Attention, Perception, & Psychophysics (2020) 82:1271–1289
https://doi.org/10.3758/s13414-019-01809-1

Using the flicker task to estimate visual working memory storage capacity

Hrag Pailian[1] · Daniel J. Simons[2] · Jeffrey Wetherhold[1] · Justin Halberda[3]

Published online: 18 July 2019

Abstract
Studies of visual working memory (VWM) typically have used a "one-shot" change detection task to arrive at a capacity estimate of three to four objects, with additional limits imposed by the precision of the information needed for each object. Unlike the one-shot task, the flicker change detection task permits measurement of VWM capacity over time and with larger numbers of objects present in the scene, but it has rarely been used to assess the capacity of VWM. We used the flicker task to examine (a) whether capacity is close to the typical three to four items when using subtly different stimuli; (b) which dependent measure provides the most meaningful estimate of the capacity of VWM in the flicker task (response time or number of changes viewed); (c) whether capacity remains fixed at three to four items for displays containing many more objects; and (d) how VWM operates over time, with repeated opportunities to encode, retain, and compare elements in a display. Four experiments using grids of simple items varying only in luminance or color revealed a range for VWM capacity limits that was largely impervious to changes in display duration, interstimulus intervals, and array size. This estimate of VWM capacity was correlated with an estimate from the more typical one-shot task, further validating the flicker task as a tool for measuring the capacity of VWM.

Keywords Visual working memory · VSTM · Capacity · Memory · Similarity · Proximity · Change detection · Top-down

Most studies estimate visual working memory (VWM) storage capacity by measuring observers' ability to detect a change to one item in an array of several items. In general, performance is nearly perfect when the display contains three or fewer simple objects, but declines steadily when displays contain four or more objects. This transition from accurate to degraded performance is often taken to reflect a storage limit of three to four unified object representations at one time (Luck & Vogel, 1997; Pashler, 1988; Phillips, 1974; Pylyshyn & Storm, 1988; Sperling, 1960; Vogel, Woodman, & Luck, 2001). The measured capacity can be smaller for complex or difficult-to-distinguish objects (Alvarez & Cavanagh, 2004; Bays & Husain, 2008), suggesting an information-based limit in addition to the item-based one (but see Awh, Barton, & Vogel, 2007; Zhang & Luck, 2008). Other accounts using this method posit a VWM system that can store an unlimited number of items with decreasing information per item as the number of items increases (Wilken & Ma, 2004).

The "one-shot" change detection task, originally developed by Phillips (1974), has been the primary method for measuring the functional capacity of VWM. However, it has drawbacks. For instance, it is poorly suited to examining how observers use VWM over time when inspecting more complex displays with many objects. Moreover, variants of this task seem to produce estimates that are unreliable across set sizes, suggesting that the task is measuring different things at different set sizes (Pailian & Halberda, 2015). The reliability of these estimates can be improved, however, when task

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13414-019-01809-1) contains supplementary material, which is available to authorized users.

✉ Hrag Pailian
pailian@fas.harvard.edu

Daniel J. Simons
dsimons@uiuc.edu

Jeffrey Wetherhold
wetherhold@post.harvard.edu

Justin Halberda
halberda@jhu.edu

[1] Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138, USA

[2] University of Illinois, Chicago, IL, USA

[3] Johns Hopkins University, Baltimore, MD, USA

demands require participants to localize the change (Pailian & Halberda, 2015; Rensink, 2014). Another task commonly used to measure the precision of the information held in working memory (Anderson, Vogel, & Awh, 2011; Bays & Husain, 2008; Bays, Catalao, & Husain, 2009; Wilken & Ma, 2004; Zhang & Luck, 2008), the "continuous report" method, suffers from similar limitations. Observers view a small number of items and after a blank display, report the feature value for one of these items (e.g., "What color was here?"). Even if alternative methods were available, such tasks should not be abandoned, but theorizing about the functional operation of VWM may be enhanced by using tasks that require more extended viewing with larger numbers of items. Real-world scenes always include more than four simple objects, and people presumably use VWM as they interact with them over time. For example, they swap attended objects into and out of working memory.

