Assessing Capacity and Control of Visual Working Memory

by

John F. Magnotti IV

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 7, 2012

Keywords: visual working memory, change detection, capacity estimation, visual search

Approved by

Jeffrey S. Katz, Chair, Alumni Professor of Psychology
Ana M. Franco-Watkins, Associate Professor of Psychology
Gopikrishna Deshpande, Assistant Professor of Electrical and Computer Engineering
Frank W. Weathers, Professor of Psychology
Thomas Denney, Ed & Peggy Reynolds Family Professor of Electrical and Computer Engineering

Abstract

This paper develops 3 experiments to answer fundamental questions about the independence of storage and control processes in visual working memory (VWM). Results from Experiment 1 showed that executive control processes are invoked even for simple color-change detection tasks using the forced-choice method. Experiment 2 increased the difficulty of the sample-probe comparison in order to assess how high levels of comparison difficulty impact estimates of VWM capacity. Experiment 3 presents a novel modification of the change detection procedure and shows independence between storage and control processes in VWM.

Results from these experiments contribute to a better characterization of VWM. A running theme throughout each experiment is that changes to the task at choice time can affect VWM capacity estimates. Rather than support the original view that VWM is a highly robust representation, the current results add to the growing body of literature suggesting the fragility of VWM (e.g., Makovski, & Jiang, 2008; Makovski, Shim, & Jiang, 2006). More unique to the current work is the focus of interference at choice time rather than at encoding or during maintenance. Taken together, these results show the importance of quantifying the joint contribution of each VWM process in order to better understand them individually.

Acknowledgments

This work would not have been possible without the consistent provision of many people. First credit belongs with my advisor, Dr. Jeffrey Katz, for his wonderful blend of structure, freedom, and encouragement. Secondly, Dr. Ana Franco-Watkins has been an enthusiastic participant in nearly all of the major milestones I've reached at Auburn University. Drs. Frank Weathers and Gopi Deshpande have also contributed of their time quite generously on this and many other projects.

Much of this work was completed in-between stimulating coffee breaks with Alex Daniel. I also benefited greatly from the categorical support of Adam Goodman and his ever-increasing speculations about old and new components of working memory. My progress in the lab was due as least in part to those that went before me, especially Drs. Bradley Sturz, Kent Bodily, Michelle Hernández, and Kelly Schmidtke, who helped to ease my transition into the lab and were excellent colleagues on numerous projects.

I am in continual debt to my wife, Jennifer, who shouldered perhaps more than her fair share of this burden, so that I might possibly succeed. Finally, to my parents, John and Valerie, who provided many carrots (and a few sticks) along the way to ensure I would finish what I started.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Our dependence on visual information in daily life is difficult to overestimate. Even the simple procedure of locating our favorite coffee mug on a shelf involves a dizzying amount of neural computation to coordinate action based on visual information readily available in the environment and that stored in memory. Questions on the nature of visual memory date back to at least Artistotle's work, *On Dreams*, composed over two millennia ago. The characterization of visual working memory (VWM) remains a major area of research for today's cognitive scientists (Logie, 2011). VWM is the set of processes involved in the active storage and processing of visual information, distinct from auditory working memory and visual long-term memory (Baddeley, 1992, 1998; Brady, Konkle, & Alvarez, 2011; Cowan, 2010; Vogel, Woodman, & Luck, 2001). A better understanding of VWM will lead to a fuller characterization of various mental illnesses that involve a VWM deficit such as Williams' syndrome (Atkinson, et al., 1997), Posttraumatic Stress Disorder (e.g., Litz, et al., 1996; Vasterling, Brailey, Constans, & Sutker, 1998), Schizophrenia (e.g., Gold, Fuller, Robinson, McMahon, Braun, & Luck, 2006; Gold, et al., 2010), and specific phobias (e.g., McGlynn, Wheeler, Wilamowska, & Katz, 2008). Additionally, performance on VWM procedures has been found to correlate with measures of general intelligence (e.g., Cowan et al., 2005; Fukuda, Vogel, Mayr, & Awh, 2010). Of *a priori* concern is the establishment of a method that is flexible enough to assess multiple aspects of VWM while maintaining enough simplicity to ensure its adoption and implementation in myriad contexts and populations.

The goal of this review and subsequent experiments is to assess the independence of storage and control processes central to VWM. Storage is defined as the number of objects that may be retained across an empty interval (a later section discusses the appropriateness of this metric). The control processes measured were those invoked at retrieval time, which may be distinct from the attentional control processes often assessed at encoding (e.g., stroop interference resolution, encoding of perceptual groups). Joint assessment of storage and control allows a fuller conceptualization of VWM, capturing important interactions and tradeoffs between these processes that may be missed when they are assessed separately. Further, by measuring individual performance correlations across the capacity and control manipulations, the relative independence of these processes can also be measured, placing constraints on how the process of VWM may be instantiated (Vogel & Awh, 2008).

The first part of this review considers experiments that demonstrate a distinction between high-capacity iconic memory and limited-capacity visual working memory. This section is primarily concerned with the methodological details necessary to ensure that a given experiment is measuring VWM rather than iconic memory. Next, attention is turned to issues in developing a discrete measure of VWM capacity and the units of such a measure. Throughout the second section, emphasis was placed on the development of capacity as a measurable quantity. The third section considers theoretical implications of the discrete storage assumption used to develop the fixed-capacity theory of VWM described in the second section. The last section of the review focuses on recent attempts to parcel out the contribution of comparison difficulty to estimates of VWM capacity. Finally, a series of experiments are proposed to assess the impact of comparison difficulty on change detection using simple objects and how capacity estimates are influenced.

Experiment 1 will test the hypothesis that control processes in VWM provide strong resistance to interference at choice time by manipulating both the memory load (number of items to encode) and comparison difficulty (number of distractors) as participants search for a single changed object. Comparison difficulty is further increased in Experiment 2 by changing the task to a search for a single unchanged object. The relative difficulty of searching through non-encoded objects was assessed by comparison to Experiment 1. Experiment 3 uses the same approach as Experiment 2, but eliminates the spatial relationship between the sample and probe arrays to test if the decreases in performance in Experiments 1 and 2 are due to changes in the spatial relationship between the two arrays rather than a change in comparison difficulty. Decoupling the spatial relationship between the arrays results in a cleaner estimate of storage and control processes of visual *object* memory. The ultimate goal of Experiment 3 is to test if individuals with a large storage capacity also have excellent resistance to interference (i.e., process independence). Although previous research has looked at individual differences for inhibiting encoding (i.e., selection) of extraneous information (e.g., Vogel, McCollough, & Machizawa, 2005), but less has been done to understand individual differences at resisting interference during retrieval (i.e., access) at sizes necessary to estimate VWM capacity (Hyun, Woodman, Vogel, Hollingworth, & Luck, 2009).

Terminology and general method details. Throughout this review, the language used to describe various methods has been standardized to accord better with current VWM research (e.g., immediate memory becomes working memory). The term visual memory is used collectively to refer to visual working memory and iconic memory for studies that did not make a distinction between them. To facilitate comparison among the methods described, trial descriptions have been standardized using the language used in Figure 1. At the start of a trial,

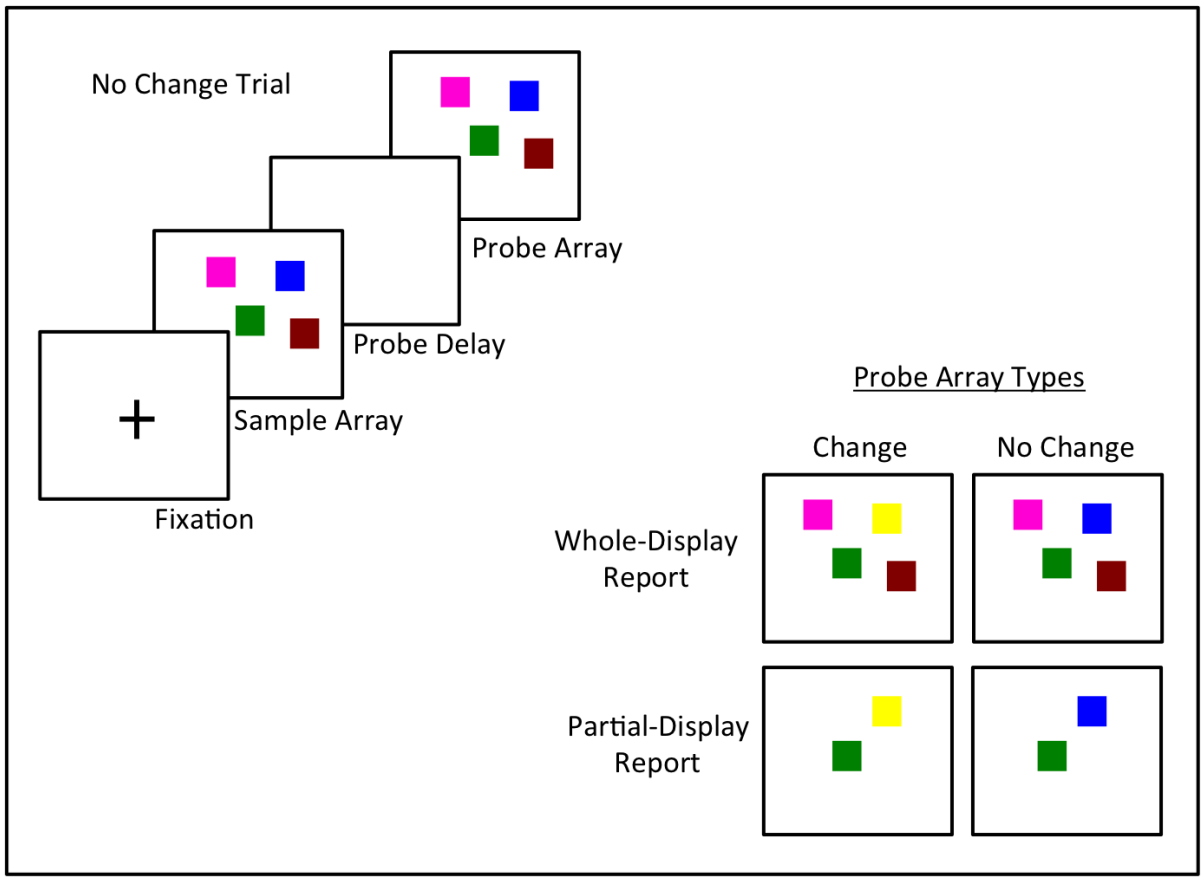


Figure 1. *Left*: Example of a typical change detection trial with no change from the sample to probe array. *Right*: common ways to present the probe array. In the yes/no procedure, participants make a binary change/no-change decision. In the forced-choice procedure, all probe arrays contain a changed item (left column), and participants select either the changed or unchanged item.

participants typically view a fixation image for a set time before viewing the sample array. The amount of time spent viewing the sample array will be referred to as the viewing time. The sample array is comprised of visual items from one or more stimulus classes (e.g., colored squares, irregular polygons, travel slides, human faces). The sample array is then followed by a probe delay that may or may not contain a masking stimulus. After the probe delay participants view a second array called the probe array and make a decision about the item(s) in the probe array (e.g., "pick the changed item"), or their relationship with the sample array (e.g., "same" or "different"). The number of items in the sample and probe arrays are called the sample set size

and probe set size, respectively, whereas the number of unique stimuli available across an experiment per stimulus type is called the stimulus set size (e.g., if an experiment used 12 colors and 4 shapes, the stimulus set sizes would be 12 and 4, regardless of how many items are displayed in the sample or probe array). If the probe set size is identical to the sample set size, the procedure is called a whole-display test, if the probe set size is less than the sample set size it is called a partial-display test or single-probe test if a single item is in the probe array (Jiang, Olsun, & Chun, 2000; Rouder, Morey, Morey, & Cowan, 2011; Wheeler & Treisman, 2002). For experiments that use a signal to indicate that participants should attend to a subset of sample array, the signal is labeled based on the difference between the time of its onset and the onset of the sample array, referred to as the stimulus onset asynchrony (SOA). The signal is a pre-cue if the signal occurs before (negative SOA) or concurrent with ($0 < SOA <$ viewing time) the onset of the sample array and a post-cue if it occurs at or after the termination of the sample array ($SOA \geq$ viewing time). For some experiments, a subset of the probe array was cued for processing. These cues appear simultaneously with the probe array unless otherwise noted.

Pashler (1988) coined the term "change detection" to describe procedures that involve relating sequential arrays displayed at the same spatial location. Luck and Vogel (1997) credit Phillips (1974) as the developer of the modern, sequential change detection procedure. In distinction to other same/different procedures, the probe delay in change detection is relatively brief, long enough only to ensure that iconic memory is not responsible for performance. A related phenomenon, "change blindness," often conceptualized as a failure to detect changes in the natural environment (Simons, & Rensink, 2005), is sometimes considered distinct from performance in the yes/no and forced-choice type procedures (for a review, see Rensink, 2002). The primary problem in relating change detection performance to the phenomenon of change

blindness relates to *how* the problem is solved—Rensink (2002) maintains that perceived correspondence between the sample and probe array (i.e., the probe array/scene is a transformation of the sample array/scene) is a hallmark of change detection, separating it from difference detection, which can be applied arbitrarily without requiring any relationship between the sample and probe array. For example, participants are detecting changes when they report that an object within a scene has been rotated (the overall scene is still intact), whereas participants are only reporting a difference when they discriminate the gender of two successively presented faces. Problematically, when arrays of objects are used, this definition makes it hard to know when a same/different procedure becomes a change detection procedure and vice versa. For the present review and subsequent experiments, the term change detection will be used in the sense of Pashler (1988), as a designation of specifics for a particular set of methods (in agreement with Gibson, Wasserman, & Luck, 2011), rather than a distinct set of cognitive processes (e.g., Rensink, 2002; Wright et al., 2010).

A brief look at early visual memory research

Whole and Partial Report Procedures. Perhaps the simplest method to assess the contents of an individual's visual memory is simply to show them an array of visual objects for a fixed amount of time and then ask them to report the identity of each object once the array is no longer visible. Early work by Sperling (1960) used both a whole report technique (i.e., list all remembered items) and a partial report technique (i.e., report only a subset of the items) to identify how much information could be extracted from visual scenes. Participants viewed arrays of consonants or consonants and digits arranged in rectangular grids (1x3, 1x5, 1x6, 2x3, 2x4, 3x3, and 3x4) with sample set sizes ranging from 3 up to 12 for 50ms before writing down all the information remembered. Across the 5 subjects, the average number of items remembered was

about 4, regardless of the sample set size, grid configuration, or the kind of stimuli used (just consonants or numbers and consonants). A second experiment using similar sample arrays but with variable viewing times of 15, 50, 150, and 500ms also produced a capacity estimate around 4 items, demonstrating that encoding limitations are unlikely to be the primary information bottleneck.

One drawback of the whole report method is its susceptibility to interference (e.g., Tulving & Arbuckle, 1966). The action of reporting the contents of visual memory may itself interfere with the contents of memory. Additionally, if the contents of visual memory decay rapidly, memory for an item may decay before an observer can finish the report. To circumvent this issue, observers can be asked to give a report of part of the sample array. In the partial report method, the subset of the sample array to be recalled may be post-cued by an auditory signal. Crucially, the observers cannot know which subset to report in advance, and therefore must attend to each subset evenly to obtain the highest accuracy. To estimate the number of items initially represented in visual memory, the proportion of items recalled from the cued subset is multiplied by the sample set size. For example, if an observer is able to report, on average, 2 items from a post-cued row of a 2x4 sample array (sample set size 8), the capacity estimate is $2/4 \times 8 = 4$. This calculation assumes that observers are attending to each subset evenly (essentially an assumption that visual "samples" are independent and identically distributed uniformly across the grid) and does not take into account any partial information that may be extracted (e.g., if a participant reports a correct letter in the wrong location, it is considered an incorrect report). Also, this estimate is a lower bound for capacity, as it estimates the minimum number of letters that would be required to obtain a given accuracy level.

In another experiment using only multi-row sample arrays, Sperling (1960) inserted an auditory cue to signal which row should be reported (different frequencies corresponded to different rows) immediately following the sample array offset. Using this partial report method, the average estimate for the number of items in visual memory increased to 9.1. To determine if interference or decay was chiefly responsible for the discrepancy between the capacity estimates from partial and whole report, the SOA between the sample array and the auditory cue was varied across trials from -50ms to 1050ms. Participants were instructed not to rehearse the letters to reduce the amount of visual information that was verbally recoded. Estimates of visual memory capacity were relatively stable for SOAs of -50ms and 50ms, but dropped sharply from 50ms to 350ms, then more slowly declined from 350ms to 1050ms. Interestingly, when the SOA was 1050ms, capacity estimates from the partial report method coincided closely with those from the whole report method. Convergent findings were also reported by Keele and Chase (1967), using SOAs out to 5000ms with alphanumeric arrays displayed at low, medium, and high luminance levels. Yeomans and Irwin (1985) showed that the partial report benefit is not a by-product of extremely brief viewing times, but rather information persists in iconic memory using viewing times approximating typical eye fixation times (e.g., 500ms).

Having demonstrated that visual memory is susceptible to rapid decay, the next experiment reported by Sperling (1960) assessed how fragile visual memory is by using post-sample array masking. Rather than the pre- and post-sample array background being dark as in the previous experiments, a new condition, called the Baxt condition after its originator, was used with a dark pre-sample array background and a white post-sample array background. Regardless of viewing time (15 or 50ms), accuracy in the Baxt condition was significantly reduced (2.8 items compared to 4.3 in the normal condition). Taken together with the previous

experiments, there seems to be a division in visual memory: a high capacity, fragile, short duration component and a lower capacity, durable, longer duration component. Sperling (1963) further showed that information enters into the durable component only after 50-100ms of viewing time, a result that has withstood much scrutiny (Vogel, Woodman, & Luck, 2006).

A potential problem with Sperling's design is the use of stimuli that maybe easily nameable and more prone to verbal rehearsal. Indeed, a short-term memory model developed by Sperling (1963), based in large part from his work with partial report of letters and digits, specified that all visual images were converted to an auditory code to facilitate rehearsal. Purdy, Eimann, and Cross (1980) also suggested that Sperling's whole and partial report techniques require participants to use long-term representations for letter identification (e.g., identifying a particular visual pattern as the letter "A"), causing interference and an underestimate of capacity. This explanation casts some doubt on the applicability of his capacity estimates to VWM; therefore a different method of assessment is needed to characterize visual storage. In support of using Sperling's data for VWM, however, in an analogous study done by Keele and Chase (1967) an error analysis designed to look for acoustic confusions in the partial report data did not indicate that participants were relying primarily on a verbal recoding of the stimuli.

Shaffer and Shiffrin (1972) suggested that the Sperling's results are not necessarily indicative of separate memory stores. Instead, they hypothesized that visual memory may be stored monolithically, with rapid decay for complex information, and relatively slow decay for simple information. The reason that masking and increased probe delay had such an effect on capacity estimates in the earlier experiments is thus due to the rate of decay of simple vs. complex information, rather than the reliance on dissociable memory stores. Whether the participant is trying to remember a large or small number of items, the information is placed into

unitary storage that shows decay proportional to the amount of information placed into storage. In a picture recognition experiment, Shaffer & Shiffrin (1972) showed that memory (and recognition confidence) for complex picture stimuli does not increase with increased time for rehearsal or removal of masking.

Yes/No Procedures. Phillips (1974) sought to explicitly test for distinctions between iconic memory and VWM using a visual-pattern same/different procedure. In this procedure, participants are instructed to report if the sample and probe arrays are the *same* or *different*. Phillips (1974) created sample arrays by lighting a subset of cells from grids of multiple sizes: 4x4, 6x6, and 8x8. Each cell had a 50% chance of being lit. These patterns are not easily nameable[1] and have less familiarity than the letters and numbers used by Sperling (1960). Participants viewed the sample array for 1s, followed by a probe delay of .02, 1, 3, or 9s, and finally the probe array. On same trials, the probe array was identical to the sample array. On different trials, the probe array differed by exactly one grid cell (the cell changed from lit to unlit or from unlit to lit). Sessions contained an equal number of trials at each probe delay and trial type, randomly ordered.

Across increasing probe delay, decreases in accuracy were dependent on the sample set size: larger sample set sizes showed proportionally larger initial decreases compared to the 4x4 grid. By the longest probe delay (9s), accuracy was relatively stable, and showed large differences between each sample set size even after converting to d', 4x4: d' = 1.38; 6x6: d' = .52; 8x8: d' = -.16. At the smallest probe delay (.02s), no accuracy differences existed between

[1]A subset of participants were asked to produce brief, accurate written descriptions of 10, 5x5 grids to measure the ease of verbal transcription. On average, participants spent 4 minutes writing verbal descriptions for each pattern.

the arrays, replicating earlier results of a large capacity visual store that quickly decays. Phillips (1974) noted that these results are also consistent with Shaffer and Shiffrin's (1972) account of unitary visual storage.

