviews*, *62*, 100–108. https://doi.org/10.1016/j.neubiorev.2016.01.003

Magen, H., Emmanouil, T.-A., McMains, S. A., Kastner, S., & Treisman, A. (2009). Attentional demands predict short-term memory load response in posterior parietal cortex. *Neuropsychologia*, *47*(8–9), 1790–1798. https://doi.org/10.1016/j.neuropsychologia.2009.02.015

McCollough, A. W., Machizawa, M. G., & Vogel, E. K. (2007). Electrophysiological Measures of Maintaining Representations in Visual Working Memory. *Cortex*, *43*(1), 77–94. https://doi.org/10.1016/S0010-9452(08)70447-7

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, *11*(1), 103–107. https://doi.org/10.1038/nn2024

McVay, J. C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin*, *136*(2), 188–197; discussion 198-207. https://doi.org/10.1037/a0018298

Mitchell, D. J., & Cusack, R. (2008). Flexible, Capacity-Limited Activity of Posterior Parietal Cortex in Perceptual as well as Visual Short-Term Memory Tasks. *Cerebral Cortex*, *18*(8), 1788–1798. https://doi.org/10.1093/cercor/bhm205

Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, *55*(1), 8–24. https://doi.org/10.1016/j.jmp.2010.08.008

Mrazek, M. D., Smallwood, J., Franklin, M. S., Chin, J. M., Baird, B., & Schooler, J. W. (2012). The role of mind-wandering in measurements of general aptitude. *Journal of Experimental Psychology. General*, *141*(4), 788–798. https://doi.org/10.1037/a0027968

Nosofsky, R. M., & Donkin, C. (2016a). Qualitative contrast between knowledge-limited mixed-state and variable-resources models of visual change detection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1507–1525. https://doi.org/10.1037/xlm0000268

Nosofsky, R. M., & Donkin, C. (2016b). Response-time evidence for mixed memory states in a sequential-presentation change-detection task. *Cognitive Psychology*, *84*, 31–62. https://doi.org/10.1016/j.cogpsych.2015.11.001

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38. https://doi.org/10.1016/j.neuroscience.2005.06.005

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

Prosser, S. (1995). Aspects of Short-term Auditory Memory as Revealed by a Recognition Task on Multi-tone Sequences. *Scandinavian Audiology*, *24*(4), 247–253. https://doi.org/10.3109/01050399509047544

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13), 21–21. https://doi.org/10.1167/12.13.21

Reddy, L., Kanwisher, N. G., & VanRullen, R. (2009). Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences*, *106*(50), 21447–21452. https://doi.org/10.1073/pnas.0907330106

Reinhart, R. M. G., Heitz, R. P., Purcell, B. A., Weigand, P. K., Schall, J. D., & Woodman, G. F. (2012). Homologous Mechanisms of Visuospatial Working Memory Maintenance in Macaque and Human: Properties and Sources. *Journal of Neuroscience*, *32*(22), 7711–7722. https://doi.org/10.1523/JNEUROSCI.0215-12.2012

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

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

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 of the United States of America*, *105*(16), 5975–5979. https://doi.org/10.1073/pnas.0711295105

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. https://doi.org/10.3758/s13423-011-0055-3

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

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

Sewell, D. K., Lilburn, S. D., & Smith, P. L. (2014). An information capacity limitation of visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(6), 2214–2242. https://doi.org/10.1037/a0037744

Sheremata, S. L., Somers, D. C., & Shomstein, S. (2018). Visual Short-Term Memory Activity in Parietal Lobe Reflects Cognitive Processes beyond Attentional Selection. *The Journal of Neuroscience*, *38*(6), 1511–1519. https://doi.org/10.1523/JNEUROSCI.1716-17.2017

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–830. https://doi.org/10.1037/a0029856

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin*, *132*(6), 946–958. https://doi.org/10.1037/0033-2909.132.6.946

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. *Current Biology*, *24*(18), 2174–2180. https://doi.org/10.1016/j.cub.2014.07.066

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

Sprague, T. C., Saproo, S., & Serences, J. T. (2015). Visual attention mitigates information loss in small- and large-scale neural codes. *Trends in Cognitive Sciences*, *19*(4), 215–226. https://doi.org/10.1016/j.tics.2015.02.005

Sprague, T. C., & Serences, J. T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nature Neuroscience*, *16*(12), 1879–1887. https://doi.org/10.1038/nn.3574

Stipacek, A., Grabner, R. H., Neuper, C., Fink, A., & Neubauer, A. C. (2003). Sensitivity of human EEG alpha band desynchronization to different working memory components and

increasing levels of memory load. *Neuroscience Letters*, *353*(3), 193–196. https://doi.org/10.1016/j.neulet.2003.09.044

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

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(10), 9–9. https://doi.org/10.1167/13.10.9

Sutterer, D. W., Foster, J. J., Adam, K. C. S., Vogel, E. K., & Awh, E. (in prep). Delay-period activity encodes multiple items stored in working memory.

Sutterer, D. W., Foster, J. J., Serences, J. T., Vogel, E. K., & Awh, E. (2018). Alpha-band oscillations track the retrieval of precise spatial representations from long-term memory. *BioRxiv*. https://doi.org/10.1101/207860

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*(6984), 751–754. https://doi.org/10.1038/nature02466

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective & Behavioral Neuroscience*, *5*(2), 144–155. https://doi.org/10.3758/CABN.5.2.144

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26. https://doi.org/10.1016/j.cogpsych.2014.01.003

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2015). Working Memory Delay Activity Predicts Individual Differences in Cognitive Abilities. *Journal of Cognitive Neuroscience*, *27*(5), 853–865. https://doi.org/10.1162/jocn_a_00765

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124–149. https://doi.org/10.1037/a0035234

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*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

Vogel, E. K., & Awh, E. (2008). How to Exploit Diversity for Scientific Gain: Using Individual Differences to Constrain Cognitive Theory. *Current Directions in Psychological Science*, *17*(2), 171–176. https://doi.org/10.1111/j.1467-8721.2008.00569.x

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. https://doi.org/10.1038/nature02447

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. https://doi.org/10.1038/nature04171

Weissman, D. H., Roberts, K. C., Visscher, K. M., & Woldorff, M. G. (2006). The neural bases of momentary lapses in attention. *Nature Neuroscience*, *9*(7), 971–978. https://doi.org/10.1038/nn1727

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135. https://doi.org/10:1167/4.12.11

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

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91–95. https://doi.org/10.1038/nature04262

Xu, Z., Adam, K. C. S., Fang, X., & Vogel, E. K. (2017). The reliability and stability of visual working memory capacity. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0886-6

Zakrzewska, M. Z., & Brzezicka, A. (2014). Working memory capacity as a moderator of load-related frontal midline theta variability in Sternberg task. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00399

Zanto, T. P., & Gazzaley, A. (2009). Neural Suppression of Irrelevant Information Underlies Optimal Working Memory Performance. *Journal of Neuroscience*, *29*(10), 3059–3066. https://doi.org/10.1523/JNEUROSCI.4621-08.2009

Zar, J. H. (2010). *Biostatistical Analysis* (5th ed.). Upper Saddle River, New Jersey: Prentice-Hall.

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