An alternative change detection method, the flicker task (Rensink, O'Regan, & Clark, 1997), has mostly been used to document the extent to which people fail to notice changes to scenes (change blindness). In this task, an original and changed version of an image alternate repeatedly until observers find the changing item. Provided the two images are separated by a brief blank screen, people take a surprisingly long time to find the change, even when the change is large and easily seen once detected (Rensink et al., 1997). Despite its extensive use in the change blindness literature, with relatively few exceptions (Lleras, Rensink, & Enns, 2005, 2007; Pailian & Halberda, 2015; Rensink, 2000) this task has not been used to assess the limits of VWM storage. Yet it has several advantages, as well as some disadvantages, over one-shot tasks for measuring the functional limits of VWM. Specifically, it allows observers to engage in an extended search through alternating displays with many objects while still providing a useful dependent measure of change detection performance (RT, or number of alternations) and a derived estimate of VWM storage capacity (K). Flicker estimates are not only highly reliable across a range of set sizes but are also highly correlated with estimates from the one-shot task (Pailian & Halberda, 2015). These results suggested that the flicker task may be a valuable addition to the one-shot task when measuring individual differences in VWM capacity. In this article, we explore a wide range of variations in the flicker task to test its broader usefulness in measuring working memory capacity.

The logic underlying the estimate of capacity in the flicker task relies on the idea that increasing display times should allow people to encode more information into VWM, provided that they still have VWM capacity available (Rensink, 2000). Once VWM is filled to capacity, providing additional display time for encoding should not improve change detection performance. Using a measure derived from the asymptote of change detection performance as a function of display time, Rensink (2000) successfully estimated the capacity of VWM for orientation and polarity of items.[1]

Using the flicker task to measure VWM capacity does, however, introduce some complications and concerns. In the one-shot task, encoding occurs only during the time-limited initial display presentation, and comparison occurs only during the longer post change display. This separation makes it straightforward to isolate the contribution of each to memory. In contrast, in the flicker task, observers encode and compare information during each display presentation. One goal of our work is to test the hypothesis that if displays are presented long enough for the encoding and comparison processes to run to completion, the VWM capacity estimates measured using the flicker task and the one-shot task will converge. If they do, then this "drawback" becomes an advantage in that two tasks with different dependent measures (i.e., accuracy in one-shot and RT in flicker), different numbers of targets (e.g., four vs. unlimited [in theory]), and different viewing conditions (e.g., single viewing for one-shot and repeated viewings with movements of attention during search for flicker) can provide converging evidence for a similar limit to VWM. Testing the convergence between these two measures, as well as the reliability of each measure, is a goal of the present work and of some of our previous work (Pailian & Halberda, 2015).

Another concern is that the flicker task involves visual search for a changing item, so deriving an estimate of VWM capacity requires assumptions about the mechanisms of visual search. Rensink (2000) assumed that participants were optimally efficient in their search for the changing item (e.g., never revisiting an already-visited item), and tested this assumption using an analysis based on the number of alternations required to identify the target. However, actual search might be less efficient than this, particularly with displays containing larger numbers of items. We instead provide a range of capacity estimates while assuming either efficient or inefficient search.

For both the flicker and one-shot tasks, capacity estimates might be inflated by grouping strategies. The discrepancy between Rensink's (2000) capacity estimates for orientation (5.5 items) and polarity (9 items) might have more to do with the nature of the stimuli than with the method itself. Observers might have stored perceptual groups or clusters of similar objects in a single VWM "slot" (see Rensink, 2000, for discussion of these issues). Here, we minimized the role of grouping by using large numbers of visually simple objects

[1] The logic of inferring VWM capacity relies on the assumption that long-term memory (LTM) does not contribute to performance in the flicker task. The validity of this assumption can be tested by extending display durations. If long-term memory additionally contributes to flicker performance once VWM is filled, then capacity will increase with increasing display times, subject to the limits of long-term memory. This type of pattern has not been observed in extant studies, suggesting negligible contributions from LTM.

that were not easily encoded verbally or grouped perceptually (differing only in luminance or color).

Across four experiments, we seek to answer the following questions: (a) Do flicker estimates of storage capacity correlate with those produced by the one-shot task? (b) What is the range of capacities in the flicker task consistent with efficient and inefficient search? (c) Which dependent measure in the flicker task provides the most appropriate measure of VWM capacity, overall response time or number of changes viewed? (d) How might estimates of the capacity of VWM change for displays with far more objects than the capacity of VWM? And, (e) How does VWM operate over time, with repeated opportunities to encode, retain, and compare elements?