A second experiment compared same/different performance using a smaller range of probe delays (.02s, .06s, .1s, .3s, and .6s) with 5x5 and 8x8 stimulus arrays. Additionally, on half the trials the probe array was displaced horizontally by the width of a single grid cell (subtending a visual angle of 0.45°). If visual memory is stored in a single store, then accuracy should decline similarly in the "move" and "still" conditions, with perhaps a difference in performance at the lowest probe delays. Instead, Phillips (1974) found that performance in the move condition was virtually unaffected by manipulations of probe delay. At probe delays less than .3s, accuracy in the move condition was significantly lower than the still condition, but performance was indistinguishable once the probe delay was at or above .3s. These data cannot be explained easily by the differential decay rate hypothesis, and Phillips concluded that a clear functional difference exists between iconic memory and VWM: high capacity iconic memory is bound to specific spatial positions. In an additional experiment, Phillips (1974) showed that the decrease in accuracy caused by moving the probe array was comparable to post-stimulus masking.

For each experiment, response time data mimicked the accuracy data—at short probe delays, response times are independent of sample set size, whereas at longer probe delays, response time was directly related to sample set size. Additionally, for masked sample arrays, response time data increased with sample set size. This pattern is consistent with iconic memory allowing parallel comparison between the in-memory array (which is a subset of the sample array plus any additional information the participant is concurrently maintaining) and the probe array, but VWM entails a serial comparison between the in-memory array and the probe array.

Considering accuracy and response time data from these previous studies together, several distinctions between iconic memory and VWM emerge in terms of storage capacity (high vs. low), spatial specificity (high vs. low), resistance to interference (low vs. high), and resistance to decay (low vs. high). These distinctions are presented in Table 1. Importantly, these differences cannot be explained by a unitary storage mechanism with decay rate based on information load.

A key difference between Phillips' approach and previous efforts is the lack of an explicit report component. Rather than require participants to recall the contents of their memory, a single decision is made that reflects the contents of their memory. Phillips' design is classified as a yes/no procedure, because participants make a binary yes (same) or no (different) response Phillips (1974). This procedure is distinguished from a forced-choice procedure (for examples in change detection, see Eng, Chen, & Jiang, 2005; Wright et al., 2010), in which an item in the probe array always contains a single item with the quality being tested (e.g., a single different item). The forced-choice procedure will be discussed in more detail later.

Although the yes/no procedure indirectly measures the contents of memory, it reduces the interference produced by reporting the contents of memory, allowing a potentially purer index into the characteristics of VWM. New complications are introduced, however, as the chance level becomes quite high (50% in standard same/different and change/no-change paradigms) and guessing strategies must be explicitly modeled (e.g., Rouder, et al., 2011). In a typical whole report study, the chance of guessing a single row of alphanumeric characters is comparatively small: about 3% per character, and .004% for a row of just 3 characters. A later section in this review will provide a more formal approach to estimating capacity, for now it suffices to note only that the previous partial-report method is not ideal.

An attempt to estimate storage capacity based on Philips (1974) results is complicated by his use of lit/unlit grid patterns. These grids can be considered as binary on/off patterns, and the procedure solved by storing only the "on" (or "off") locations instead of remembering item identities. By relying on location information, participants can extract configural information about the sample array, rather than storing individual items independently (one of the assumptions mentioned above). For example, if a contiguous 4x4 region is uniformly lit or unlit, participants may remember a filled "square" of white or black, rather than 16 discrete items. If a change occurs to a cell within this square, the shape is no longer uniformly filled in the probe array, and the change may be detected based on a change to the pattern, rather than to a single cell's state. A simplistic capacity estimate that assumes independent storage of all items in the sample array might conclude that the participant had correctly stored all 16 items. Although the grids used by Phillips (1974) were not uniform, pattern information could still be extracted. Results from texture segregation (e.g., Julesz, 1981; Treisman & Gelade, 1980), symmetry detection (e.g., Huang & Pashler, 2002; Pashler, 1990), and location-change detection (e.g., Huang, 2010; Jiang, et al., 2000; Wheeler & Treisman, 2002) confirm that pattern information can be encoded quickly and may not be subject to the same limitations as object feature information (Huang & Pashler, 2007; Huang, Treisman, & Pashler, 2007). In short, there is no simple way to derive a capacity estimate from the grids used by Phillips without knowing precisely the encoding strategy used by participants (cf. Linke, Vicente-Grabovetsky, Mitchell, & Cusack, 2011, for related issues even when more discrete stimuli are used).

Table 1. Summary of functional distinctions between iconic memory and visual working memory (Phillips, 1974; Sperling, 1960).

Characteristic	Iconic Memory	Visual Working Memory
Storage Capacity	Extremely large	Around 4
Spatial Specificity	Within 0.45°	Not bound to position
Resistance to Interference	Very sensitive to masking	Not sensitive to masking
Resistance to Decay	Large decay within 100ms	Little decay within 9s
Access Method	Parallel	Serial/Limited Capacity Parallel

Estimating visual working memory capacity using change detection

Capacity estimation in whole-display report procedures. More than five years after Phillips' demonstration of functional differences between iconic memory and VWM, Purdy, et al. (1980) combined the alphanumeric arrays from Sperling's (1960) work with the yes/no recognition procedure from Phillips (1974). In the Purdy et al. procedure, participants viewed a 4x4 sample array of letters for 100ms, followed by a probe delay of 100, 250, 500, or 5000ms (probe delay was a between-subjects manipulation) before being presented with a 4x4 probe array for 100ms. The probe array could contain 0 (the probe array is identical to the sample array), 1, 4, or 16 changed letters. Somewhat problematically, Purdy and colleagues gave participants only 6 same trials, and 2 of each level of change trials, although 58 participants completed the experiment. The primary hypothesis in their study was that less than perfect partial report data from the iconic memory conditions in Sperling's (1960) experiments were due to the increased interference of the specific reporting method. Participants in the current procedure made a single key response after the offset of the probe array to indicate "same" or "different" (Purdy et al., 1980).

Across the levels of change manipulation, participants were most accurate when all 16 items changed (nearly 100% correct across all probe delays) followed by the no-change trials (around 80% correct across all probe delays). Performance was poor at the 100ms probe delay for trials

containing only 1 or 4 changed items (around 40% correct), but showed a puzzling increase for these trials at the 5s probe delay (up to 60% correct). Purdy et al. (1980) noted that if participants were relying on iconic memory then performance should be equivalent across the level of change manipulation. Despite a significant interaction between change level and probe delay, the authors focused on the overall accuracy across probe delay. This analysis showed no difference between any probe delay, a result certainly at odds with the steep differences between probe delays in earlier studies Purdy et al. (1980). The authors concluded that iconic memory remains intact out to at least 5s when a recognition procedure is used in place of a recall procedure.

A problem with the procedure from Purdy et al. (1980) is the use of probe delays at or above 100ms. Compared to the earlier studies (e.g., Phillips, 1974; Sperling, 1960), this relatively long probe delay may have been beyond the timespan of iconic memory, requiring participants to rely on VWM instead. If participants were relying on VWM, then the steady performance across probe delay actually agrees with Sperling's (1960) results and strengthens the claim that VWM is highly resistant to decay.

Suggesting that participants were using VWM explains the performance across the change-level variable, but fails to explain why performance rose across probe delay for trials with 1 or 4 changes. One way to explain this result is to suggest that the increased probe delay may allow time for rehearsal of the sample array. Using a list-memory picture recognition procedure, Intraub (1980) showed that recognition performance is nearly equivalent between images viewed for 6s and those viewed for only 110ms but followed by a 5890ms interstimulus interval (ISI; the time between successive images in the sample). If some form of rehearsal is playing a role, then by definition of VWM, Purdy et al. (1980) were measuring performance based on VWM, rather than iconic memory.

Feeling confident that the procedure was actually solved using VWM, a capacity estimate can be derived from the reported accuracies, as in Pashler (1988). The primary logic of Pashler's discrete capacity formula is to estimate the minimum number of items that must be remembered in order to achieve a given hit rate. For example, if a participant is able to extract 5 of 16 items from a 1-item change trial from Purdy et al. (1980), the probability that the changed item is in memory is 5/16 = .3125. The probability that an item is not in memory *and* guessed correctly is 11/16 × *g*, where *g* is the guessing rate, assumed to be equal to the false alarm rate. The total hit rate is therefore the sum of the proportion of correctly remembered trials and the proportion of correctly guessed trials: 5/16 + 11/16 × *g*. For a false alarm rate around .2 (corresponding to the 80% correct on no-change trials), a capacity of five items corresponds to an obtained hit rate of .45, very close to the .42 and .43 hit rates for probe delays of 100 and 250ms from Purdy et al. (1980). The general formula for hit rate, *H*, is

$$H = \frac{k}{N} + \frac{N-k}{N} g$$

Where *k* is the estimate of capacity, *N* is the sample set size, and *g* the guessing rate, assumed to be identical to the false alarm rate, as before. The formula is only valid for $k \leq N$, because *N* establishes an upper limit. With a bit of re-arranging, we can estimate *k* directly:

$$k = \frac{N(h-g)}{1-g}$$

For completeness, note that $g < 1$ and $h \geq g$ are also required to produce usable capacity estimates. In the studies reviewed here, these latter constraints are not an issue.

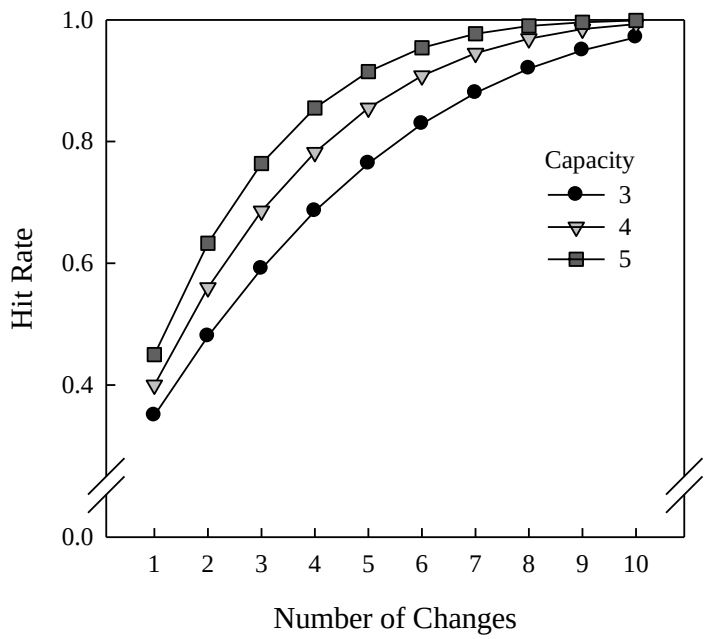


Figure 2. Hit rates for change detection with variable number of changes in a 16-item array, calculated according to Gibson et al., (2011) for capacities of 3, 4, and 5. Hit rate for a single changed item is identical to the formula given by Pashler (1988). False alarm rate is assumed to be 0.2 across all change levels. Random guess rate is calculated as (16 – Number of Changes)/16.

For the case with more than one item changing, Gibson, Wasserman, and Luck (2011, see supplement for details) present a generalization of Pashler's formula. With multiple changes, the probability of *failing* to remember each changed-item is separately estimated: $(N - k) / N$ for the first changed item, and $\{[N - (i - 1)] - k\} / [N - (i - 1)]$ for the i^{th} changed item, up to the number of changed items, c. The joint probability of failing to remember all the changed items, assuming maintenance of a single item is independent from another (something not to be assumed lightly, cf. Johnson, Spencer, Luck, & Schöner, 2009; Lin & Luck, 2009) is

$$\prod_{i=1}^{c} \frac{[N - (i - 1)] - k}{N - (i - 1)}$$

This formulation is valid only for $k \leq N - (c - 1)$, because $N - (c - 1)$ establishes an upper limit on estimated capacity. The probability of remembering at least one changed item is the complement of not remembering all the items (the events are mutually exclusive), and the observed hit rate is again the sum of the proportion of correctly remembered trials and the

proportion of those correctly guessed (the product of false alarm rate and the joint probability specified above).

With this formulation, VWM capacity can be estimated across the different change levels in Purdy et al. (1980). Figure 2 shows predicted hit rates for capacities of 3, 4, and 5, when the change level is varied systematically from 1 to 10 and the false alarm rate is fixed at 0.2. Note that even a capacity of 1 would predict perfect performance for the 16-item change condition, because the $k \leq N - (c - 1)$ constraint is not satisfied for any choice of $k > 1$.

Beyond providing a closed-form estimate of VWM capacity, Pashler (1988) demonstrated that Phillips' (1974) results can be extended to familiar characters and are not an artifact of using unfamiliar, binary stimuli. Across several experiments, Pashler manipulated combinations of probe delay (34ms, 67ms, 217ms), viewing time (100ms, 300ms, 500ms), masking (using both a checkerboard mask and a homogenous white field), and ease of identifying the stimuli (reflected v. upright letters). Accuracy dropped off precipitously when probe delay exceeded 34ms but remained constant from 67ms to 217ms. Viewing time had only modest effects, but masking with both the checkerboard and homogenous masks produced marked decreases in accuracy at even the shortest probe delays. The difference between masked and unmasked performance greatly decreased, but still remained, at longer viewing times, suggesting that VWM may not be entirely robust to interference from irrelevant stimuli or that the mask partially disrupts the consolidation of iconic memory into VWM. The final set of experiments showed that increasing the difficulty of identifying the letters did not significantly impact change detection performance, suggesting that neither categorical information nor familiarity facilitate change detection (or that category/familiarity information is not available). Under conditions similar to Sperling's (1960) whole report procedure (2x5 stimulus array, 300-

ms viewing time, no mask, 67-ms probe delay), Pashler reported an estimated VWM capacity of 4.24 in the change detection procedure, remarkably similar to the 4.3-letter estimate given by Sperling. This convergence across different reporting methods (not to mention nearly 30 years of advances in experiment technology) provides strong evidence for a fixed limit on VWM capacity.

Luck and Vogel (1997) extended the 4-item capacity limit in VWM using change detection with colored squares (stimulus set size 7, sample set sizes up to 12) at viewing times of 100 or 500ms (see Figure 3, a and b). They also showed no effect of adding a concurrent auditory load (silently repeating digits) or reducing decision load by cuing an item in the probe array. In the cued condition, participants decided if the cued item had changed, without considering the status of the other items. The limitations on performance are thus firmly confined to the storage and comparison stages of that procedure, even with simple colors.

Determining the unit of storage capacity. All the past studies just summarized used stimuli defined by values along a single dimension (e.g., shape, color), leading to concerns about the generalizability of the current capacity constraint. Work in visual target search has repeatedly demonstrated that searching for conjunction targets (i.e., a target defined by a specific combination of features) is significantly more difficult than a search for a single feature (Treisman & Gelade, 1980; Wolfe, 1998), and perhaps VWM capacity is determined by the number of features being remembered rather than by the number of objects. Additional experiments by Luck and Vogel (1997) tested this hypothesis by having participants look for changes in objects defined by conjunctions of color and orientation, and compared accuracy in this conjunction condition to accuracy in objects defined by color or orientation alone. Across sample set sizes of 2, 4, and 6, no accuracy differences were found between the conjunction and

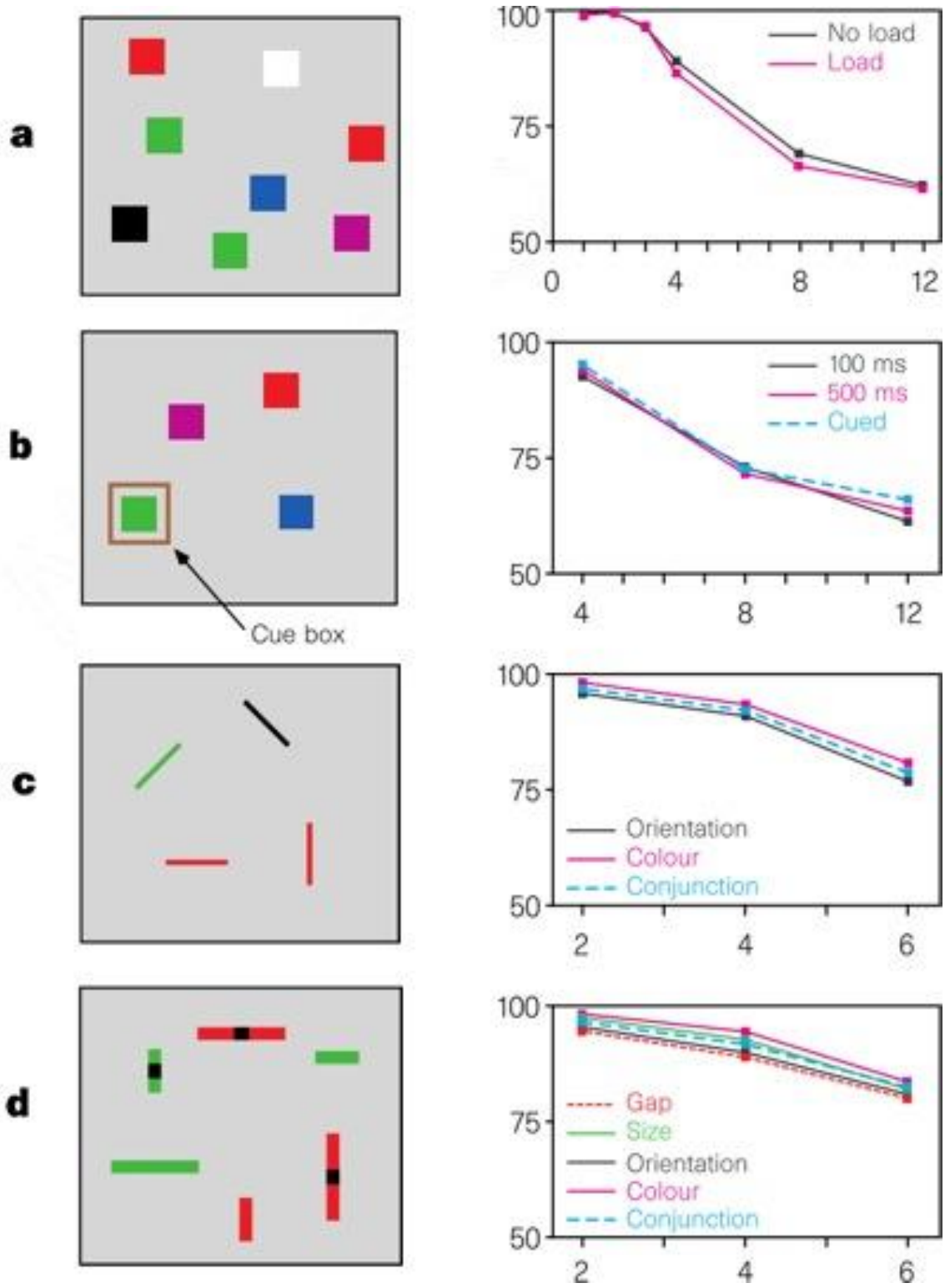


Figure 3. Example probe arrays and results from Luck and Vogel (1997). See text for details.

single-feature conditions (Figure 3, c). To demonstrate the extremity of the object-storage hypothesis, another experiment showed equivalent performance for quadruple conjunctions of gap (a black bar bisecting a colored rectangle), color, orientation, and size compared to performance for objects defined by only one of those features (Figure 3, d). As mentioned earlier, the distinction between single-feature and conjunction stimuli is the presence of multiple stimulus attributes. Within the visual memory literature, both of these types of stimuli are called "objects," but it is suggested that conjunction stimuli require additional processing to ensure accurate maintenance of binding information (the information that connects the separate features to the same object) across dimensions (e.g., Treisman, 1996). Even single-feature stimuli, though, may be bound to a specific location or serial order if required by the procedure (Avons, Ward, & Russo, 2001; Nairne & Neath, 2001). In the current context, a bound object is a set of visual features displayed at a common location, distinguishable from another set of identical features solely via location information. If the sample and probe array contain no duplicate stimuli, then the visual features need not be bound to specific locations. Extensive theoretical accounts of visual attention suggest that object features may be accessed via location information (Huang & Pashler, 2002), but also that features can be encoded without precise location information (e.g., Treisman, 1998; Treisman & Schmidt, 1982).