General methods

The experiments in this article were conducted using different technologies and equipment. Experiment 2 (conducted from 2001 to 2002) was coded using Vision Shell software and presented on Macintosh iMac computers with built-in CRT monitors. This experiment, conducted many years ago, used smaller samples than would be typical now, but which were common in visual cognition research at the time. Returning to our questions of interest years later (Experiment 3: 2009, Experiment 1: 2014–2015, Experiment 4: 2015–2016), we replicated and extended the earlier results. Experiments 1, 3, and 4 were coded using MATLAB and Psychophysics Toolbox and were presented on Macintosh iMac computers with LCD monitors. The experiments are presented out of temporal sequence for the purpose of clearer exposition.

Experiment 1

Perhaps the most straightforward way to compare capacity estimates derived from the flicker task and the one-shot task is to have the same observers complete both tasks.

Method

Observers ($n = 31$) from Johns Hopkins University participated in a two-block, within-subjects experiment, with the block order counterbalanced across observers. During one block of 80 trials, observers performed the flicker change detection task with arrays of four or nine dots, a display duration of 700 ms, and an interstimulus interval (ISI) of 900 ms. The observer's task was to press a button to stop the alternations as soon as they spotted the changing item (see Fig. 1a). After pressing the space bar to end the alternations, they clicked on the item that had been changing in a static image of the array.

For the other block of trials, observers performed a one-shot change detection task with arrays of four or nine dots,

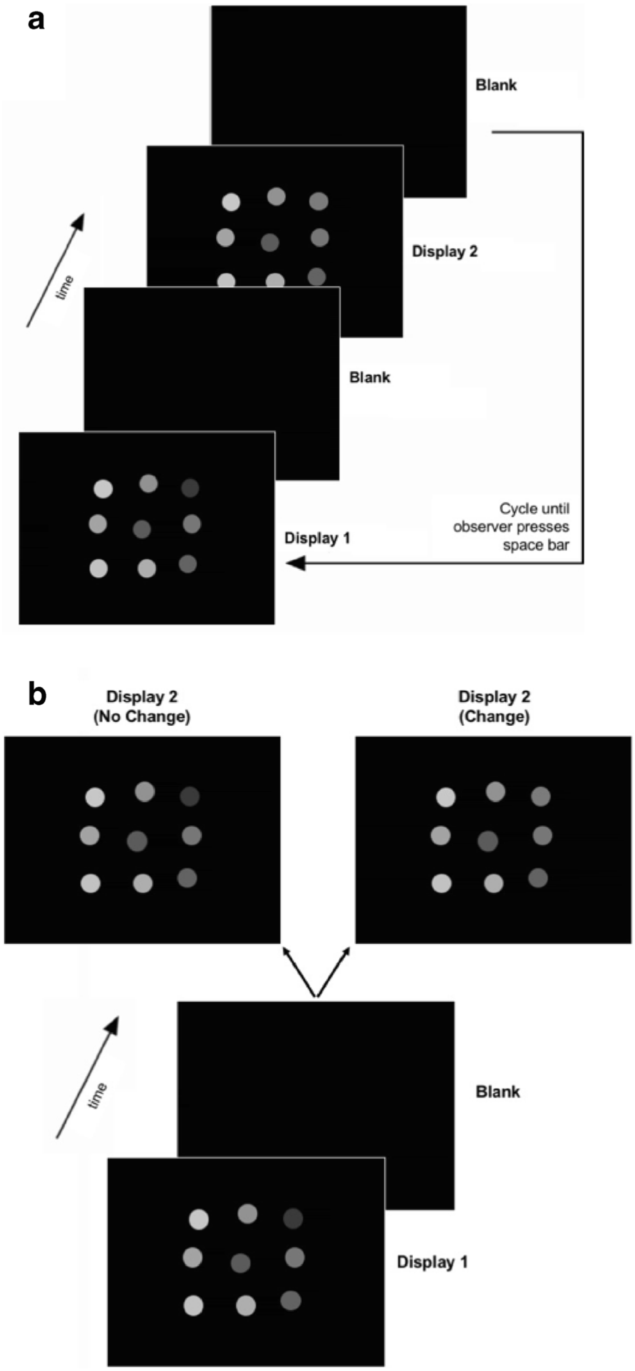

Fig. 1 The trial structure for (**a**) flicker and (**b**) one-shot experiments involving a luminance change among grayscale dots

using settings typical of one-shot tasks (e.g., Luck & Vogel, 1997; Vogel et al., 2001)—a first display duration of 100 ms, an ISI of 900 ms, and a second (comparison) display duration of up to 2,000 ms (see Fig. 1b). For the one-shot block, 50% of the trials included a change to a single item, and the observer's task on each trial was to determine whether or not a change had occurred.

In order to create displays that were similar to those typically used in one-shot experiments, the position of each dot was set to a grid position (mean center-to-center distance = 2.21 degrees of visual angle, diameter of each dot = 1.5 degrees of visual angle) with an additional random jitter (range: −.25 to +.25 degrees of visual angle in both horizontal and vertical directions). We used this same algorithm for the displays for both the flicker and one-shot blocks.

Each dot was randomly assigned one of 30 equally spaced luminance values ranging from 4.65 to 61.51 cd/m^2 (possible values ranged from 55 to 200 in 5-volt intervals). In both flicker and one-shot, the changed screen was identical to the first, except that one dot changed luminance by ±14 shades (70 volts). For these relatively large changes, most items could either increase or decrease in luminance within the allowable range, but not both. For those cases in which the changed item was capable of becoming either lighter or darker, the direction of change was selected at random before the trial and maintained throughout the trial.

For the flicker block, response times were transformed into estimates of VWM storage capacity (K) using an equation described in the results section. For the one-shot block, change detection accuracy was transformed into an estimate of VWM storage capacity (K) using Pashler's (1988) equation. This K estimate is derived from the hit rate and false-alarm rate for detecting a single change in the array, but does not require observers to localize the changed item.

Observers received course credit for participating. Five additional observers produced an unusually large number of wrong responses in the one-shot block, yielding negative values for K estimates. Given that it is not possible to have a negative storage capacity, data from these observers were removed from analysis.

Results and discussion

Capacity for the flicker block was calculated separately for each set size. All trials for which observers selected the wrong item (6.0% of trials) and all trials with response times more than two standard deviations above or below an observer's mean for that block (5.2% of trials, range: 1.52%–9.46% across blocks) were eliminated from further analyses. Across both set sizes, observers detected changes after about 4.66 s (*SD* = 1.65 s) on average.

Estimating capacity in the flicker task

Both response latency and the number of alternations required for detection allow us to derive an estimate of the capacity of VWM from the flicker task: the estimated number of items held and compared during each display duration. Because information about the display must be stored in VWM over the course of each ISI/ blank screen, the capacity of VWM places an upper bound on performance in the flicker task—just like in the more traditional one-shot task. By including additional display time, other factors—such as crowding in the display, possible saccades made during encoding, and the time required to decide that no change has been detected—can be eliminated as limiting factors in the flicker task. Once sufficient display time has been provided, these processes should reach completion, and additional display time should no longer improve performance because no additional information can be held in VWM across the ISI.

Adapting earlier work (Rensink, 2000), we developed a method to transform flicker response times (RTs) into estimates of VWM storage capacity (K). Recently, we have used this method to assess VWM capacity in adults (Pailian & Halberda, 2015) and in children (Pailian, Libertus, Feigenson, & Halberda, 2016). Our approach is based on the assumption that the amount of time it takes to find a target item is determined by two factors: (1) non-search-related activity and (2) search-related activity:

$$\mathrm{RT} = \mathrm{non{-}search{-}related\ RT} + \mathrm{search{-}related\ RT}. \qquad (1)$$

Non-search-related RT is intended to capture the time taken to perform executive control processes both before and after search that are not directly related to active search. These will include, but are not limited to, (1) processing the initial visual display, (2) initiating search, (3) storing the first set of items, (4) verifying that the target is indeed changing, and (5) initiating and executing a button press. At a minimum, this non-search-related RT must be greater than the display duration + ISI because no change can be detected during the first display on time and subsequent ISI. An empirical approach to estimating all non-search-related contributions to response latency is to compute the average of each observer's fastest detections for each trial type. This approach is motivated by the intuition that, on occasion, the observer will happen to find the changing item in the very first set of items stored in memory and compared. Thus, the fastest response times can serve as an estimate of the time required for these non-search-related activities.[2]

Next, the remaining search-related activity will allow us to estimate VWM capacity. Straightforwardly, the search-related activity can be estimated as the display duration + ISI multiplied by the number of items that must be searched, divided by VWM capacity (K):