Accurate performance in the conjunction condition, however, need not be based on bound objects. Instead, if several of the dimensions from the quadruple-conjunction condition are separable (for a discussion of separable vs. non-separable dimensions, see Treisman & Gormican, 1988), then they may have been stored in dissociable feature maps (e.g., Treisman & Gelade, 1980; Treisman and Sato, 1990). For example, if the sample array contains conjunctions of color and orientation, then color and orientation information may be encoded and stored in

separate feature maps with independent capacities. Observers then compare location-color and location-orientation pairs between the sample and probe arrays to make a decision.

To explicitly contrast feature vs. object storage, Luck and Vogel (1997) presented participants with arrays of 2-color squares (an inner square of one color, and an outer square of another). A sample set size of 4 now corresponds to 8 colors. If the number of features sets the capacity of VWM, then performance with 4, 2-color squares should be comparable to performance with 8, single-color squares. On the other hand, if VWM stores bound objects, performance with 4, 2-color squares should be comparable to performance with 4 single-color squares. Across sample set sizes of 2, 4, and 6 and using a viewing time of 100ms, change detection accuracy for two-color squares was indistinguishable from that of single-color squares. Luck and Vogel concluded that VWM contains bound objects rather than unbound features, and the unit of capacity is therefore objects, not features. Although not discussed in detail by Luck and Vogel, an important point is the low stimulus set size used: only 4 colors made up each display, even when 6, 2-color squares were displayed. This point does not mean definitively that participants were not storing 4 bound objects, but the claim that features can be added to objects without cost may not be tenable.

Wheeler and Treisman (2002) explicitly tested the cost of encoding unique features across objects. Interestingly, Wheeler and Treisman failed to replicate the results from the 2-color squares condition in Luck and Vogel (1997). The 2-color squares used by Wheeler and Treisman varied in the composition, but included the inner/outer composition used by Luck and Vogel. As mentioned, the important difference is that Wheeler and Treisman did not repeat colors within the sample array and the color changes in the probe array were to a unique color (i.e., a color not displayed elsewhere in the sample or probe array), negating any potential feature-repetition

benefit. When unique colors were used, the number of colors (features) rather than the number of objects determined accuracy. A second experiment using only 4 colors that were allowed to repeat also failed to replicate Luck and Vogel's (1997) result that capacity is determined by object count. The only concrete explanation for the failure to replicate given by Wheeler and Treisman relates to the nature of the articulatory suppression: participants in the Wheeler and Treisman study repeated the phrase "Coca Cola" throughout the sample array and probe delay, whereas participants in the Luck and Vogel study silently maintained two numbers. Wheeler and Treisman were not confident that this difference was responsible for the discrepancy, and other studies have also failed to find any appreciable effect of articulatory suppression on detecting changes in colored squares (e.g., Eng, Chen, & Jiang, 2005; Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001). Anecdotally, Wheeler and Treisman noted that some participants reported only attending to the inner color rather than to the whole 2-color object, which may explain why accuracy was dependent on the number of colors, rather than the number of objects. At the extremely short viewing times used in these studies (e.g., less than 500ms), encoding strategy likely plays a large role (Bays, Catalao, & Husain, 2009; Linke et al., 2011).

Additional experiments from Wheeler and Treisman (2002) assessed memory for bound objects vs. individual features by intermixing color changes with location changes. For the location changes, an item could move to a new, previously unused location (logically similar to a color change), or two items could switch places. In order to detect the latter kind of location change, participants must successfully maintain a bound representation. The different types of changes were presented in 4 blocks: color only, location only, either color only or location only (hereafter, the 2-feature condition), and the binding condition. The 2-feature condition replicates the logic of the conjunctions from Luck and Vogel (1997) and may be solved without explicit

binding of color and location. By comparing the 2-feature condition with the binding condition the additional cost beyond feature storage of maintaining bound representations can be assessed. If the number of visual objects determines VWM capacity, then accuracy should not differ between these conditions. Accuracy across sample set sizes of 3 and 6 was highest for the location only (*M* = 97%) and for location changes in the either condition (*M* = 94%). Accuracy in the color only and for color changes in the either condition were identical (both *M*s = 90%). In the critical binding condition, accuracy was significantly lower than all other conditions (*M* = 78%). The same pattern held for both sample sizes. These results seem to support the notion that capacity is highest when individual feature maps are used, rather than when bound object representations are required. To be sure the results were not due to using location as a test dimension, an additional experiment using color and shape tests replicated the basic effect (Wheeler & Treisman, 2002).

Although the binding condition clearly lowered accuracy, how it lowered accuracy is unclear. The difficulty could arise from at least 3 distinct sources: a) a lack of adequate encoding time, b) limited storage capacity for color-location bindings, or c) an increase in the difficulty of comparing bound object representations. The excellent performance in the 2-feature condition casts doubt on the first explanation, and so Wheeler and Treisman (2002) implemented a single-probe procedure of both the location-color and shape-color conditions. Results showed that performance in the critical binding condition (remembering location and color or shape and color) was no different than remembering any feature individually.

Instead of a reduction in capacity, an increase in comparison difficulty may explain the reduced accuracy for the binding condition. In a whole-display report, participants must compare each on-screen item with the items in memory before making a decision. In the binding

condition, this comparison is difficult because the stimuli are feature conjunctions and accuracy diminishes in the change detection procedure, just as reaction time increases in target searches for conjunction stimuli (e.g., Treisman & Gelade, 1980). To ensure the difference was not due to a reduction in the number of decisions being made, Wheeler and Treisman ran whole-display conditions with a single probe item cued at test. Using a cue in the whole-display report did not increase accuracy in the binding condition, showing that just the presence of additional items at decision time reduces performance. If attention is required to bind and maintain features and locations (Treisman & Gelade, 1980), the presentation of conjunction items in the probe array may reduce attention to the in-memory items causing them to become unbound and detection to fail. In the single-probe condition, the amount of complexity is not sufficient to cause the previously bound features to "fall apart" (Wheeler and Treisman, 2002). Rather than support the notion that VWM capacity is set by the objects, Wheeler and Treisman suggested that features are encoded in parallel into separate feature maps, and when mandated by the procedure, an attention-demanding binding process invoked. Searching for a change in an object defined by feature conjunctions reduces accuracy analogous to the increased reaction time in conjunction searches.

Johnson, Hollingworth, & Luck (2008) noted that a procedural difference between the 2-feature and binding conditions used by Wheeler and Treisman (2002) may have led to the wrong conclusion about the need for attention to maintain object-location bindings. In the 2-feature condition two items either changed location or changed color, but in the binding condition colors were re-assigned to previously used locations. Location changes in the 2-feature condition therefore provide strong changes to the overall stimulus configuration, leading to increased accuracy in the either condition, rather than reduced accuracy in the binding condition. Johnson,

et al. (2008) showed that when changes to the configuration are minimized (using color-orientation bindings rather than color-location bindings) attention can be diverted without causing objects to unbind, even in the whole-display report condition. An attempt to exactly replicate the binding condition from Wheeler and Treisman failed to show a difference between the binding and 2-feature conditions. Taken together, the results suggest that comparison difficulty does not uniquely impact bound object representations and focused attention is not always needed to maintain bound objects. Additionally, configural information may be used during the whole-display report when locations from the sample array are used in the probe array.

Capacity estimation in single-probe report procedures. Several of the previous studies converted raw accuracies to d' values to facilitate comparison between whole-display and single-probe report. To test if putative VWM capacity is identical between the tasks, it is useful to consider how capacity may be estimated in the single-probe procedure, to complement the whole-report formula developed by Pashler (1988). The formula offered by Pashler (1988) assumes that sample set size = probe set size, and is not appropriate for the single-probe procedure (Rouder et al., 2011). Instead, Cowan and others have suggested a discrete capacity formula for the single-probe procedure (Cowan, 2001; Cowan et al., 2005; Rouder et al., 2008; Rouder et al., 2011). As before, assuming that participants store *k* items on a given trial, the hit rate is the probability of correctly remembering the single probe item (*k/N)* plus the probability of guessing when the probe is not remembered [1 – *k/N*] * *g*, where *N* is the sample set size and *g* the guessing rate. The difference between the Cowan and Pashler estimates is the estimated guessing rate. In Pashler's formula, the guessing rate is set to be the empirical false alarm rate. Cowan and colleagues have suggested for the single-probe procedure that false alarms occur

only when the probe is not remembered ($1 - k/N$) and the participant guesses: $f = [1 - k/N] \times g$. Combining the hit rate and false alarm rate allows for a closed-form capacity expression:

$$k = N\ (h - f)$$

The constraints of $k \leq N$ and $h \geq f$ apply to Cowan's formula as well.

The debate on whether VWM stores features or objects continues to be hashed out, with recent research showing that features may be accessed without reference to an enclosing object (Bays, Wu, & Husain, 2011) and that features of a single object may be remembered/lost independently (Fougnie, & Alvarez, *in press*). Additionally, more tests have failed to replicate the initial effect from Luck and Vogel (1997) that features can be added to objects without a reduction in capacity (e.g., Olson & Jiang, 2002; Fougnie, Asplund, & Marois, 2010). Experiments that separate encoding processes from maintenance and storage, however, suggest that although *encoding* may operate on individual features, VWM proper (i.e., the set of processes responsible for storage and processing of currently available visual information) requires bound object representations (e.g., Quinlan, & Cohen, 2011; Woodman, & Vogel, 2008; Zhang, Johnson, Woodman, & Luck, *in press*). Essentially, features may be bound (and conversely not bound) into an object independently, but the process of consolidation into VWM necessarily creates objects, and only through these objects may feature information be accessed. Future work is needed to dissociate the neural basis of these processes to add support to the hypothesized representational differences (i.e., the brain selects features for encoding, but stores and retrieves bound objects). Change detection inherently requires encoding, storage, maintenance, retrieval, and comparison processes, requiring experimental precision to dissociate failures in one stage from another. The set of experiments proposed at the end of this review

explicitly equate encoding and storage demands while manipulating retrieval and comparison difficulty, allowing more pure assessment of the retrieval and comparison process.

Visual working memory as a flexible, continuous resource

Although Pashler's discrete capacity formula remains in use and has been given a recent analytical update (Rouder et al., 2011), it is notable that Pashler (1988) originally suggested that it was "admittedly crude in admitting no partial information and in attributing all errors to the maintenance, rather than the comparison, process" (p. 370) and that "the discreteness implied by the model is, no doubt, not strictly correct" (p. 372). The existence of partial (e.g., non-discrete) information is discussed in the current section, and the assessment of the comparison process, which forms the major impetus for the current set of experiments, is reviewed in the next section.

Whether the fundamental constraint on VWM capacity is considered feature or object-based, a great amount of current research is attempting to come up with accurate measures of *the* capacity of VWM. Having a single number with the explanatory power to predict performance across a broad swath of visual procedures is immensely appealing (cf., Cowan, 2001; 2010; Rouder, et al., 2008). In the name of capacity estimation, hierarchical Bayesian models have been developed (Morey, 2011), adjustments have been proposed for reducing variance (Kyllingsbaek & Bundesen, 2009) and increasing sensitivity (Ihssen, Linden, & Shapiro, 2010; Makovski, Watson, Koutstaal, & Jiang, 2010), nonhuman procedures have been developed (Elmore, et al., 2011; Gibson, et al., 2011; Heyselaar, Johnston, & Pare, 2011; Wright et al., 2010), and fMRI and EEG work has been conducted to find neural regions that correlate with VWM capacity estimates (Gao, Yin, Xu, Shui, & Shen, 2011; Song & Jiang, 2006; Todd & Marois, 2004; Xu, 2007; Xu & Chun, 2006). All of this work, however, presupposes that VWM capacity is, in fact, an estimable, discrete quantity.

An alternative account of VWM holds that capacity is an emergent quantity that results from selective allocation of attention processes (e.g., Bays, Wu, & Husain, 2011; Bays & Husain, 2008; Wilken & Ma, 2004). If attention is evenly distributed, then observers will extract *partial* information from *each* item, with precision proportional to the information in the display. Under this continuous model of VWM, there is no fixed capacity limit that underlies performance across a wide range of procedures; instead, a common resource allocation strategy may explain the observed performance similarities. Problematically, the predicted performance differences between the continuous and discrete models for standard change detection procedures are minimal in many situations—a lower fixed capacity estimate can be modeled as more noise in a continuous representation (cf. Elmore, et al., 2011). Furthermore, neurological evidence has been provided in support of both the discrete (e.g., Buschman, Siegel, Roy, & Miller, 2011) and the continuous (e.g., Ma, Beck, & Pouget, 2008) accounts of VWM. One way to move forward is to consider as *de facto* the simpler (in terms of how many parameters are typically involved) fixed capacity model, so long as it makes sounds predictions and adequately explains extant data (Cowan & Rouder, 2009; Rouder, et al., 2008). Three specific predictions and relevant data will be covered in this section: a) perfect performance when the sample set size is below capacity, b) no partial information is available, and c) objects are stored with a fixed resolution.

Ceiling performance. Wilken and Ma (2004) derived receiver operating characteristic (ROC) curves showing that the fixed-capacity assumption is equivalent to a threshold assumption. If an object's activation strength reaches the threshold, that object is stored completely, otherwise nothing about the object is stored. Beyond just being a threshold model, the fixed capacity view of VWM also specifies that a specific number of items can reach the

threshold at any given time. If the current task requires less than the capacity, then all the items are stored and the threshold assumption predicts that performance should be perfect.

Despite typical capacity estimates around 4, participants often make mistakes even when fewer than 4 items are in the sample array (e.g., Luck & Vogel, 1997). Under the fixed-capacity model, less than perfect performance with a 3-item sample set size indicates that capacity is necessarily less than three. Puzzlingly, the same participant may do quite well with 8-item sample sizes, arriving at a capacity estimate closer to five. At face value, these data pose a problem for VWM capacity as a fixed quantity. The important issue, however, is to determine whether the errors at lower sample set sizes are caused by a reduction in storage capacity or other, unrelated phenomena (e.g., blinking, response confusion).

Rouder et al. (2008; 2011) added an attention parameter to account for all-or-none attention. Although the single parameter is estimated across all trials (rather than a separate attention parameter per sample set size), the modification adequately accounts for shifts in ceiling-level performance without allowing the shape of the predicted accuracy function to vary widely. Ensuring that a model cannot fit arbitrary data patterns provides a shield against claims of the model being overly general (Roberts, & Pashler, 2000). The fixed capacity with binary attention model was used recently to compare estimates of VWM capacity in pigeons and humans (Gibson, et al., 2011). By controlling for (hypothesized) attentional lapses, Gibson et al. found stable capacity estimates over a range of change conditions for pigeons that could be directly compared to capacity estimates for humans. Without controlling for the relatively low performance by the pigeons in the simplest condition, the estimate for the pigeons may have been artificially low, potentially contaminated by non-task-related issues. The attentional lapse rate was much higher for pigeons than humans and the average human capacity of 3 items

(nearly 2 for the pigeons) is well within the normative range, indicating some face validity for the parameter.

Partial information effects. If VWM stores whole, discrete objects, then observers should be unable to report feature information from non-remembered objects. On the other hand, if VWM stores partial information across multiple objects, then observers should be able to provide a low-precision estimate (i.e., one with a high standard deviation) of each item in the sample array. Crucially, the precision should be inversely proportional to information (e.g., number of items and their complexity) in the sample array. To test VWM precision, Zhang and Luck (2008) used a color recall task in which participants selected a color from a color wheel to report on their memory of the sample array (cf. Wilken & Ma, 2004). After briefly viewing the sample array and a blank probe delay, participants saw the color wheel enclosing the locations from the sample array, now demarcated by square outlines. One of the outlines was thicker than the others, indicating the recall location. Zhang & Luck (2008) showed that performance was consistent with participants having excellent memory for a fixed number of items and zero information about other items. They modeled performance as a 2-component mixture distribution: a Gaussian for precision of stored information (the noise coming from encoding, not maintenance) and a uniform distribution to represent guessing when the items were not stored. Comparing performance for sample set sizes of 3 and 6 showed that the width of the Gaussian stayed constant, whereas the contribution of the uniform distribution increased, consistent with the discrete, fixed resolution view of VWM. Similar results were obtained with an additional experiment using an analogous shape recall task (Zhang & Luck, 2008). Zhang and Luck (2009) extended the mixture modeling approach to show that these memory representations do not exhibit slow decay over time, but rather exhibit fixed resolution, all-or-none storage. Precision

for color memory (as measured by the width of the Gaussian component in the mixture model) was statistically equivalent across probe delays of 1, 4, and 10 seconds, but the contribution of the uniform guessing distribution increased significantly from the 4 to 10-second probe delay.

Bays and Husain (2008) assessed precision of VWM by varying the magnitude of location and orientation changes. The fixed-precision assumption assumes that performance at below-capacity sample set sizes should be not be affected by the magnitude of changes (assuming the change is supraliminal). Instead, Bays and Husain showed that accuracy was affected by change magnitude at all sample set sizes and relative precision was best fit by a power law relationship with inverse sample set size. To measure resource allocation, eye fixations were specifically controlled. Participants were given instructions to make saccades in a given direction from fixation. Eye tracking data showed a strong precision benefit for the saccade target (i.e, higher-resolution memory for the upcoming fixation), further suggesting that participants flexibly allocate attentional resources, rather than extract a fixed number of items from each display.

Cowan and Rouder (2009) fit the data from Bays and Husain under the fixed-capacity, discrete VWM assumption (with the attention parameter from Rouder, et al., 2008; 2011) and showed equivalent or better fits for the discrete over the continuous model. To account for the performance difference at below-capacity set sizes, a "slot averaging" mechanism (Zhang & Luck, 2008) was added. The premise of slot averaging is that if a given display has fewer items than the capacity limit, duplicate items will be placed in the empty slots. The duplicates are averaged upon retrieval, allowing a sharper (higher resolution) representation than would be achievable if all slots contained a unique object. Bays and colleagues have argued that the discrete models of VWM cannot account for the reduced precision in serial presentation tasks

and there is insufficient evidence to justify adding the attention parameter and slot averaging mechanism (e.g., Bays & Husain, 2009; 2011; Bays, Marshall, & Husain, 2011; Gorgoraptis, Catalao, & Bays, 2011). In defense of the discrete view, more work has been done to show a neurological basis for slot averaging using in ERP and behavioral designs (e.g., Anderson, Vogel, & Awh, 2011) and the capacity-estimation formulas have received a more formal treatment (e.g., Morey, 2011; Rouder et al., 2011). In sum, the debate on the existence of partial information surrounds the number and validity of model parameters, as both the discrete and continuous models can explain a wide range of data.

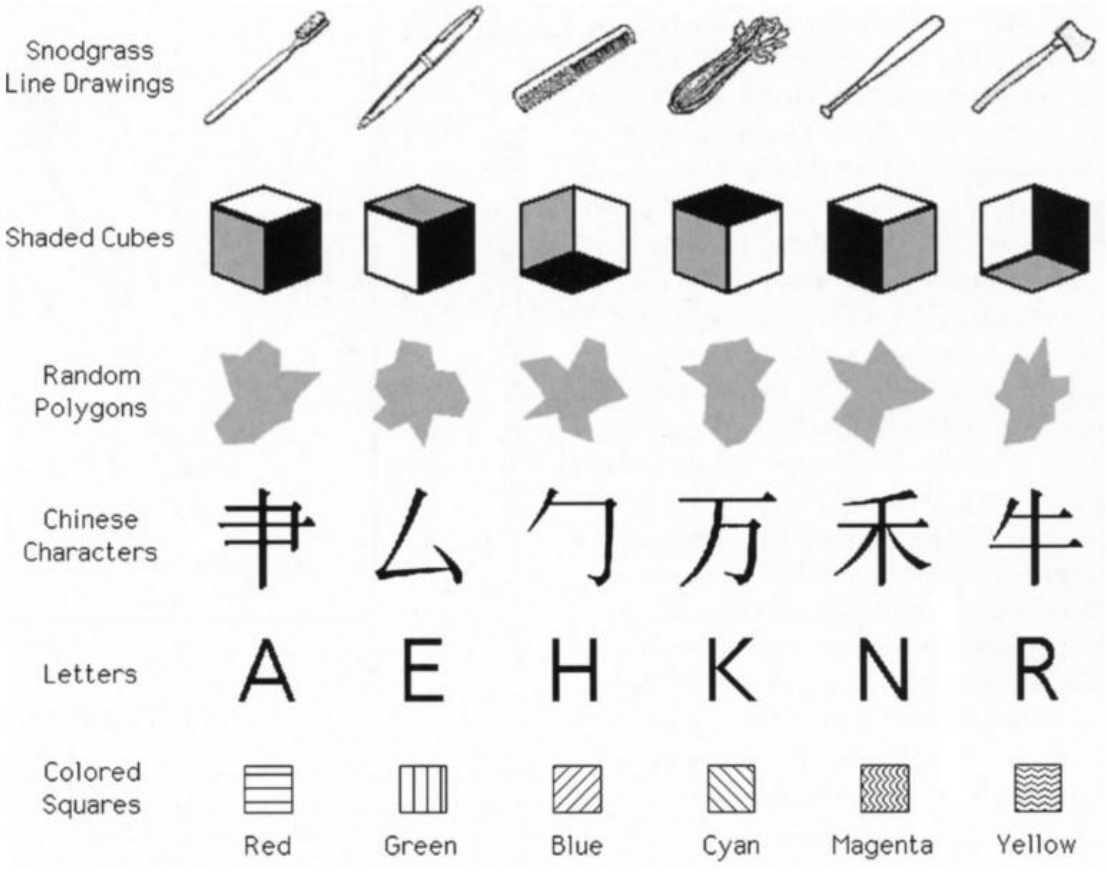


Figure 4. Example stimuli used by Alvarez and Cavanagh (2004) to assess the impact of object complexity on change detection accuracy.

Fixed object resolution. The assumption of fixed-resolution object representations follows directly from the assumption of all-or-none storage, as discussed in the previous section. The issue in the current section is to present data on how the complexity of the sample array impacts capacity estimation. In the strictest view of all-or-none storage, each object is stored without error. Object representations should have a snapshot-like quality, the representation exactly reproducing the features of the stored object. Taking into account noise during encoding, the discrete model does not require that VWM store exact replicas of the items from the sample array, only that the resolution of the stored items be independent of the number of items presented. As previously mentioned, slot averaging allows an object's resolution to be increased, therefore the discrete model only predicts that resolution will not decrease when the number of items in the sample array continues to grow past capacity. For example, given a VWM capacity of 4, items from a 2-item sample array will be stored with double precision. Increasing the sample set size to 4 means the items will be stored with single precision. The prediction of the model is that increasing the sample set size to 12 items will not change the precision of the 4 items stored. The continuous view of VWM argues that the resolution of each stored item should decline across sample set size according to a power law.

Alvarez and Cavanagh (2004) tested the fixed-resolution assumption directly by assessing change detection accuracy with stimuli of high and low degrees of visual information/complexity (e.g., shaded cubes and Chinese characters vs. colored squares and letters). Information load was measured independently by looking at target search slopes (steeper slopes indicating a more difficult search). In addition to calculating a capacity estimate for each stimulus class, they also estimated a 75% threshold criterion, which is the sample set size corresponding to 75% accuracy. Across sample set sizes from 1 to 15, performance in the task

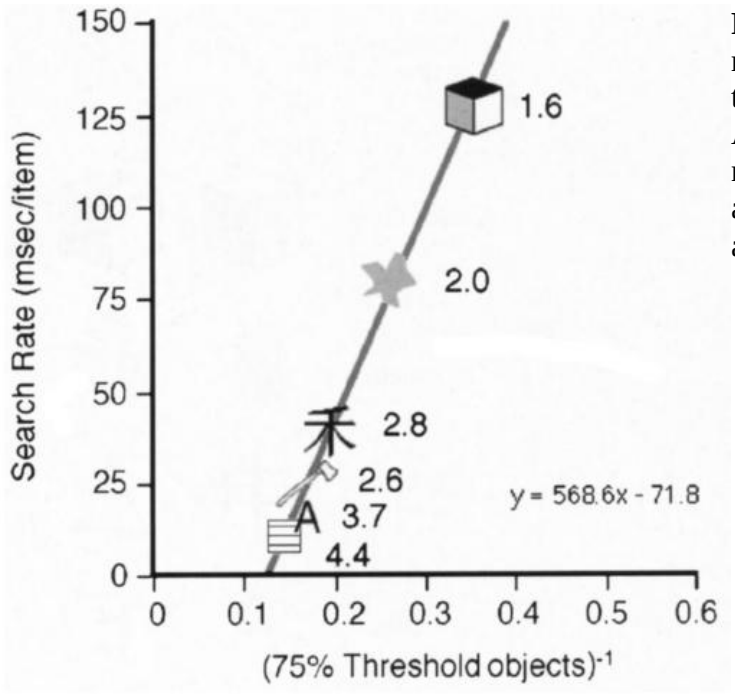


Figure 5. Relationship between search rate and the inverse 75% threshold for the different stimulus classes in Alvarez and Cavanagh (2004). The numbers beside each class exemplar are capacity estimates calculated according to Pashler (1988).

was clearly impacted by the information load in the task, with the linear regression of search rate on 1 / (75% threshold) explaining over 99% of the variability in search rate. Control experiments manipulated sample-probe item similarity by using a rotated block numbers 2 and 5, and showed again that the inverse 75% threshold had a strong linear relationship with search rate. An additional test with longer viewing times (out to 850ms) showed that performance with the more complex objects was not limited by encoding time. A similar study by Eng, Chen, and Jiang (2005) showed that allowing arbitrary increases to viewing time reduced, but did not eliminate, the correlation between information load and VWM capacity. The authors suggested that setting viewing time between 1x or 2x the search time (assuming participants are conducting self-terminating searches through half the objects on average) may be a useful way to ensure adequate encoding time in change detection. Taken together, these results suggest that the resolution of VWM is limited (as suggest by the previous section on fixed-resolution), and

perhaps more complex objects take up more slots, leading to an effective reduction in VWM capacity.

Assessing capacity and comparison difficulty in change detection

Instead of concluding that VWM capacity is modulated by object complexity, it is also possible that accuracy is being reduced by other factors, and the reduction of capacity is a side effect of the reduced accuracy. As noted earlier, change detection performance is limited by encoding, storage, maintenance, retrieval, and comparison processes. A failure anywhere in this pipeline would lead to reduced capacity estimates. Eng et al. (2005) showed that neither encoding limits or maintenance (via a probe delay manipulation) fully explain the relationship between object complexity and VWM capacity. This still leaves the possibility that a difficulty in the retrieval or comparison process is responsible for the lowered capacity estimates, as well as the search slopes. In accord with this hypothesis, Awh, Barton, and Vogel (2007) argued that the increased object complexity in the earlier studies was confounded with increased sample-test similarity. Using stimuli similar to Alvarez and Cavanagh (2004), Awh et al. (2007) also manipulated the kind of change that occurred--either a within-category change (e.g., cube to cube), or a between-category change (e.g., cube to Chinese character). At a given sample set size, the between-category change condition still requires the same amount of information to be stored, but it makes the comparison much easier because the similarity of items from different categories is less than the similarity of items within a category. Additionally, Awh et al. also used a single-probe report procedure, reducing the impact of a counting strategy and further reducing the perceptual complexity of the probe array. Using this design, change detection performance using simple objects was significantly correlated with the between-category change condition only. Additional studies using mixed stimulus classes (e.g., sample arrays comprised of

Chinese characters and shaded cubes) within a given trial again found no decrease in capacity estimates, but instead that limitations in object resolution constrained performance (Barton, Ester, Awh, 2009). Importantly, resolution was shown to be independent of the complexity of the sample array, arguing in favor of a discrete, fixed-resolution VWM.

A potential problem with the Awh et al. (2007) procedure is that the cross-category changes may have encouraged participants to code the category of the objects rather than the visual features at each location (Olsson & Poom, 2005). Hollingworth (2003) noted that at least one study using cross-category changes (Simons, 1996) found a decrement in performance when a verbal distractor procedure was given, suggesting that verbal category labels may play a role. Importantly, Hollingworth found no effect of verbal interference in a natural object change detection procedure for within-category changes. If comparison difficulty between the sample and probe arrays is to explain the reduced accuracy (and therefore capacity estimates) for detecting changes in complex objects, it should be possible to show the impact of comparison difficulty using simple objects and within-category changes.

Chapter 2: Experiments

Experimental overview

The goals of the following experiments were a) to assess if increased comparison difficulty significantly impairs performance in a color change detection procedure and b) to what extent individual VWM capacity estimates are distinct from resistance to interference at retrieval. A vital part of accomplishing the latter step was the development of a principled fixed-capacity model to provide a lower bound estimate of VWM capacity for forced-choice change detection with varying probe set size. Using colored squares increases the likelihood that the sample/probe comparison is not resolution limited. The forced-choice procedure was used to ensure that participants searched for an item on every trial and to remove the possibility that participants develop a bias toward a given response (e.g., always reporting "change"). Additionally, results from Makovski et al. (2010) suggest that the forced-choice procedure may be particularly sensitive to comparison difficulty.

Previous work looking at the impact of probe set size also repeated stimuli in the sample array which may have increased reliance on location information, rather than dissociating object and location memory (e.g., Jiang et al., 2000; Theeuwes, 2004; Davis & Leow, 2005). Other work that did not repeat stimuli (e.g., Hyun, et al., 2009) used sample set sizes too small to obtain useable capacity estimates (because of the upper limit on capacity estimates noted above). Finally, no work has been done to derive discrete capacity estimates for the forced-choice procedure for probe set sizes other than 2, and there is reason to doubt the validity of a currently used method because of its failure to account for storage dependencies (e.g., Eng et al., 2005; Elmore et al., 2011).

Experiment 1

The goal of Experiment 1 was to test whether increasing the number of items in the probe array modifies change detection accuracy. Trial types were created to replicate the accuracy decline across sample set size (e.g., Figure 3, a, right panel) and to show the added difficulty of having to compare items in memory with additional on-screen items. Although this effect has been measured by previous studies (e.g., Makovski, et al., 2010), the current method assesses a broader range of the parameter space without repeating colors within the sample array. These modifications provide a stronger demonstration of the impact of comparison difficulty on array sizes large enough to estimate VWM capacity and without straightforward color-grouping encoding strategies.

Method

Participants

Participants (N=24) were students enrolled in a psychology class at Auburn University, ages ranging from 18 to 24. Participants received course credit for participation. Parental consent was obtained for all participants 18 years of age; all other participants provided informed consent prior to the beginning of the experiment. The Auburn University Institutional Review Board approved all protocol details.

Apparatus

Participants were tested individually in an unlit room. They were seated 30-cm away from a 17-in touchscreen LCD monitor (1280x1024, 60Hz). Experimental events were controlled with custom software written in C++ using OpenGL. All responses were recorded using the touchscreen.

Stimuli

The stimuli were 14 colored squares created in Adobe Illustrator ®: aqua (RGB: 59, 255, 255), blue (RGB: 59, 59, 255), gray (RGB: 128, 128, 128), green (RGB: 32, 132, 33), lime (RGB: 0, 255, 1), navy (RGB: 1, 0, 128), orange (RGB: 254, 165, 0), magenta (RGB: 255, 0, 254), red (RGB: 237, 27, 46), white (RGB: 255, 255, 255), yellow (RGB: 254, 242, 0), yellow-green (RGB: 149, 197, 49), purple (135, 50, 180), and brown (RGB: 139, 69, 19). Each square subtended a visual angle of 3.25° x 3.25°. At the start of each trial, a small white fixation cross (subtending a visual angle of 0.6° x 0.6°) was displayed at the center of screen. Sample and probe array stimuli were displayed at imaginary points on one of three invisible concentric rings of radius 4.90°, 9.78°, and 14.63°, containing stimulus locations every 90°, 45°, or 30°, respectively for a total of 24 possible stimulus locations. Stimuli were jittered according to a uniform [–0.5°, 0.5°] distribution along each axis before the presentation of the sample array.

Procedure

Participants read procedure instructions on screen, and had an opportunity to ask questions about the procedure. The instruction screen read:

> On the next screen you will see multiple colored squares for a short amount of time. After they disappear some of them will reappear. Your task is to touch the square that is CHANGED from the first display. Accuracy is more important than speed in this task. You will not receive any feedback about your performance, simply try your best.
>
> Touch anywhere on the screen to begin.

Participants completed 6 practice trials before beginning the first block. Prior to each block the instructions were repeated and participants were allowed to ask questions about the procedure. Each 480-trial session included 2 blocks of 240 trials. Each block contained 20 repetitions of the 12 trial types, defined by a specific sample and probe set size pair (e.g., 4:3, representing 4 items in the sample array and 3 items in the probe array). For sample set sizes 4, 5, and 6, there are 3 (4:4, 4:3, and 4:2), 4 (5:5, 5:4, 5:3, and 5:2), and 5 (6:6, 6:5, 6:4, 6:3, and 6:2) probe set sizes, respectively, leading to the 12 trial types. Trial types were ordered randomly throughout each block.

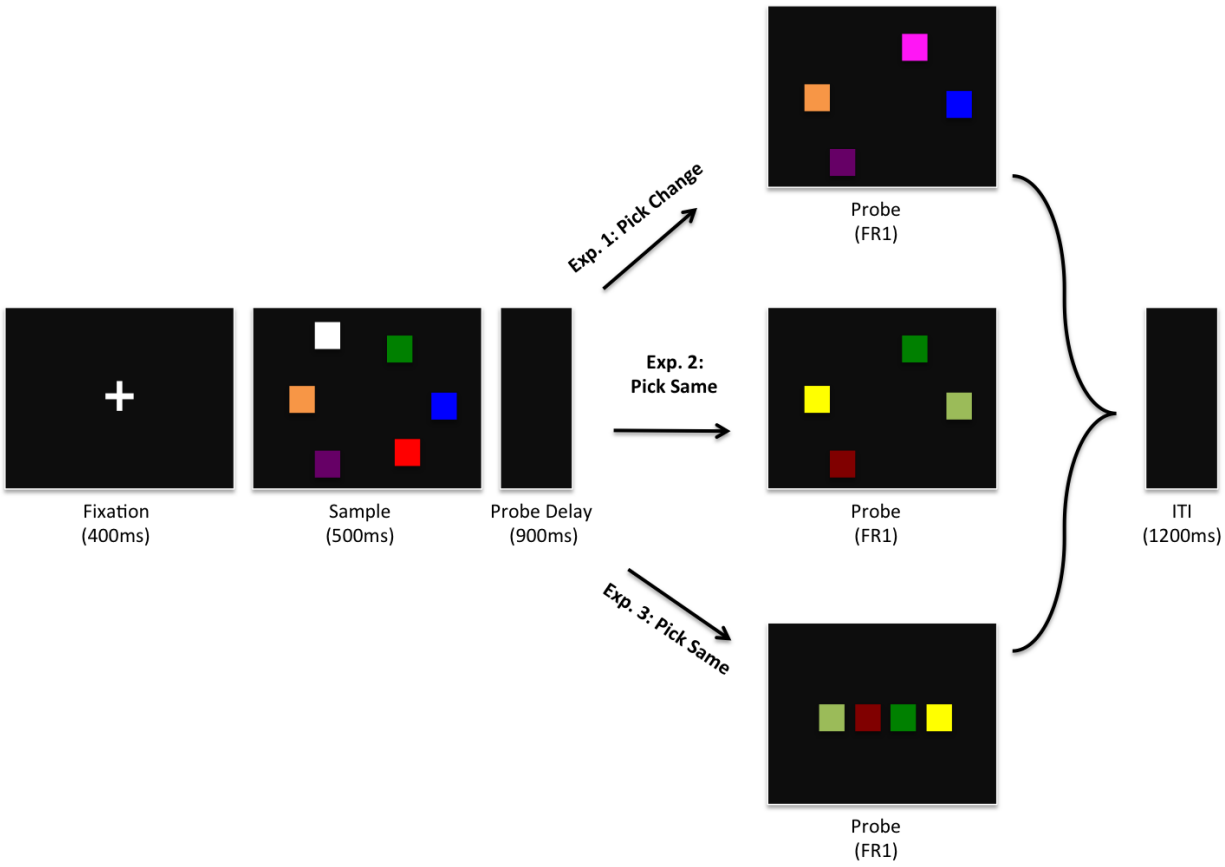


Figure 6. Example trial progression for each experiment. Note that the primary differences across experiments are the probe array composition (single changed item vs. single same item) and the probe array arrangement (re-using locations vs. horizontal alignment). FR: fixed ratio (touch response). ITI: intertrial interval.

At the start of each trial (Figure 6), a small fixation cross appeared for 400ms, followed by the presentation of the sample array containing either 4, 5, or 6 uniquely-colored squares for 500ms, then a blank probe delay of 900ms, and finally the probe array. Probe arrays contained a subset (or possibly the entire set) of items from sample array (e.g., for sample set size 5, potential probe set sizes were 5, 4, 3, and 2). After the participants made a choice, a 1200-ms intertrial interval commenced. No feedback was provided after choices and each block was separated by a short break.

A principled approach to forced-choice capacity estimation

As part of analyzing data from Experiment 1, discrete, fixed-capacity formulas were developed for the forced-choice partial- and whole-display report conditions. The formulas derived for the yes/no procedure are not applicable because the forced-choice procedure always contains a change, requiring the participant to choose an item from the probe array. Consider a 4-item sample array and a 2-item probe array that contains a single change. If the participant stores 3 items from the sample array in memory, there is a 25% chance that the changed item is not stored. Even if the participant has not stored the changed item, the participant can determine which item changed via an elimination rule—choose the only not remembered item. This strategy will not work in the yes/no procedure, because there is a 50% chance that no item changed. To estimate capacity in the forced-choice procedure, Eng, et al. (2005) present a formula based on Pashler (1988). To miss a forced-choice trial, the participant must not store either item, which Eng and colleagues report as being equal to the square of the probability that just the changed item is not stored. If we let ~C be the event that the changed item is not stored and ~U be the probability the unchanged item is not stored, their formula calculates P(~C and

~U) = P(~C) × P(~U). Because the participants do not know which item will change, it is reasonable to assume the marginal distributions are equivalent, P(~C) = P(~U). Eng et al. then conclude that P(~C and ~U) = P(~C)^2. The assumption, however, of the independence of the events ~C and ~U is not reasonable for the fixed-capacity model.

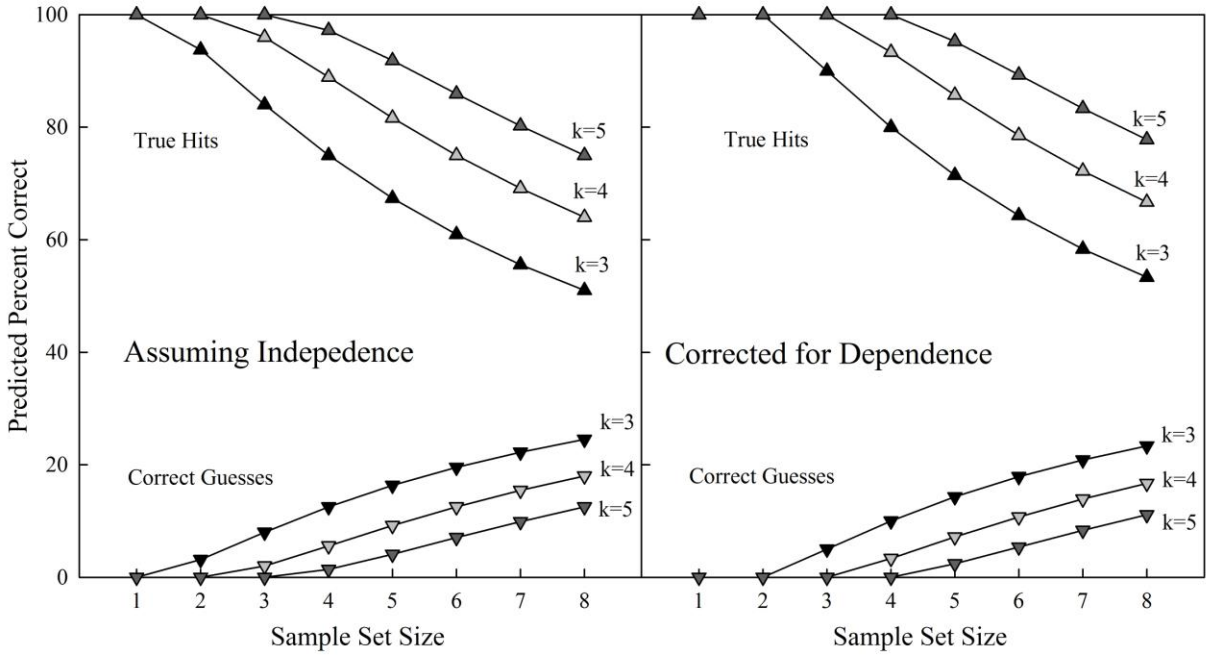
If we assume that participants store *k* items from each sample array, then the probability of having stored the unchanged item in the probe array must be updated based on the assumption that the changed item has *not* been stored. If VWM stores exactly *k* items (no more and no less), the probability the changed item was not stored is P(~C) = (*N* – *k*) / *N*, as before. To find the joint probability that the changed and unchanged item are not in memory requires updating the probability of the unchanged item not being stored based on assuming the changed item was not stored:

$$P(\sim U \cap \sim C) = P(\sim U | \sim C) \times P(\sim C) = \frac{(N-1)-k}{N-1} \times \frac{N-k}{N}$$

The important point is that P(~U | ~C) ≠ P(~U) and therefore independence is violated. Predicted hit rate is then calculated as the sum of the correctly remembered trials and those not remembered but correctly guessed (correct guessing rate is set to 0.5):

$$\mathrm{P(H)} = \left\{1 - \left[\frac{(N-1)-k}{N-1} \times \frac{N-k}{N}\right]\right\} + \left[\frac{(N-1)-k}{N-1} \times \frac{N-k}{N}\right] \times 0.5$$

Figure 7 shows the predicted hit rate and guess-correct rate for the formula from Eng et al., (2005; left panel) and the current formula (right panel). The current formula implies that performance should be at ceiling for $k \geq N - 1$, rather than for $k \geq N$. This result captures the intuition that participants can use an elimination rule to determine the changed item if they are able to store the unchanged items displayed at choice time. Although one could argue that less-



than-ideal observers may fail to utilize the elimination rule, the current estimate still serves to establish the true lower bound for the number of stored items in the forced-choice procedure.