[2] We estimated non-search-related activity by the average of the three fastest RTs within a condition. This was informed by previous work (Pailian & Halberda, 2015; Pailian et al., 2016, Fig. 5) where the three fastest RTs were found to be a stable and useful estimate. In additional analyses of the current experiments, we have also made the number of trials dependent on the estimated capacity and array size, with the understanding being that as the array size becomes larger, the number of trials on which an observer is likely to find the changing target on their first active search becomes smaller. These analyses returned similar results to those reported here. We also conducted additional analyses using the single fastest detection and the two fastest detections. Besides being more volatile than using the three fastest detections, results were qualitatively the same. We hope that future modeling work will refine these parameters and how they are fit. For the present we chose to continue our approach of using the three fastest RTs to estimate non-search-related activity.

$$\mathrm{RT} = \text{non-search-related RT} + \frac{(\text{display duration} + \text{ISI}) \times (\text{number of items searched})}{\mathrm{K}} \tag{2}$$

That is, response time will be influenced by the temporal parameters of the task—including the amount of time that a single display is presented on the screen (display duration) and the amount of time that separates the two alternating displays (ISI). These parameters interact with an individual's search rate, dependent on their search strategy (i.e., number of items that must be searched), scaled by their capacity (K) to determine how long that person will need to search before finding the target.

Search efficiency is an important factor for estimating VWM capacity in the flicker task. One straightforward way of understanding search efficiency is by the average number of items that observers search through prior to identifying the target item (e.g., do they revisit items or do they work serially through items one by one). Consistent with models of visual search, if search is efficient (e.g., random without replacement), observers will, on average, have to search through roughly half of the display (set size + 1/2) before coming upon the target item (Johnson & Kotz, 1977; Rensink, 2000; Treisman & Gelade, 1980). Alternatively, if search is inefficient (e.g., random with replacement, allowing for repeats), observers will, on average, have to search through a number of items equal to the set size presented in the display before finding the target item (Horowitz & Wolfe, 1998, 2003; Johnson & Kotz, 1977). Assuming that there is some forgetting during the ISI, the resulting search in the flicker task likely falls somewhere between these two extremes of efficient and inefficient search (i.e., observers might revisit some items that were previously stored but forgotten during the ISI, or they may revisit items that were visited earlier in their search). Because using VWM in natural scenes likely involves searches among many items (where items will sometimes be revisited), attention to how visual search interacts with VWM capacity may be a positive aspect of using the flicker task to study VWM rather than a simple confound. In the present analyses, we will provide an estimate of the upper and lower bounds for VWM capacity based on efficient and inefficient visual search. As we will see, providing this range allows a for a more honest reporting of the range of possible capacities, but in no way hinders our ability to generate productive individual difference measures of this capacity.

Lastly, search rate will be affected by individual differences in the capacity of VWM. In order to find the changing stimulus, observers must load a number of items into VWM (storage capacity, K), maintain these during the blank ISI, and then compare them to the items presented in the subsequent display. An observer who has a VWM storage capacity of four items will, on average, be able to detect a target that is part of a four-item display within a single unfolding of events. In contrast, an observer who can maintain only one item in VWM will, on average, take more than twice as long to find the target item; they will have to repeat the encoding, maintenance, and comparison processes (as well as switch attention between items in the display) until they come upon the target item. The logic of using the flicker task is that sufficient display durations will allow VWM capacity to be the dominant factor in determining typical search times.

The contribution of all of the above factors to flicker response times can be mathematically expressed using the following equations:

Assuming search is efficient:

$$\mathrm{RT} = \text{non-search-related RT} + \frac{(\text{display duration} + \text{ISI}) \times [(\mathrm{N}+1)/2]}{\mathrm{K}} \tag{3a}$$

Assuming search is inefficient:

$$\mathrm{RT} = \text{non-search-related RT} + \frac{(\text{display duration} + \text{ISI}) \times (\mathrm{N})}{\mathrm{K}} \tag{3b}$$

These formulas can be rearranged to provide equations for estimating VWM storage capacity (K):

Assuming search is efficient:

$$\mathrm{K}_{\text{efficient}} = \frac{(\text{display duration} + \text{ISI}) \times [(\mathrm{N}+1)/2]}{\text{RT} - \text{non-search-related RT}} \tag{4a}$$

Assuming search is inefficient:

$$\mathrm{K}_{\text{inefficient}} = \frac{(\text{display duration} + \text{ISI}) \times (\mathrm{N})}{\text{RT} - \text{non-search-related RT}} \tag{4b}$$

We used Equations 4a and 4b to compute estimates of storage capacity. Given that the values for efficient and inefficient search differ by a constant ratio, regression analyses performed on these estimates will yield identical results, and the individual differences documented by one type of search will be identical to the other. Therefore, to reduce repetition, we present analyses using a single "K" variable that can stand for either efficient or inefficient K.