Figure 7. Estimated values from two forced-choice capacity models. The left panel shows the formula used by Eng et al (2005) and Elmore et al (2011) that incorrectly assumes independent storage of items. The right panel shows the correction for dependence. For each graph, *k* represents capacity.

A combinatorial approach was adopted for clarity as part of generalizing the capacity formula to larger probe set sizes. First, note that De Morgan's Law allows us to rewrite the event (~U and ~C) as ~(U or C). For a sample set size of *N* and capacity of *k*, the probability of not storing either the unchanged or changed item is the probability of choosing *k* of the $N - 2$ items remaining: P(~Store) = $\binom{N-2}{k} / \binom{N}{k}$. The hit rate formula for probe set size 2 then becomes, P(H) = P(Store) + P(~Store)**g*, with P(Store) = 1 – P(~Store) and *g* the guess-correct rate (0.5 for probe set size 2).

When more than 2 items are in the probe array, optimal observers can use information about the unchanged items to guide their search of the probe array. For example, given a sample set size of 6, a probe set size of 3, and a capacity of 4, the observer is guaranteed to recognize at least one of the items in the probe array. The recognition of this item allows the observer to exclude it from the set of possible guesses, similar to how a student may eliminate known incorrect options on a multiple-choice test. To account for this strategy, the guessing rate must be updated based on the number of items in the probe array not stored. Letting *i* be the number of items recalled from a probe array with probe set size *p*, the guess correct rate becomes:

$$g(N,p,k) = \sum_{i=\max(0,-1\times(N-p-k))}^{\min(k,p-1)} \frac{1}{p-i} \frac{\binom{N-p}{k-i} \times \binom{p-1}{i}}{\binom{N}{k}}$$

Looking within the summation, the first term captures the guess probability. If an observer failed to store any of the items that are in the probe array, this value reduces to 1/*p*. The terms in the numerator consider first the number of ways to select items that are not in the probe array and second the selection of unchanged items from the probe array. The denominator represents the total number of ways to select *k* items from the sample array. Considering the bounds on the summation, the initial value for *i* reflects the fact that on trials where $N - p < k$, some items (exactly –1*(N – p – k)) from the probe array must have been encoded. The terminal value for *i* ensures that the number of items recalled be no more than an individual's capacity as well as the total number of unchanged items in the probe array. In words, the above equation considers all possible ways to not store the changed item, and then weights those combinations by an appropriate guessing factor. It should be noted that disregarding the guess-correct rate [1/($p - i$)], the summation above for $p = 2$ is just P(~Store) from the earlier formulation. The general hit rate formula is therefore written as P(H) = P(Store) + g(*N*, *p*, *k*). As suggested by Rouder et al. (2008;

2011), an attention parameter should be used to account for lapses in attention. The general forced-choice hit rate formula with attentional lapse parameter, $a \in [0, 1]$, becomes:

$$P(H) = a\big(P(Store) + g(N, p, k)\big) + (1 - a)\frac{1}{p}$$

Figure 8 shows how the formula builds expected hit rates across probe set size and capacity, fixing sample set size at 6 for illustrative purposes. Considering first the know rate calculation, these lines show no effect of probe set size and a linear effect of capacity. As expected, only an observer with capacity 6 will "know" the answer on all the trials, but a 0-capacity observer will never "know" the right answer. The top right panel shows the guess-correct rate. This rate is the proportion of trials that are answered correctly based on a guessing strategy. The primary feature of the current capacity estimation model is the heavy influence of probe set size on guessing strategy, as well as its interaction with capacity. For individuals with a 6-item capacity, none of trials are guessed correctly, as a consequence of having a perfect know rate. At capacity 5 there is still no effect of probe set size as the optimal observer can use the elimination rule to guess correctly regardless of probe set size. Interestingly, the proportion of guess-correct trials for capacity 1 is identical to that for the capacity 0 line. This counter-intuitive result arises because an observer with capacity 1 guesses on 5 out of 6 trials with perfect accuracy on 1/6 (via the elimination rule) and 50% accuracy on the remaining 4 out of 6 trials. Thus, 1/6 + 2/6 = 0.5 proportion of guess-correct trials. For the 0-capacity observer, guessing correct results from 50% accuracy on all 6 trials.

The lower panel of Figure 8 shows the overall predicted hit rates. These curves take into account an attentional lapse parameter (set at 0.9) that serves to mix contributions of the "know" and "guess-correct" trials with random guessing trials in which the guess rate is just 1 / probe set size. For the overall hit rate graph, note the identical predictions for capacities of 5 and 6. This

identity reinforces the previously mentioned upper limit on capacity estimates, such that $k \leq N - 1$. The shape of these overall hit rates show the impact probe set size should have on accuracy according to the fixed-capacity model. The large differences underscore the importance of accurately estimating the guessing parameter when assessing VWM capacity.

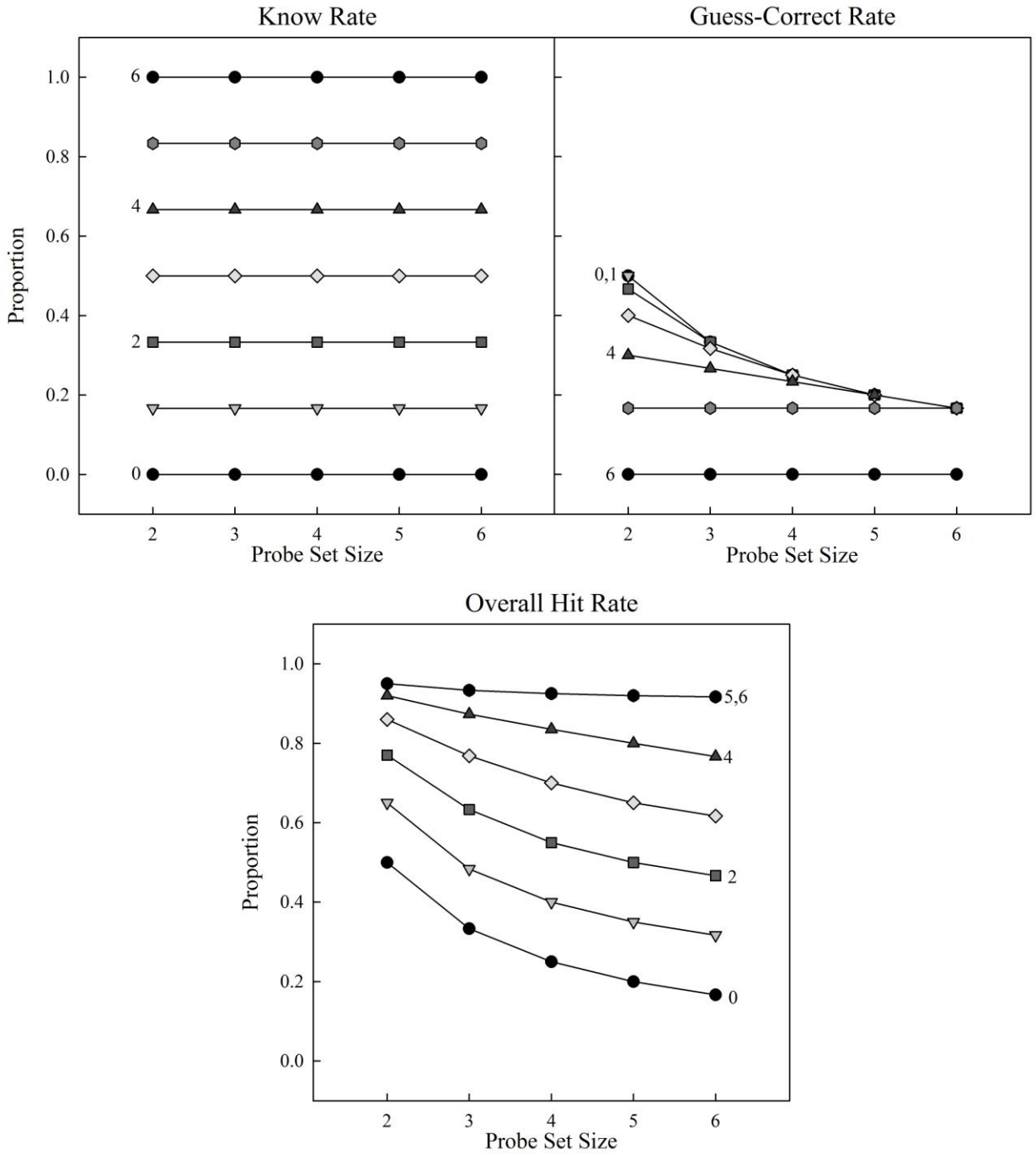


Figure 8. Predicted values for "know," "guess," and the overall hit rate for the forced-choice change detection task across varying probe set sizes. For the overall hit rate, the attention parameter was set at .9. Capacity values are listed beside the curves. The 0-capacity line represents chance performance. Values separated by commas denote lines that superimpose.

Fitting the fixed-capacity model to the current data. In each of the following experiments, estimates for the capacity and attention parameters were fit to each participant's average accuracy data using a two-step process. Initial attempts at simultaneous optimization of both parameters showed too much sensitivity to initial parameter guesses. To counteract this instability, capacity was set to successive values from 0 to 10 and the value of the attention parameter that produced the best fit (i.e., the one with the highest log-likelihood) was stored. The fits for each of these 11 (one for each potential capacity) models were then compared and the capacity and attention parameter from the overall best fitting model were retained. The instability of the estimates raises a larger question about the use of such techniques for estimating VWM capacity, as a lower (or higher) capacity estimate combined with a higher (or lower) value for the attention parameter can be produce very similar log-likelihood values. This finding is not unique to the current procedure or the broader study of VWM, but serves as a cautionary note to users interested in fitting models to cognitive processes.

As a slight departure from other capacity estimation techniques (e.g., Rouder et al., 2011), the current study considers only integer-valued capacities. Only integer values of capacity were allowed for two reasons. On theoretical grounds, it is hard to justify the strong object hypothesis (i.e., that participants are storing whole objects) if the fixed-capacity estimate is not a whole number. Even if participants are storing individual features, it is difficult to describe how part of a feature value could be stored (in the current study, the "objects" contain only a single feature, so this distinction is, in practice, moot). The second reason is related to implementation: the combinatoric approach taken to model VWM selects whole objects from the sample array into memory and does not work with arbitrary rational values.

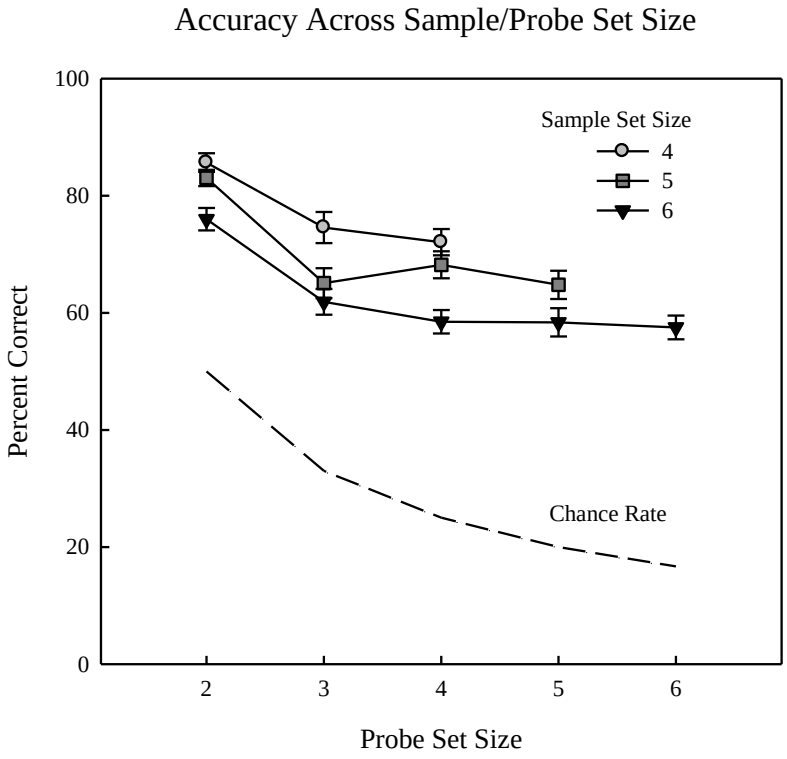


Figure 9. Changes in accuracy across sample set size (separate lines) and probe set size for Experiment 1. The dashed line represents chance performance, calculated as 1 / probe set size. Error bars are ±SE.

Experiment 1 Results and Discussion

Mean accuracy data from Experiment 1 are shown in Figure 9. Across increases in sample set size and probe set size accuracy declined steadily, but was always well above the chance rate line. Across probe set size, sample set sizes 4 and 6 show smooth negatively accelerated slopes indicating that the effect of probe set size may not be strictly linear. Accuracy for sample set size 5 shows less consistent effects across probe set size, but shows the same general pattern as 4 and 6.

To measure the reliability of accuracy changes across sample and probe array size, a 3x3 Repeated Measures (RM) Analysis of Variance (ANOVA) of Probe Set Size (2, 3, 4) and Sample Set Size (4, 5, 6) on Accuracy was conducted. Note the truncated range of probe set sizes used because the factors may not be fully crossed in this design (e.g., there cannot be a probe set

size of 5 with a sample set size of only 4, but see Experiment 3). The RM ANOVA yielded significant effects of sample set size [$F(2, 44) = 58.695$, $p < .001$, $\eta_p^2 = .727$], probe set size, [$F(2, 44) = 70.956$, $p < .001$, $\eta_p^2 = .763$], but not the interaction [$F(4, 88) = 2.076$, $p = .091$]. The lack of any interaction suggests that any added difficulty caused by increasing probe set size is independent of the sample set size within the range of probe set sizes analyzed (2, 3, and 4). Trend contrasts for sample set size on accuracy show a reliable linear trend [$F(1, 22) = 119.56$, $p < .001$, $\eta_p^2 = .845$], but not a quadratic trends [$F(1, 22) = .467$, $p = .502$]. Trend contrasts for probe set size on accuracy show reliable linear [$F(1,22) = 113.690$, $p < .001$, $\eta_p^2 = .838$] and quadratic trends [$F(1, 22) = 28.877$, $p < .001$, $\eta_p^2 = .568$]. Because these trend analyses only consider a subset of the sample/probe set size pairs, these conclusions need to be confirmed across the rest of the data.

In order to leverage all of the probe set sizes, polynomial trend contrasts were conducted across probe set size on accuracy within each sample set size. For sample set size 4, the contrasts showed a clear linear effect of probe set size [$F(1, 22) = 30.999$, $p < .001$, $\eta_p^2 = .585$], but an unreliable quadratic effect [$F(1, 22) = 3.994$, $p = .058$]. For sample set size 5, strong evidence was observed for linear [$F(1, 22) = 46.934$, $p < .001$, $\eta_p^2 = .681$], quadratic [$F(1, 22) = 17.049$, $p < .001$, $\eta_p^2 = .437$], and the cubic trend [$F(1, 22) = 27.062$, $p < .001$, $\eta_p^2 = .552$]. The presence of a quadratic trend shows that the effect of increasing probe set size is nonlinear on accuracy, indicative of proportional (rather than absolute) increase of comparison difficulty as probe set size increases, or the use of configural information at larger probe set sizes. Although the cubic trend was reliable and had a moderate effect size, there is no straightforward explanation for it. Perhaps the increase in comparison difficulty and decrease in chance rate cause a quadratic decline in accuracy, but the increase in configural information at the largest probe set size leads

to an overall increase. Sample set size 6 mimicked the pattern for sample set size 5, showing reliable linear [$F(1, 22) = 94.519$, $p < .001$, $\eta_p^2 = .811$], quadratic [$F(1, 22) = 32.782$, $p < .001$, $\eta_p^2 = .598$] and cubic trends [$F(1, 22) = 4.792$, $p = .039$, $\eta_p^2 = .179$]. The quartic trend was not reliable [$F(1, 22) = .075$, $p = .787$]. The nuanced explanation for the cubic trend in sample set size 5 is bolstered by the presence of the same trend for sample set size 6. It is unlikely to be idiosyncratic if it appears on independent slices of the dataset with reasonable effect sizes.

Although the contrasts show the nonlinear effect of probe set size, they do not directly address the issue of whether participants are encoding configural information from the sample display. One way to measure the benefit of configural information is to compare accuracy between the last two probe set sizes within each sample set size (e.g., the change from 6:5 to 6:6). If there is an increase in accuracy, then the individual is said to have benefited from the presence of configural information (or a configural strategy). A decrease would mean that whatever benefit there may be from configural information is overshadowed by the increase in the number of items in the probe array. The decrease may be due to increased comparison difficulty or more simply the corresponding decrease in the chance rate. The left panel of Figure 10 shows the configural benefit for participants in Experiment 1. The first thing to notice is the great degree of variability among participants—some show benefits of configural information, whereas others show zero or negative effects.

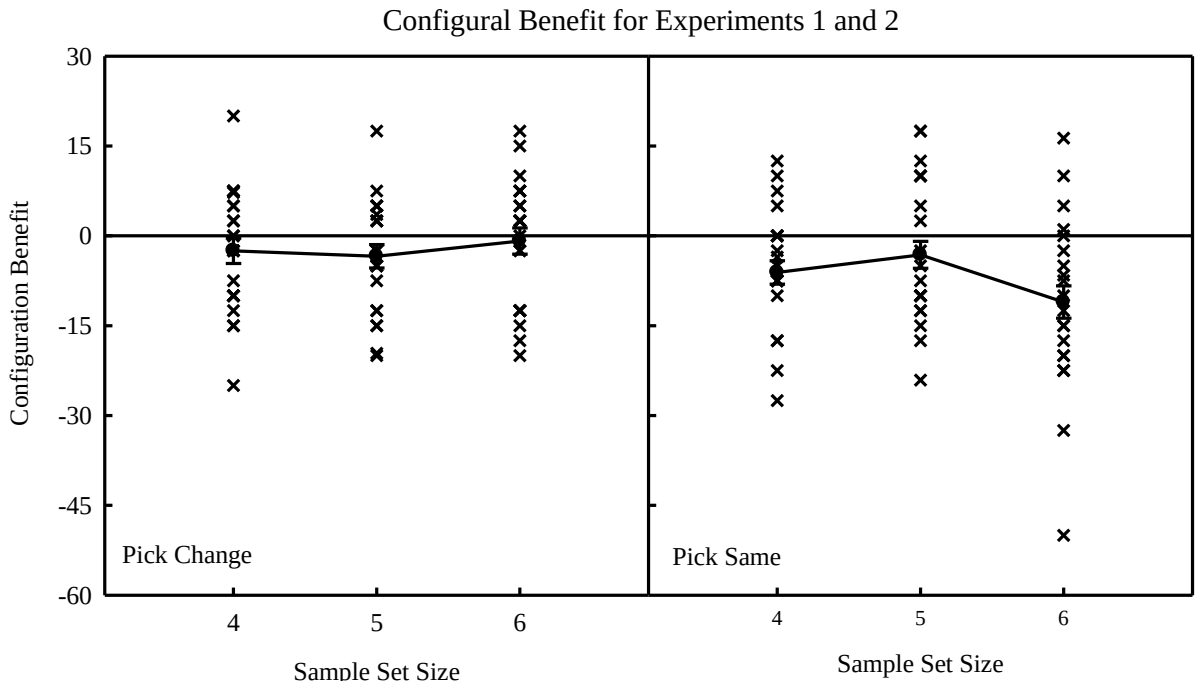


Figure 10. Estimate of configural benefit across Experiments 1 (left panel) and 2 (right panel). Configural benefit is calculated as the difference between accuracy for the highest and second highest probe set size within each sample set size. Positive values indicate an increase in accuracy for whole display report trials as compared to probe arrays with only one unused location from the sample array. Individual x symbols represent individual values, and the filled circles show the mean for each sample set size. Error bars are ±SE.

To test if individuals are using a consistent configural strategy, the configural benefit across sample set size was analyzed for linear correlations. No sample set size pair showed evidence of a reliable linear relationships (all rs < .333, ps > .12) The lack of a correlation between any of the sample set sizes suggests that participants either were not or could not apply a consistent configural/item-specific encoding strategy across sample set sizes.

A flaw in the accuracy-based measure, however, is that it does not take into account the decreased chance value at larger probe set sizes. Alternatively, if capacity estimates at larger probe set sizes are consistently higher than those at smaller probe set sizes, this would show that increasing the number of comparison options increases performance despite a decrease in chance

level (which is taken into account by the estimation process). Figure 11 compares capacity estimates calculated on specific subsets of the trials. Sample set size was allowed to vary (denoted as N in the figure), and probe set size was group as all trials (labeled All Trials) or fixed at either 2 (e.g., 4:2; labeled N:2 Trials) or matched to the sample set size (e.g., 4:4; labeled N:N Trials). The solid line depicts an equivalent capacity estimate for each trial subset. Overall, capacity estimates ranged from 2 to 5 regardless of the subset considered, with some variation in the means: $M_{N:2} = 2.74$, $M_{N:N} = 3.65$, $M_{All} = 3.22$. Notably, the trial type with the predicted highest comparison difficulty actually has the highest capacity estimate, strengthening the conclusion that configural processing may be playing a role in the change detection task, even at lower sample set sizes. To confirm this difference, a one-way RM ANOVA of Trial Subset (N:2 vs. N:N) on capacity revealed a significant main effect [$F(1, 22) = 11.152$, $p = .003$, $\eta_p^2 = .336$]. An analogous non-parametric test bore the same pattern of results. The all trials condition was not explicitly tested as it includes N:2 and N:N trials[2].

The first panel compares capacity estimates from trials with the lowest comparison difficulty (N:2) to those with the highest (N:N). For most individuals, the capacity estimate for N:N trials was higher than the estimate from N:2 trials, in agreement with the above test on the means. The lack of a significant correlation between the estimates [$r = .268$ $p > .05$] gives evidence for a dissociation between performance at low and high probe set sizes. Although null results should not be interpreted too strongly, the lack of correlation is consistent with the view that capacity and control of VWM involve independent processes. The existence of configural

[2] Including the capacity estimates from all the trials as an additional level in the RM ANOVA did not change the pattern of results, nor was a reliable difference found between the N:2 or N:N capacity estimates and the estimates calculated across all trials.

benefits in this task, however, suggests that perhaps capacity is independent of strategy (whether one benefits from configural information) at larger set sizes.

The right and lower panels compare the subsets with the overall capacity estimates. Interestingly, there is no reliable correlation between overall capacity estimates and the N:2 trials [$r = .161$, $p = .463$], although the linear relationship between N:N estimate and overall estimates is reliable [$r = .736$, $p < .001$]. This pattern shows that the overall capacity estimates are more consistent with the N:N estimate rather than the N:2 estimate, perhaps because the additional probe set sizes included in the overall estimates are more similar to N:N trials than N:2 trials.

As a whole, results from Experiment 1 shows that if we want to distinguish executive control of VWM from capacity of VWM, we need to ensure there are no alternative strategies that work at specific sample/probe set size pairs (e.g., an increasing reliance on configural information as sample set size increases). Importantly, the fixed capacity model fails to provide an accurate picture of performance in this task, as demonstrated by the increase in capacity from N:2 to N:N trials. Before abandoning this assumption, however, Experiment 2 will try and reduce the apparent configural benefit in Experiment 1 by altering the correspondence between the identities of items in the sample and probe array. If the configural benefit is caused by repetition of item identities (rather than item locations) than the effect should be eliminated. If the configural benefit was just masking an increase in comparison difficulty, though, the effect may be reversed (N:N trials harder than N:2).

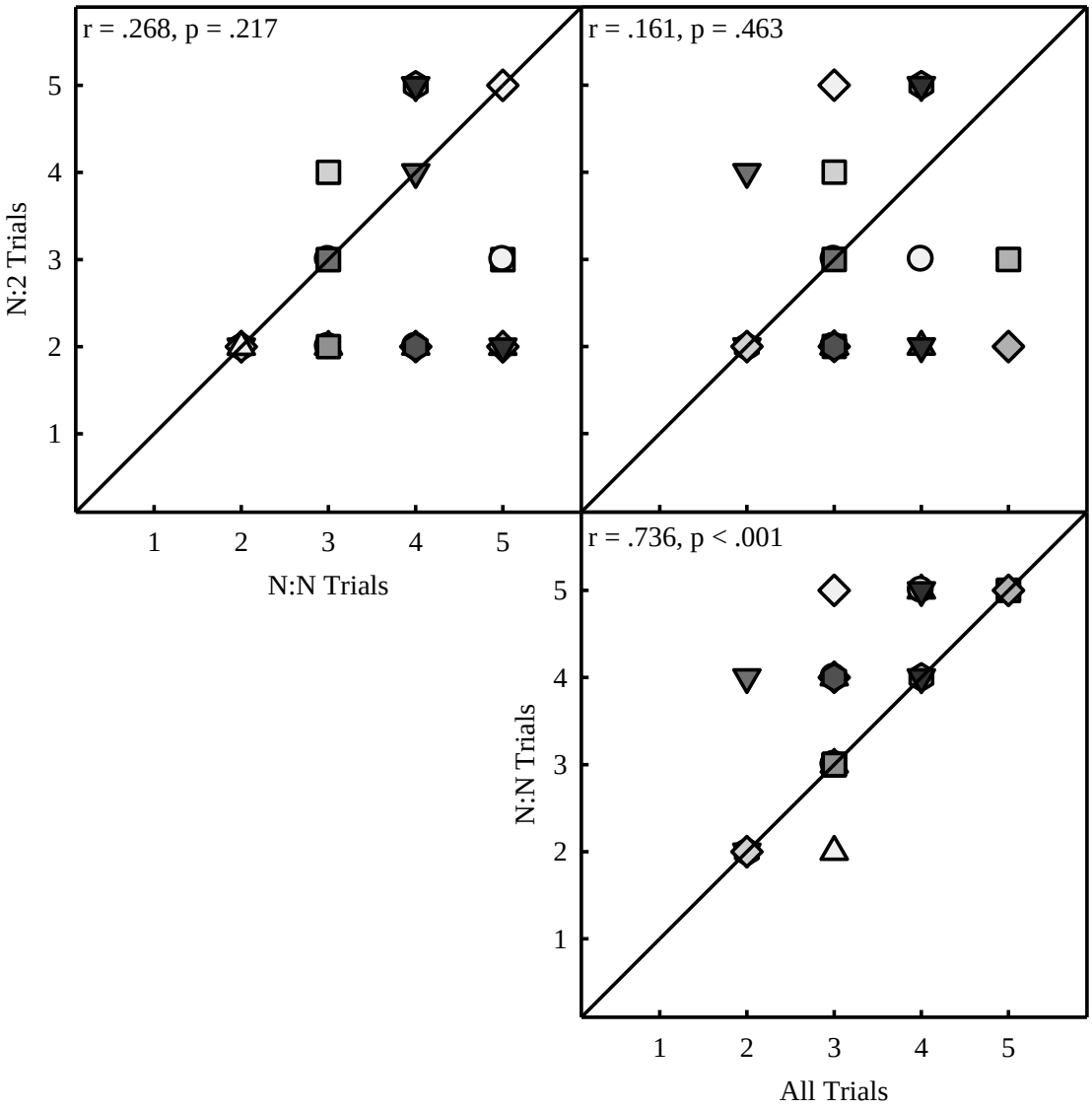


Figure 11. Relationship between capacity estimates calculated on a specific subset of sample/probe set size pairings, or for all trials for Experiment 1. N:2 indicates trials with all sample set sizes, but only probe set size 2 (i.e., 4:2, 5:2, and 6:2), and N:N indicates trials with the same number of items in the sample and probe array (i.e., 4:4, 5:5, and 6:6).

Experiment 2

As underscored by the somewhat involved capacity estimation process, a complication in understanding how probe set size impacts change detection is that unchanged probe array items may also be stored in memory. If the items in the probe array are already stored in memory (i.e., they have a 1-to-1 correspondence with items in memory), it likely impacts how the array is processed, similar to priming effects in visual search (Maljkovic & Nakayama, 2000). Consider a trial with 4 items in the sample array and 4 items in the probe array, with one item having changed in color. Assuming the participants have correctly encoded and stored the items from the sample array, the 3 unchanged items in the probe array correspond to items in memory. In contrast, a trial with 6 items in the sample and probe arrays likely contains several items that appear "new" in the probe array, as they are not able to accurately store all 6 items from the sample array. Thus, at larger probe set sizes there are not only more comparison to make, but the proportion of changed/unchanged comparisons is also different. Before describing Experiment 2 (which ensures only 1 item is unchanged from the sample array) some related results are considered.

Results from an earlier study using the flicker procedure demonstrate the impact of presenting new items in the probe array (Rensink, 1999). In the flicker procedure, the sample and probe array continuously alternate (with a filled interval in-between) until the participant reports that they have detected the change. Instead of changing a single object, however, Rensink changed all but one object and measured the time it took participants to locate the unchanged item. Rather than a VWM capacity around 4 items, the search slopes indicated a capacity of about 1. This striking result has lead researchers to suppose that VWM capacity differs for changed v. unchanged objects (Rensink, 1999, 2001; Theeuwes, 2004).

To test this hypothesis, Theeuwes and colleagues (2004; Pinto, Olivers, & Theeuwes, 2006) used the flicker procedure to compare detection times for static targets with changing distractors (no-change-target condition) and changing` targets with static distractors (change-target condition). Except at the lowest delay intervals (likely involving iconic memory or a motion filter, cf. Davis & Leow, 2005) targets were detected more efficiently in the change-target condition than in the no-change-target conditions. Together, these results from the flicker procedure suggest that the presence of previously non-stored information in the probe array severely degrades performance.

Hyun, et al. (2009) assessed comparison difficulty by manipulating whether participants looked for "any difference" or "any sameness" across the sample and probe array using a whole-display report. In the any-difference condition, probe arrays contained 0 or 1 changed items, but in the any-sameness condition, probe arrays contained 0 or 1 unchanged items (the rest of the items changed). Across sample/probe set sizes of 1 to 4, accuracy was nearly perfect except for the any-sameness condition with 1 unchanged item, which decreased about 13% for each additional item. Response time data showed that both any-sameness conditions produced steeper set size functions than the any-difference conditions, though set size functions were not flat in any condition. In additional experiments, measurement of N2pc latency (a negative waveform appearing over the hemisphere contralateral to the shift of attention, shown to underlie covert shifts of attention, cf., Luck & Hillyard, 1994) and saccade onset suggested that changes are detected by an unlimited capacity mechanism. Manual response time data, however, suggested that the processes controlling the behavioral response are limited in capacity (i.e., increases in response time with increases in sample/probe set size). Therefore, unlike visual search, in change detection the *target* is not immediately brought into attentional focus when detected. Instead, a

separate process must bring the target into focus so it can be acted upon (Hyun, et al., 2009). Combining these results with the previously mentioned work with the flicker paradigm, the any-sameness procedure should be a much better assessment of executive control and capacity than the normative change detection procedure.

The primary goal of Experiment 2 was to eliminate item-identity repetition as a potential source of the increase in capacity as probe set size increased. To this end, an analogue of the any sameness procedure (Hyun et al., 2009) was used. The procedure was designed to be the conceptual opposite of the forced-choice "pick the change" condition; participants search for the only unchanged object and "pick the same" amidst a field of changed objects. Distinct from Hyun et al. (2009), the current design uses sample set sizes above normative capacity estimates to ensure that capacity estimates are not heavily constrained by ceiling effects (cf. Rouder, et al., 2011 and the constraints on the capacity formulas presented above), as well as manipulating probe set size. By comparing changes in VWM capacity (rather than accuracy), we can more directly measure the dependence of VWM capacity and control.

Method

Participants

Participants (N=28) were students enrolled in a psychology class at Auburn University, ages ranging from 18 to 24. None of the participants completed Experiment 1. Participants were consented and compensated as in Experiment 1. The Auburn University Institutional Review Board approved all protocol details.

Apparatus & Stimuli

Participants were tested under the same conditions and using the same materials and stimuli as in Experiment 1.

Procedure

The procedure is identical to Experiment 1, except that probe arrays will contain a single *unchanged* item amidst several *changed* items (rather than a single changed item amidst unchanged items; see Figure 6). Participants will read on-screen instructions similar to Experiment 1, except they are instructed to look for the square that is the same, rather than changed:

> On the next screen you will see multiple colored squares for a short amount of time. After they disappear several squares will reappear. Your task is to touch the square that is the SAME from the first display. Accuracy is more important than speed in this task. You will not receive any feedback about your performance, simply try your best.
>
> Touch anywhere on the screen to begin.

To ensure that the only difference between Experiments 1 and 2 is the added difficulty of comparing novel items, the sample arrays and trial order for Experiment 2 were identical those used in Experiment 1. As in Experiment 1, participants completed 2 blocks of 240 trials, with 20 repetitions of each of the 12 trial types (defined as in Experiment 1), block order counter balanced across participants.

Experiment 2 results

Figure 12 compares the accuracy data from the pick-the-change (Experiment 1) and pick-the-same (Experiment 2) procedures. At the lowest sample set size there appears to be no effect

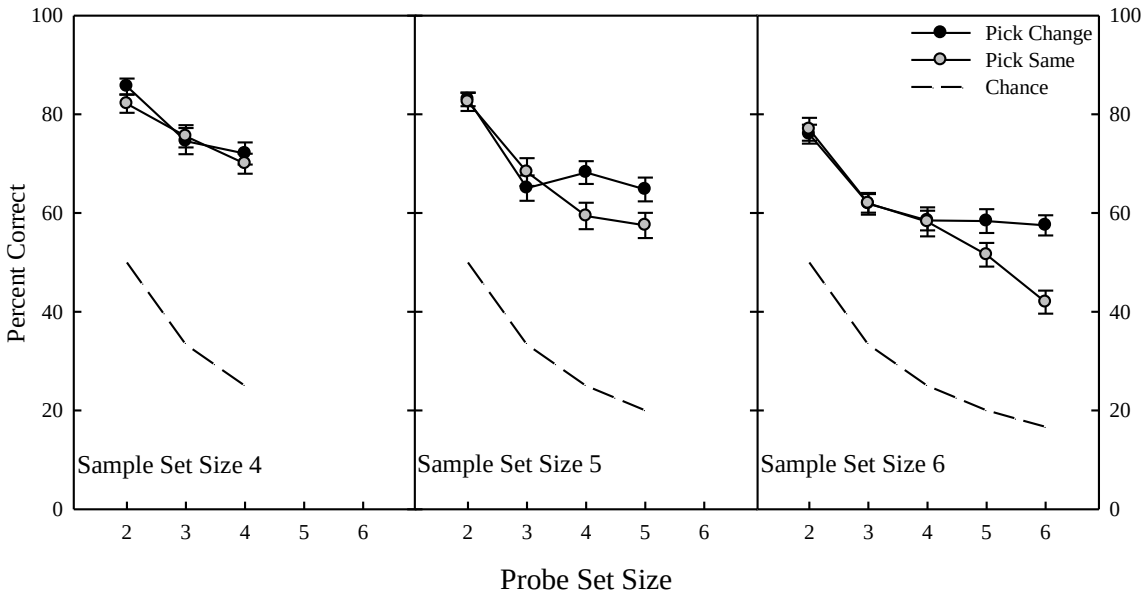


Figure 12. Changes in accuracy across sample set size (separate panels) and probe set size for Experiments 1 (Pick Change) and 2 (Pick Same). The dashed line represents chance performance, calculated as 1 / probe set size. Error bars are ±SE.

of procedure, but the high sample set sizes show a clear interaction between sample/probe set size and procedure. Specifically, when the sample and probe set size were large, participants in the pick-the-same experiment had lower accuracy. To assess the reliability of these effects, a 2x3x3 mixed ANOVA of Experiment (1 v. 2), Sample Set Size (4, 5, or 6), and Probe Set Size (2, 3, or 4) on accuracy was conducted. As before, only fully crossed variables may be entered into the ANOVA, so higher sample/probe set size combinations are not entered into this ANOVA. The ANOVA yielded significant main effects of sample set size [$F(2, 98) = 152.128$, $p < .001$, $\eta^2_p = .756$] and probe set size [$F(2, 98) = 182.412$, $p < .001$, $\eta^2_p = .788$], two-way interactions between sample and probe set size [$F(4, 196) = 4.255$, $p = .003$, $\eta^2_p = .080$] and experiment and probe set size [$F(2, 98) = 4.032$, $p = .021$, $\eta^2_p = .076$], as well as the three-way interaction of sample set size, probe set size and experiment [$F(4, 196) = 2.587$, $p = .038$, $\eta^2_p = .050$]. The main effect of experiment [$F(1, 49) = 1.214$, $p = .276$] and the interaction of

experiment and sample [$F(2,98) = .439$, $p = .646$] were not reliable. Although the measured effect sizes for each of the significant interactions are small (the largest being .08), they confirm the pattern in Figure 12. The data not entered into the ANOVA (5:5, 6:5, and 6:6 trials) visually strengthen the case for the observed interactions.

As with Experiment 1, polynomial trend contrasts were conducted to assess how increasing probe set size impacts accuracy within each sample set size. For sample set size 4, the analysis showed a strong linear effect [$F(1, 27) = 55.568$, $p < .001$, $\eta^2_p = .673$], but no quadratic effect [$F(1, 27) = .010$, $p = .923$]. This mirrors the effects for the same sample set size in Experiment 1, showing that at the smallest sample set sizes there is little difference between the tasks. Moving to sample set size 5, there was evidence of linear [$F(1, 27) = 200.387$, $p < .001$, $\eta^2_p = .881$] and quadratic [$F(1, 27) = 16.166$, $p < .001$, $\eta^2_p = .375$] trends, but no cubic trend [$F(1, 27) = .003$, $p = .959$]. The lack of a cubic trend is a departure from Experiment 1, and suggests that increasing the number of changing items has a larger impact on performance than increasing the number of unchanging items. At sample set size 6, there was strong evidence for linear [$F(1, 27) = 352.534$, $p < .001$, $\eta^2_p = .929$] and cubic trends [$F(1, 27) = 16.258$, $p < .001$, $\eta^2_p = .376$], but not quadratic or quartic trends [both Fs$(1, 27) < .17$, ps $> .2$]. The presence of a cubic trend without a quadratic trend shows that accuracy initially decreased, but then leveled off before again decreasing at the largest probe set sizes. The differences between Experiment 1 and 2 at the larger sample set sizes suggest that removal of the identity correspondence between the sample and probe array has eliminated the benefit for large probe arrays.

Returning to Figure 12, note that as probe set size increases, accuracy continues to decrease in Experiment 2, but levels off for Experiment 1. This pattern suggests that location repetition without identity repetition may not provide a strong configural benefit (cf. Figure 10,

right panel). Indeed, a 2x3 mixed ANOVA of Experiment (1 or 2) and Sample Set Size (4, 5, 6) on configural benefit showed only a main effect of experiment [$F(1, 49) = 6.303$, $p = .015$, $\eta^2_p = .114$], but no reliable effect of sample set size [$F(2, 98) = 0.694$, $p = .502$] or the experiment by sample set size interaction [$F(2, 98) = 2.670$, $p = .074$]. Because the sample set sizes have differing amounts of configural information, any effect could be masked by inconsistencies across sample set sizes. To test this possibility, follow-up ANOVAs (identical to independent samples *t*-tests) were conducted within the sample set sizes between the experiments. Participants in Experiment 2 showed a stronger negative effect ($M = -11.04$) than those in Experiment 1 ($M = -0.87$) at the largest sample set size [$F(1, 49) = 8.05$, $p = .007$, $\eta^2_p = .141$], but not at the lower sample set sizes [both Fs$(1, 49) < 1.6$, ps $> .21$]. The difference at sample set size 6 remained significant even after removing the participant with an extreme value (-50) for configural benefit ($p = .010$) or adopting a more stringent false positive criterion to account for the post-hoc follow up tests (i.e., a Bonferroni-corrected α-level of .0167 for each of the 3 comparisons). This result suggests that the benefit for repeating items from the sample array in the probe array is at least partially related to the identity of the items, not just their locations. Because the locations for items in the sample and probe array were equated for Experiments 1 and 2, the only explanation for the decrease in accuracy is that the probe array contains a single same item rather than a single changed item. Experiment 3 will also manipulate the location of items in the probe array to test if location repetition is providing some benefit.

As mentioned earlier, assessing accuracy may be misleading as it does not take into account chance levels. Figure 13 shows how probe set size impacts capacity estimates for Experiment 2 (cf. Figure 11). Notice now that the capacity estimates for N:2 trials are almost always greater than or equal to that for the N:N trials, the reverse trend from the pick-the-change

task in Experiment 1. For Experiment 1, one reason for the increased capacity estimate for N:N trials was the use of configural information to increase accuracy (and consequently the estimate capacity). In the current experiment, a reduction in the configural information available parsimoniously explains the reduction in capacity, but cannot explain why capacity dips below that on N:2 trials. Because the current capacity estimation formula takes into account the chance rate, the lower capacity estimates for N:N trials (M=2.46) than N:2 trials (M=3.29) must be due to something beyond changes in the chance rate. The difference is reliable [$t(27) = 3.191$, $p = 0.004$], and it seems reasonable to conclude that an increase in comparison difficulty is behind the significantly degraded performance. If comparison difficulty is the cause of the decrease, then the relationship between capacity on N:2 trials and N:N trials can serve as an indicator of the dependence of VWM control and capacity. The correlation between capacity on N:2 and N:N trials is not reliable ($r = .213$, $p = 0.278$). If the size of VWM capacity is indistinguishable from an ability to control information, this correlation should be strong and positive.

Although the decrease in capacity from N:2 to N:N trials was reliable, the magnitude of the decrease may not have been large enough to find a significant correlation. In the next experiment, the procedure was modified to assess a broader range of both sample and probe set sizes, and more participants were recruited to increase the power of the approach.

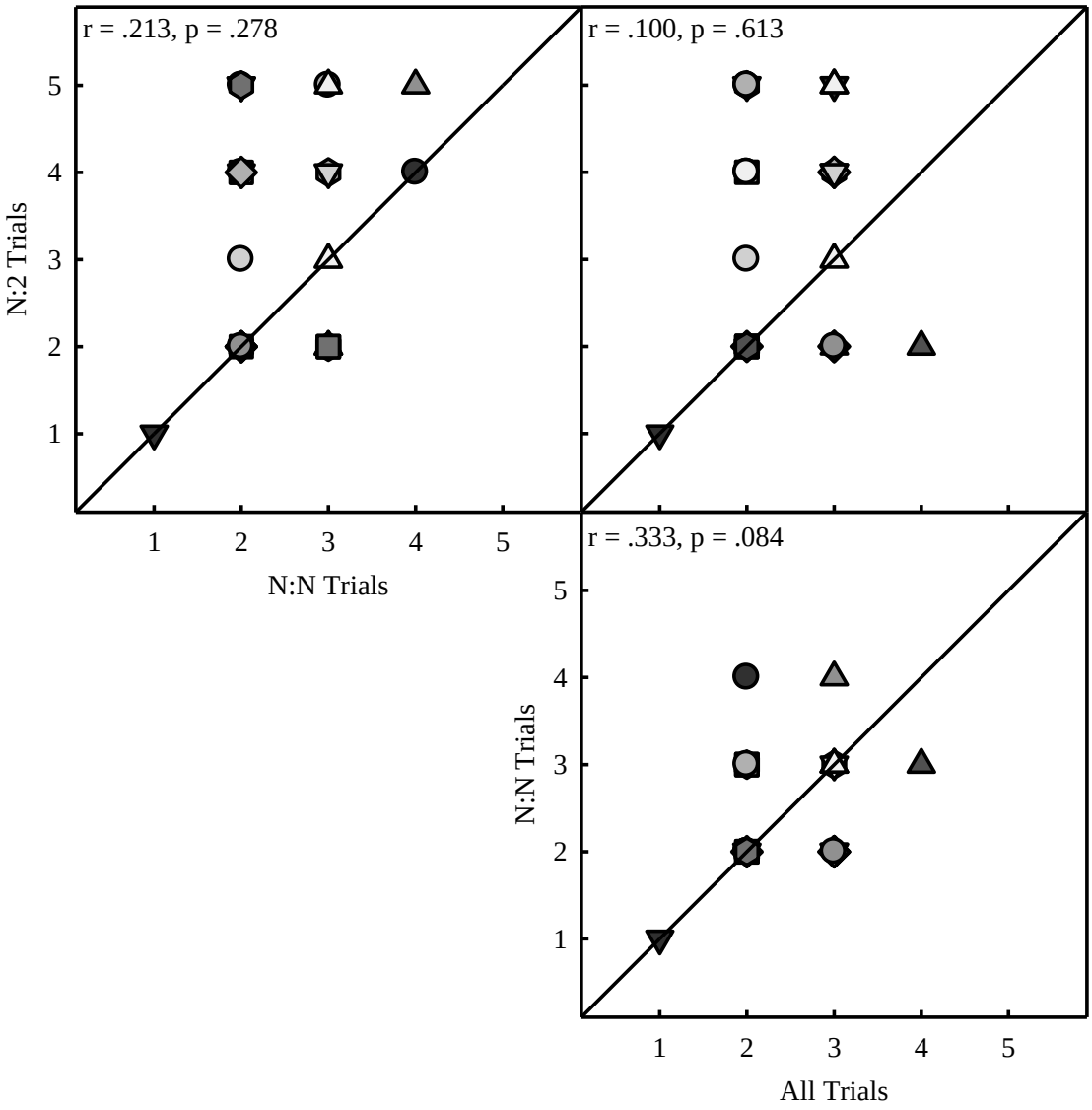


Figure 13. Relationship between capacity estimates calculated on a specific subset of sample/probe set size pairings, or all trials for Experiment 2. N:2 indicates trials with any sample set size and probe set size 2, and N:N indicates trials with the same number of items in the sample and probe array.

Experiment 3

The goal of this final experiment was to see if a relationship between capacity and control would emerge if a wider range of estimates on both variables could be achieved. The procedure from Experiment 2 was modified (cf. Figure 6) to ensure that the estimates of VWM capacity are not contaminated by non-identity information or influenced by location-based strategies. To ensure only identity information can be used to solve the task, location-correspondence between the sample and probe arrays needs to be disrupted. Although none of the previous experiments required participants to encode location information, some research suggests that location information (specific and relative) may be encoded regardless of task instruction (e.g., Jiang et al, 2000). If participants encode location information as part of encoding the colors, changing the probe set size confounds changes in comparison difficulty with changes in location information (cf. Johnson, et al., 2008).

The bottom path in Figure 6 shows the lineup task employed in the current experiment. This design has a clear advantage over the previous procedure for measuring comparison difficulty, as the original version cannot increase probe set size beyond sample set size without causing concomitant changes in location information. Rather than using randomized locations that would allow for (actual or perceived) overlap in the sample and probe locations, the stimuli in the probe array are always presented horizontally at a fixed height across the screen. These changes allow estimates of VWM capacity not confounded by changes in the available location information. This design combines the memory load aspect of change detection with a search/scanning component from traditional visual search tasks used to assess comparison difficulty (e.g., Alvarez & Cavanagh, 2004).

Method

Participants

Experiment 3 involved a new set of participants (N=40), who were students enrolled in a psychology class at Auburn University, ages ranging from 18 to 24. More participants were needed in this experiment because of the exploratory nature of the procedure and to allow for sufficient power to assess capacity and control dependence on an individual basis. Participants were consented and compensated as in the first two experiments. The Auburn University Institutional Review Board approved all protocol details.

Apparatus & Stimuli

Participants were tested under the same conditions and using the same materials and stimuli as in Experiment 2. For Experiment 3, the probe array was displayed horizontally, centered at fixation, with an inter-item distance of 1.6°, each item jittered vertically according to a uniform [-0.5°, 0.5°] distribution (see Figure 6, bottom probe array).

Procedure

As in Experiments 1 and 2, participants will read on-screen instructions describing the task (identical to Experiment 2) as a search for a single object that stays the "same" across the two displays. A wider range of probe set sizes (i.e., 2, 4, 6, and 8) was used to assess performance as comparison difficulty changes. To ensure the increased comparison difficulty does not lead to floor effects, smaller sample set sizes (i.e., 2, 4, and 6) was used to reduce memory load. As an attempt to equate encoding and storage demands with the previous experiments, sample and probe arrays from Experiment 2 were used whenever possible. When not possible (e.g., the current experiment has no sample set size 5 trials), the trials for the current experiment involved either subsets or supersets of the sample/probe arrays from Experiment 2.

Table 2 gives a detailed mapping of the trial types from Experiment 2 to Experiment 3. Participants will again complete 6 practice trials before completing 2 blocks of 240 trials (balanced across the 12 trial types and ordered randomly). All other procedural details are identical to Experiment 2.

Table 2. Mapping of trial types between Experiment 2 and Experiment 3. Bolded trial types involve addition/subtraction of items from the sample and/or probe array. Plain type entries will remain unchanged across experiments.

Exp. 2 Trial Type → Exp. 3 Trial Type		
4:2 → 4:2	**5:2 → 2:2**	6:2 → 6:2
4:3 → 4:6	**5:3 → 2:6**	**6:3 → 4:8**
4:4 → 4:4	**5:4 → 2:4**	6:4 → 6:4
	5:5 → 2:8	**6:5 → 6:8**
		6:6 → 6:6

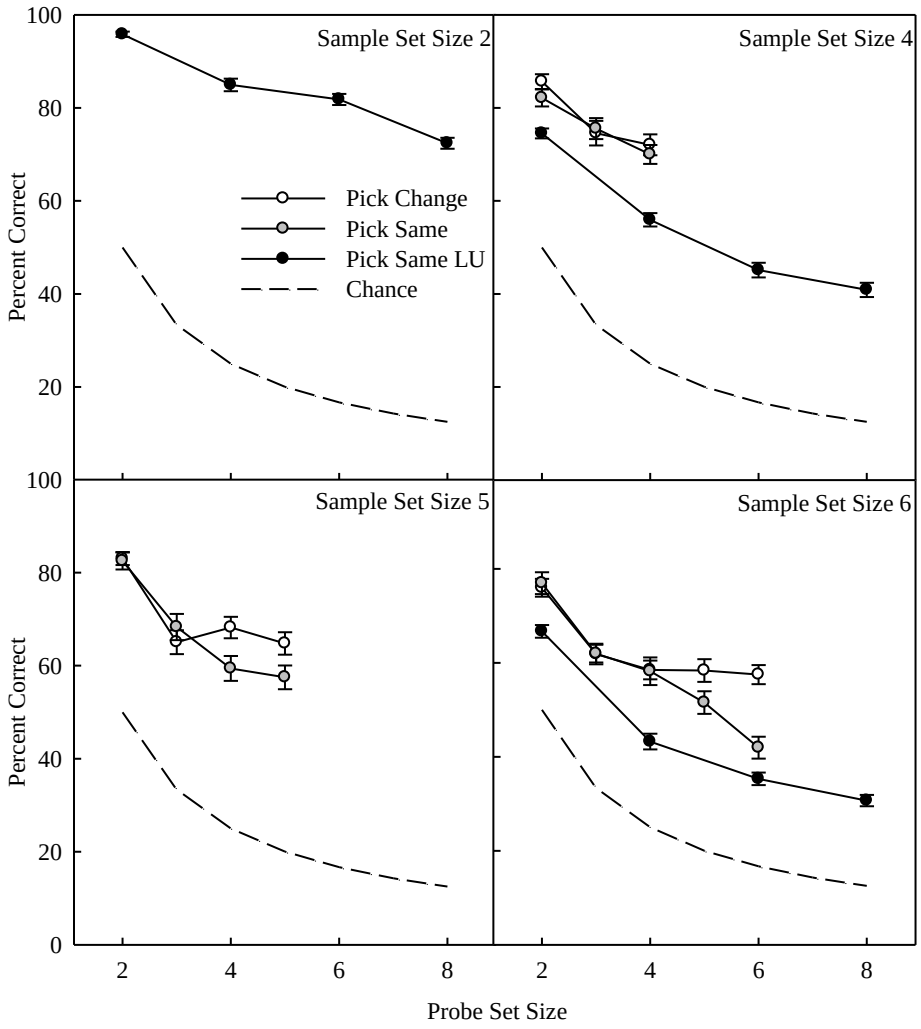


Figure 14. Accuracy across sample set size (separate panels) and probe set size for each experiment (separate lines). Error bars are ±SE.

Experiment 3 Results and Discussion

Figure 14 shows how accuracy changes across both sample set size and probe set size for each experiment. Noticeably, performance for the lineup task (black circles) in Experiment 3 ranges from nearly perfect on 2:2 trials (top-left panel) to around 40% for 6:8 trials. The strong decrease within sample set size 2 shows the impact of comparison difficulty even when all the sample elements should fit into VWM. For sample set sizes 4 and 6, the curve for Experiment 3 is always below that for the other experiments. Because the trial types are not directly comparable, a large mixed ANOVA was not possible. Instead, several one-way ANOVAs were conducted across Experiment (1, 2, or 3) on Accuracy for the overlapping sample/probe set size pairs (4:2, 4:4, 6:2, 6:4, and 6:6). Even with a stricter significance criterion (Bonferroni-corrected alpha level for 5 tests is .008), all the ANOVAs showed reliable differences among the experiments [all Fs(2 ,88) >12, ps < .001]. Tukey follow-up tests showed that accuracy for Experiment 3 was significantly lower than for Experiments 1 or 2 for conditions 4:2, 4:4, 6:2, and 6:4 (all ps < .001). For 6:6 trials, the difference between Experiments 2 and 3 dissipated, but now Experiment 1 showed a reliable increase compared to 2 and 3 (p < .001). The difference between Experiment 2 and 3 was no longer reliable on 6:6 trials (p = .064). These results suggest there is a decrement due solely to breaking the location-correspondence between the sample and probe arrays. This finding is somewhat surprising as VWM is theorized to be robust to changes in location (cf. Table 1). Once the sample and probe set size become large, however, the increased difficulty due to the Pick the Same procedure swamps any difference due to the loss of location correspondence and produces equivalent accuracy in Experiments 2 and 3.

To quantify the degree of comparison difficulty, capacity was assessed across sample set size within each probe set size. Because there is no longer any location- or identity-

correspondence between the sample and probe arrays, participants cannot use information about unchanged objects to solve the task (e.g., an elimination strategy). Figure 15 shows how capacity estimates changed across probe set size for the pick the same lineup task. Mean capacity estimates decrease smoothly across probe set size, showing that increasing the number of choice options negatively impacts performance. A RM ANOVA on the capacity estimates across probe set size showed a reliable effect of probe set size [$F(3, 117) = 21.832$, $p < .001$, $\eta^2_p = .359$]. Trend contrasts suggest a linear effect of probe set size [$F(1, 39) = 51.882$, $p < .001$, $\eta^2_p = .571$], with unclear evidence for a quadratic component [$F(1, 39) = 3.594$, $p = .065$, $\eta^2_p = .084$], and no evidence for a cubic trend [$F(1, 39) = .043$, $p = .837$]. If the general pattern in Figure 15 survives extrapolation beyond 8-item probe arrays, the quadratic component should emerge.

With clear evidence for the role of comparison difficulty in this task, the original question about the relationship between VWM capacity and control can be addressed. To avoid ceiling effects capacity estimates were obtained for sample set sizes 4 and 6 only. Additionally, to reduce the impact of comparison difficulty in our capacity estimate, only probe set size 2 will be used. VWM control was quantified by measuring the change in accuracy across increasing probe set sizes for sub-capacity sample set sizes (e.g., sample set size 2). Similar to the search rates discussed previously, the slope of the regression line relating probe set size and accuracy was used as a dependent measure of control of VWM. If participants are able to control which information is entered into VWM (or equivalently, decide which information stays), there should be little to no impact of increasing the size of the probe array. This measure of control is distinct, however, from search rates obtained from target search designs because they contain only a single target item (e.g., Alvarez & Cavanagh, 2004). Informal testing in our lab showed that with a single sample array item, accuracy was nearly perfect and search rate data also failed to

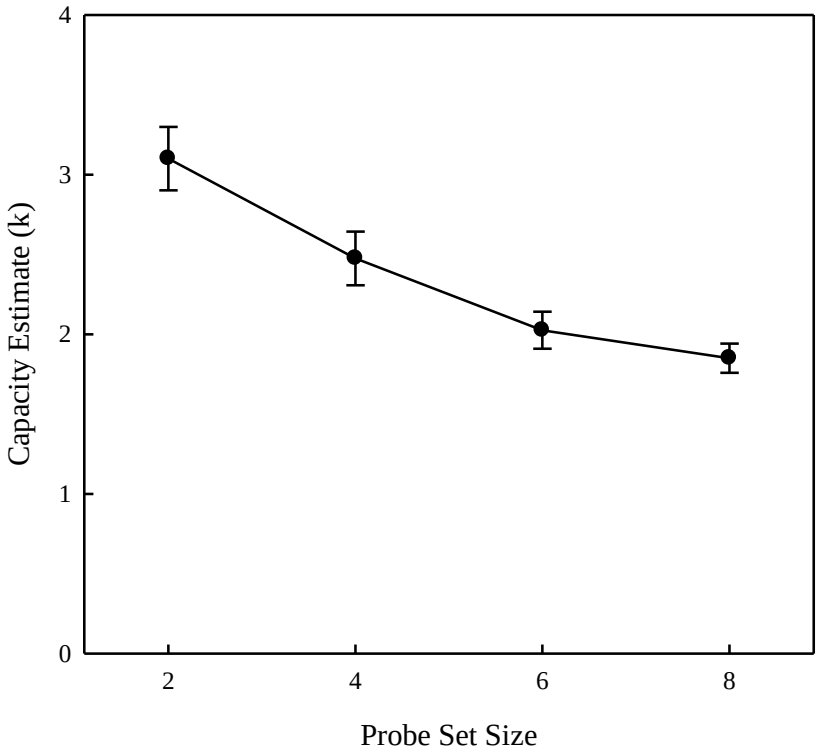


Figure 15. Mean estimated capacity across probe set size. Capacity is calculated for each participant across sample set size for each probe set size. Error bars are ±SE.

produce the large inter-subject differences needed to assess the relationship between capacity and control processes of interest. When participants are searching for a single item in memory, the processes and strategies involved may be qualitatively different than those invoked under a small memory load.

Figure 16 shows the resulting scatter plot for participants in Experiment 3. A simple linear regression model showed that control had no linear relationship with capacity, ($\beta = 0.205$, $t(38) = 1.294$, $p = .204$). This result suggests, *ex facie*, that VWM capacity is distinct from resistance to interference within individuals, as each process accounts for an unreliable 4% of the variance in the other. A possible objection to this conclusion is that the restricted range of each variable obscured any true relationship. Figure 17 looks at both the conditional distribution of

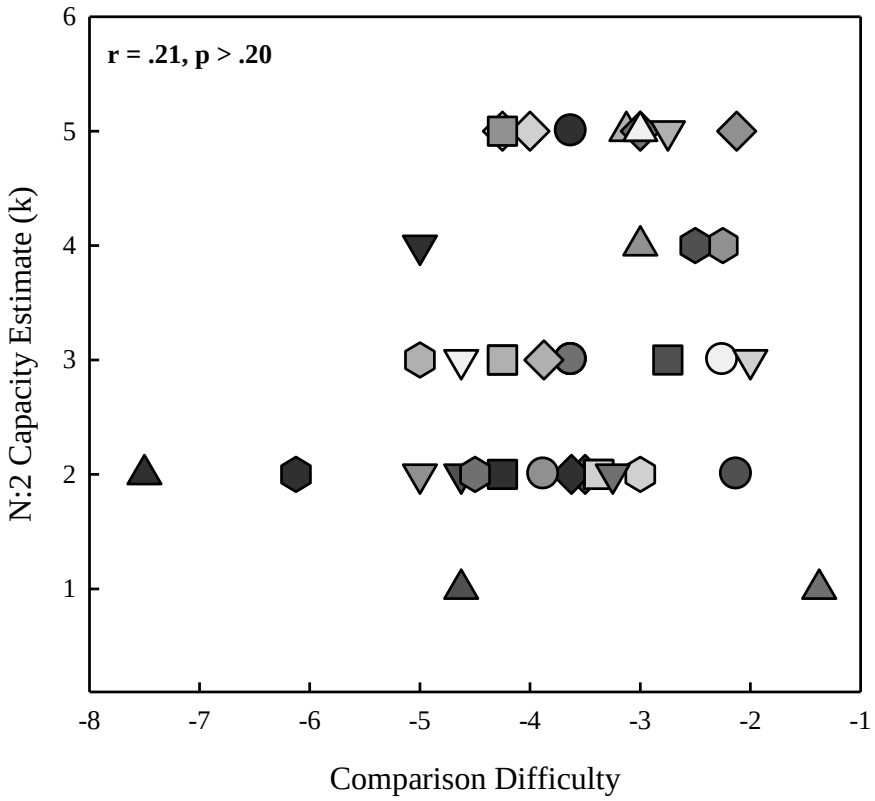


Figure 16. Relationship between VWM capacity and control for Experiment 3. The vertical axis is the capacity estimate calculated across sample set size, but restricted to probe set size 2. The horizontal axis is the slope of the regression line fit to accuracy across probe set size, but restricted to sample set size 2.

comparison difficulty given capacity as well as the marginal distribution of comparison difficulty. The top panel shows that the around 80% of participants had estimated capacities of 2, 3, or 5. The distributions within each of these capacities, however, are reasonably similar—if participants with larger capacities were able to control access to their VWM better, their distribution of comparison difficulties should be shifted toward 0. Instead, the median comparison difficulty remains between -3 and -4 for all of these capacities. The bottom panel in Figure 17 shows the marginal distribution of comparison difficulty. Most of the values are between -4 and -2, which may be too small of a range to observe a slight dependence. If the two processes were entirely dependent, however, there should still be an observed relationship, despite the somewhat sparse sampling across the two variables.

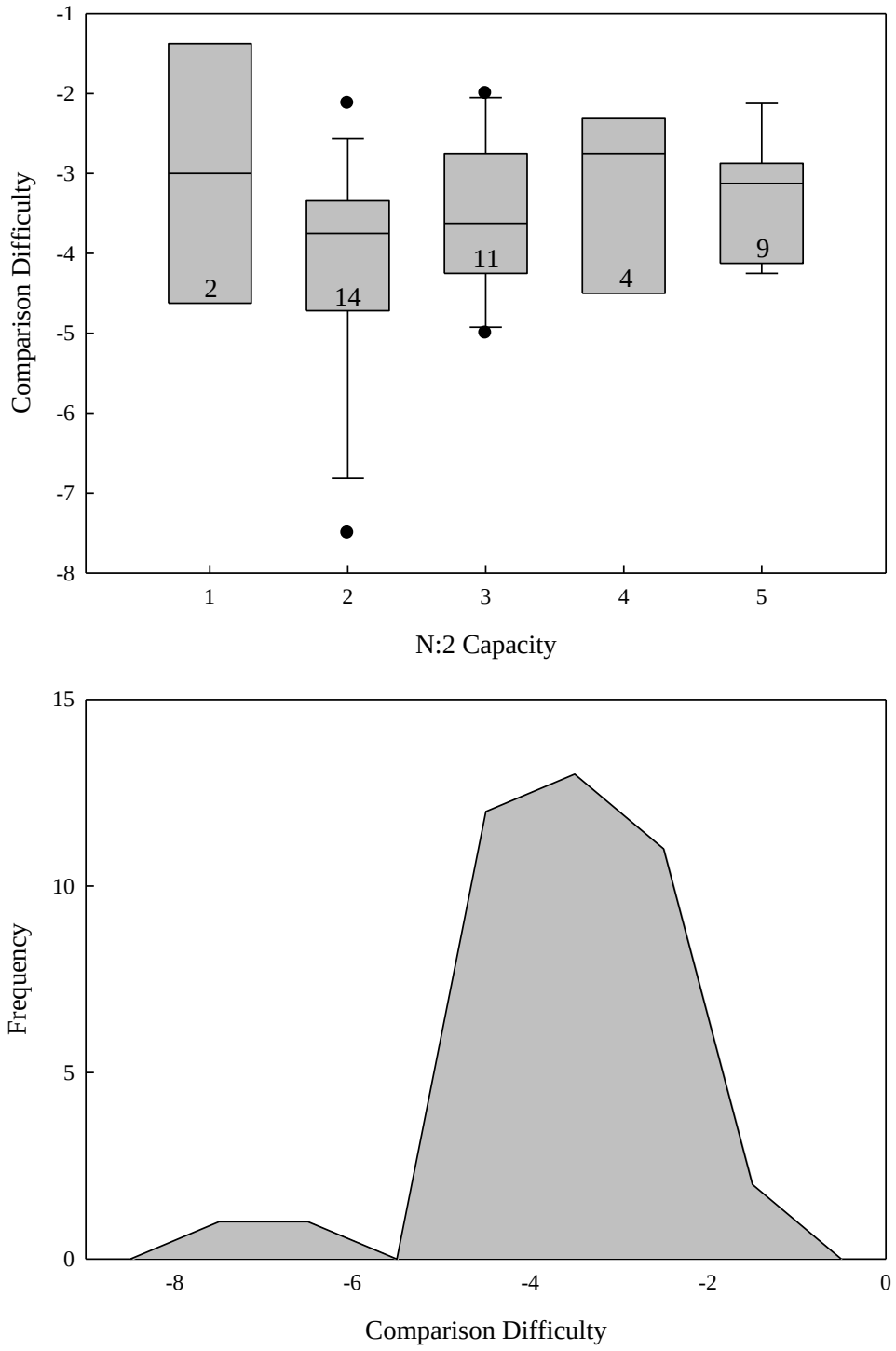


Figure 17. *Top*. Boxplot showing the distribution of comparison difficulties within each capacity. The numbers within each boxplot indicate the number of observations summarized by each plot. Filled points represents values beyond the 1st or 3rd quartile by 1.5 times the inter-quartile range. *Bottom*. A frequency polygon showing the marginal distribution of comparison difficulty.

General Discussion

Across three experiments, the effect of varying probe set sizes on VWM capacity was assessed. The broad goal of these experiments was to develop a test to independently measure VWM capacity and executive control and assess their dependence upon one another. The change detection task was used because previous work had suggested that it is amenable to capacity estimation and that comparison difficulty may play a significant role in the task (e.g., Awh et al., 2007; Luck & Vogel, 1997). The primary conclusion from the set of experiments is that comparison difficulty may significantly impact capacity estimates even for simple color stimuli and the degree of its impact is unrelated to an independent measure of VWM capacity.

In Experiment 1, sample and probe set size were manipulated to look for changes in VWM capacity estimates. Starting with the fixed-capacity assumption, a capacity-estimation technique was derived for the forced-choice procedure that corrects an error in an earlier derivation. The technique was generalized to include varied probe set sizes and an attention parameter. Results showed that capacity estimates derived from the forced-choice change detection procedure are dependent on the probe set size used, reflecting a configural benefit when the sample and probe set sizes match. No relationship was found between capacity estimates using the probe set sizes (i.e., N:2 trials) and those from higher probe set sizes (i.e., N:N trials).

In the next experiment, the change detection procedure was replaced with a conceptually opposite task, in which participants picked the only unchanged object from an array of changed objects. Again, sample and probe set size were manipulated and capacity estimates compared across changing probe set size. As in Experiment 1, results demonstrated an effect of probe set size on estimated capacity, but in the opposite direction: capacity estimates significantly

decreased from N:2 to N:N trials. These results show a memory cost for comparing even simple colors between memory and an on-screen array. Participants were not able to perform the comparison operation without using some VWM storage.

Results from Experiment 3 demonstrated how visual object comparison processes operate without the aid of spatial or identity repetition. Because encoding and storage demands were roughly equated across the last two experiments (and in the case of sample set size 2, were made less demanding), the differences between Experiments 2 and 3 (at equivalent sample/probe set sizes) are related primarily to the removal of spatial cues from the task. The design of the task allowed for estimation of VWM capacity and the impact of comparison difficulty (a proxy variable for VWM control) on independent subsets of the trials. No relationship was found between an individual's measure of VWM capacity and the degree of comparison difficulty. If storage and comparison processes are dependent, high accuracy with low probe set sizes (low comparison difficulty) should predict high accuracy with high probe set sizes (high comparison difficulty) and similarly for low accuracy in either condition. The main thrust of this finding is that having more slots in VWM does not imply that those slots are any more robust to interfering stimuli. The question remains what participants

Zhang and Luck (2008) suggest that when observers are given arrays with fewer items than their VWM capacity, they can increase the precision of their representation via a slot-averaging process. If this averaging is occurring, individuals with higher VWM capacities will have an easier time picking out the target (the one unchanged color) despite the presence of similar distracters. If comparison difficulty is primarily caused by inter-item similarity (e.g., Awh, et al., 2007), then slot-averaging should lead to higher accuracy for large probe arrays. The lack of a strong positive relationship between VWM capacity and comparison difficulty argues

against the slot-averaging mechanism. Alternatively, if the comparison process is distinct from VWM capacity, there need be no relationship between the measure of comparison difficulty and VWM capacity.

Implications for object capacity estimates

Results from these experiments contribute to three specific questions for characterizing VWM. Primarily, these results show that individuals can have good control (near zero comparison difficulty) over the contents of VWM without necessarily having a large storage capacity. The development of a single task sensitive to the differences in these processes will continue to shed light on basic questions about VWM, as well as suggest ways to refine how VWM deficits are characterized in clinical populations (e.g., is it primarily a storage deficit or a control deficit). For example, the current task may provide strong evidence that any memory deficits associated with PTSD (e.g., Vasterling et al., 1998) are confined to differences in interference, rather than VWM capacity *per se*. Using other change detection tasks or estimation techniques cannot discern between these two processes.

Secondly, the capacity estimation techniques developed for these tasks will allow for comparison within and across species. The adjustment to the current method for estimating capacity in the forced-choice method (e.g., Eng, et al., 2005) adds credibility to using this task, especially as it may be a simpler alternative to the yes/no paradigm (e.g., for nonhuman use, e.g., Elmore et al., 2011). As the primary way to test theories is to compare the predictions they make, it is imperative to have well-principled derivations of each model's predictions (cf. Rouder et al., 2011).

Data from this experiment also have direct bearing on the applicability of the strong-object hypothesis to situations without location correspondence, which is more likely to be the

case in real-world settings. For example, when trying to pick out a friend in a crowded room, it is unlikely that past spatial relationships play a large role in helping identify your friend. The current results show that breaking location-correspondence between a sample and test array can lead to decreased memory for the sample items. Further investigation along these lines will show the extent to which an object's identity information is separable from where it is located (in both a relative and absolute sense) in space.

Conclusions

A running theme throughout each experiment is that procedural and perceptual changes at choice time can have large impacts on capacity estimates. Rather than suggest that VWM is a highly robust representation, the current data add to the growing body of literature suggesting the fragility of VWM (e.g., Makovski, & Jiang, 2008; Makovski, Shim, & Jiang, 2006). Unique to the current work is the focus of interference at choice time rather than at encoding or during maintenance. Taken together, these results emphasize the importance of quantifying the joint contribution of each VWM process in order to better understand them individually. The procedure and estimation technique presented provide the way to measure these two separate aspects of VWM, providing an avenue toward a fuller understanding of the functions of human and nonhuman visual memory.

References

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.

Atkinson, J., King, J., Braddick, O., Nokes, L., Anker, S., & Braddick, F. (1997). A specific deficit of dorsal stream function in Williams' syndrome. *Neuroreport*, *8*(8), 1919-22.

Avons, S. E., Ward, G., & Russo, R. (2001). The dangers of taking capacity limits too literally. *Behavioral and Brain Sciences*, *24*(01), 114-115.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological science*, *18*(7), 622-8.

Baddeley, A. D. (1992). Working memory. *Science*, *255*(5044), 556-559.

Baddeley, A. D. (1998). Recent developments in working memory. *Current Opinion in Neurobiology*, *8*(2), 234-8.

Bays, P., & Catalao, R. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*, 1-11.

Bays, P. M., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, *11*, 1-15.

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622-31.

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

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.

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.

Cowan, N., Elliott, E. M., Saults, J. S., 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.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51-57.

Davis, G., & Leow, M. C. (2005). Blindness for Unchanging Absence of Motion Targets in the Filtering. *Psychological Science*, *16*(1), 80-82.

Elmore, L. C., Ma, W. J., Magnotti, J. F., Leising, K. J., Passaro, A. D., Katz, J. S., & Wright, A. A. (2011). Visual short-term memory compared in rhesus monkeys and humans. *Current Biology*, *21*(11), 975-979.

Eng, H. Y., Chen, D., & Jiang, Y. V. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic bulletin & review*, *12*(6), 1127-33.

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

Fougnie, Daryl, Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12), 1-11.

Fukuda, K., Awh, E., & Vogel, E. K. (2010). Discrete capacity limits in visual working memory. *Current Opinion in Neurobiology*, *20*(2), 177-82.

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), 673-9. doi:10.3758/17.5.673

Gao, Z., Yin, J., Xu, H., Shui, R., & Shen, M. (2011). Tracking object number or information load in visual working memory: Revisiting the cognitive implication of contralateral delay activity. *Biological Psychology*, 1-7.

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), 570-7.

Gibson, B., Wasserman, E., & Luck, S. J. (2011). Qualitative similarities in the visual short-term memory of pigeons and people. *Psychonomic bulletin & review*.

Harrison, A., Jolicoeur, P., & Marois, R. (2010). "What" and "where" in the intraparietal sulcus: an FMRI study of object identity and location in visual short-term memory. *Cerebral Cortex*, *20*(10), 2478-85.

Heyselaar, E., Johnston, K., & Paré, M. (2011). A change detection approach to study visual working memory of the macaque monkey. *Journal of Vision*, *11*(3), 1-10.

Hollingworth, A. (2003). Failures of retrieval and comparison constrain change detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 388-403. doi:10.1037/0096-1523.29.2.388

Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual Cognition*, *14*(4-8), 781-807.

Huang, L. (2010). Characterizing the nature of visual conscious access : The distinction between features and locations. *Journal of Vision*, *10*, 1-17. doi:10.1167/10.10.24.Introduction

Huang, L., & Pashler, H. (2007). A Boolean map theory of visual attention. Psychological review, 114(3), 599-631. doi:10.1037/0033-295X.114.3.599

Huang, L., Treisman, A. M., & Pashler, H. (2007). Characterizing the limits of human visual awareness. *Science (New York, N.Y.)*, *317*(5839), 823-5.

Hyun, J.-seok, Woodman, Geoffrey F, Vogel, Edward K, Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(4), 1140-60.

Ihssen, N., Linden, D. E. J., & Shapiro, K. L. (2010). Improving visual short-term memory by sequencing the stimulus array. *Psychonomic Bulletin & Review*, *17*(5), 680-6.

Intraub, H. (1980). Presentation rate and the representation of briefly glimpsed pictures in memory. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(1), 1-12.

Jiang, Y. V., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of experimental psychology. Learning, memory, and cognition*, *26*(3), 683-702.

Kyllingsbaek, S., & Bundesen, C. (2009). Changing change detection: improving the reliability of measures of visual short-term memory capacity. *Psychonomic Bulletin & Review*, *16*(6), 1000-10. doi:10.3758/PBR.16.6.1000

Lin, P.-H., & Luck, S. J. (2009). The Influence of Similarity on Visual Working Memory Representations. *Visual Cognition*, *17*(3), 356-372.

Linke, a C., Vicente-Grabovetsky, a, Mitchell, D. J., & Cusack, R. (2011). Encoding strategy accounts for individual differences in change detection measures of VSTM. *Neuropsychologia*, *49*(6), 1476-86. doi:10.1016/j.neuropsychologia.2010.11.034

Litz, B. T., Weathers, F. W., Monaco, V., Herman, D. S., Wulfsohn, M., Marx, B., & Keane, T. M. (1996). Attention, arousal, and memory in posttraumatic stress disorder. *Journal of Traumatic Stress*, *9*(3), 497-519.

Logie, R. H. (2011). The Functional Organization and Capacity Limits of Working Memory. *Current Directions in Psychological Science*, *20*(4), 240-245.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279-81. doi:10.1038/36846

Ma, W. J., Beck, J. M., & Pouget, A. (2008). Spiking networks for Bayesian inference and choice. *Current Opinion in Neurobiology*, *18*(2), 217-22.

Makovski, T., Watson, L. M., Koutstaal, W., & Jiang, Y. V. (2010). Method matters: systematic effects of testing procedure on visual working memory sensitivity. *Journal of Experimental Psychology Learning Memory and Cognition*, *36*(6), 1466-1479.

Makovski, T., & Jiang, Y. V. (2008). Proactive interference from items previously stored in visual working memory. *Memory & Cognition*, *36*(1), 43-52. doi:10.3758/MC.36.1.43

Makovski, T., Shim, W. M., & Jiang, Y. V. (2006). Interference from filled delays on visual change detection. *Journal of vision*, *6*(12), 1459-70. doi:10.1167/6.12.11

McGlynn, F. D., Wheeler, S. A, Wilamowska, Z. A, & Katz, J. S. (2008). Detection of change in threat-related and innocuous scenes among snake-fearful and snake-tolerant participants: Data from the flicker task. *Journal of Anxiety Disorders*, *22*(3), 515-523.

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

Nairne, J. S., & Neath, I. (2001). Long-term memory span. *Behavioral and Brain Sciences*, *24*(01), 134-135.

Olson, I. R., & Jiang, Y. (2002). Is visual short-term memory object based? Rejection of the "strong-object" hypothesis. *Perception & Psychophysics*, *64*(7), 1055-67.

Olsson, H., & Poom, L. (2005). Visual memory needs categories. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(24), 8776-80.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*(4), 369-78.

Phillips, W. (1974). On the distinction between sensory storage and short-term visual memory. *Attention, Perception, & Psychophysics*, *16*(2), 283–290.

Quinlan, P., & Cohen, D. (2011). Object-based representations govern both the storage of information in visual short-term memory and the retrieval of information from it. *Psychonomic Bulletin & Review*, *18*, 316-323.

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

Rensink, R. A. (2002). Change Detection. *Annual Review of Psychology*, *53*, 245-277.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367.

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*, 324-330.

Shaffer, W. O., & Shiffrin, R. M. (1972). Rehearsal and storage of visual information. *Journal of Experimental Psychology*, *92*(2), 292-6.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends in Cognitive Sciences*, *9*(1), 16-20. doi:10.1016/j.tics.2004.11.006

Song, J.-H., & Jiang, Y. V. (2006). Visual working memory for simple and complex features: an fMRI study. *NeuroImage*, *30*(3), 963-72. doi:10.1016/j.neuroimage.2005.10.006

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*(6984), 751-4. doi:10.1038/nature02466

Treisman, A. M. (1998). Feature binding, attention and object perception. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *353*(1373), 1295-306. doi:10.1098/rstb.1998.0284

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97-136. doi:10.1016/0010-0285(80)90005-5

Treisman, A. M., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. Psychological Review, 95(1), 15. American Psychological Association.

Treisman, A. M., & Sato, S. (1990). Conjunction search revisited. *Journal of experimental psychology. Human perception and performance*, *16*(3), 459-78.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*(1), 107-141. doi:10.1016/0010-0285(82)90006-8

Vasterling, J. J., Brailey, K., Constans, J. I., & Sutker, P. B. (1998). Attention and memory dysfunction in posttraumatic stress disorder. *Neuropsychology*, *12*(1), 125-33.

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. doi:10.1111/j.1467-8721.2008.00569.x

Vogel, E. K., Woodman, Geoffrey F, & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 92.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(6), 1436-51. doi:10.1037/0096-1523.32.6.1436

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*(1), 48-64. doi:10.1037//0096-3445.131.1.48

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120-1135. doi:10.1167/4.12.11

Wolfe, J. (2011). Searching for many things at the same time: Saved by a log. *Journal of Vision*, *11*(11), 1293-1293. doi:10.1167/11.11.1293

Woodman, G. F., & Vogel, E. K. (2008). Selective storage and maintenance of an object's features in visual working memory. *Psychonomic Bulletin & Review*, *15*(1), 223-229.

Xu, Y. (2007). The role of the superior intraparietal sulcus in supporting visual short-term memory for multifeature objects. *The Journal of Neuroscience*, *27*(43), 11676-86.

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91-5. doi:10.1038/nature04262

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233-5. doi:10.1038/nature06860

Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological science*, *20*(4), 423-8. doi:10.1111/j.1467-9280.2009.02322.x

Zhang, W., Johnson, J. S., Woodman, G. F., & Luck, S. J. (*in press*). Features and conjunctions in visual working memory. In J. M. Wolfe & L. C. Robertson (Eds.), *From Perception to Consciousness: Searching with Anne Treisman* (Vol. 5). New York, USA: Oxford University Press.