First, we examined response times. As predicted, participants took longer and required more exposures to detect the changing item for displays consisting of nine items compared with four items, $t(30) = -11.67$, $p < .001$ (see Fig. 2a). They also took longer to process the display, engage in search, and make a decision for the larger set size as indicated by a statistically significant difference in our estimate of non-search-related activity, $t(30) = -5.60$, $p < .001$ (see Fig. 2b). Despite these differences, and as predicted, K estimated from the flicker task was not significantly different between set

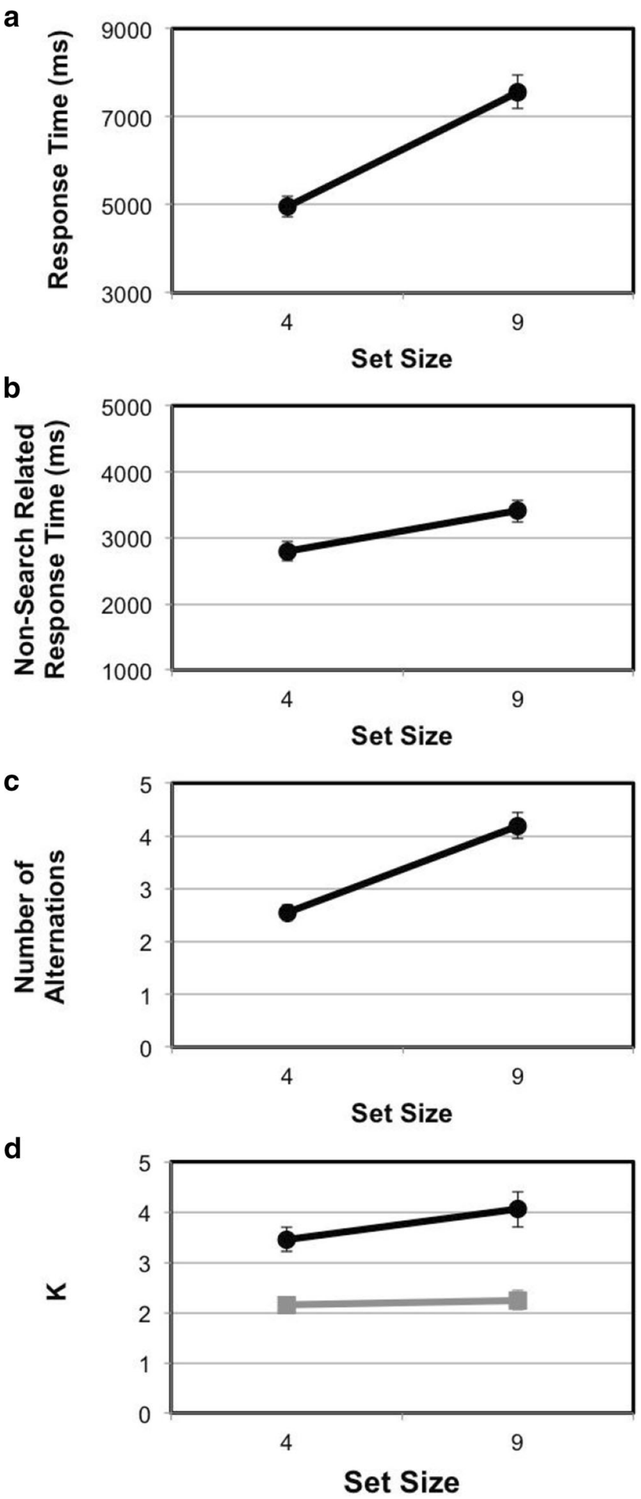

Fig. 2 Results of Experiment 4 (array size = varied, display duration = 700 ms, ISI = 900 ms). **a** Mean response time (RT) in ms (±*SE*) for the two set sizes. **b** Non-search-related activity as operationalized by the average of three fastest change detection RT's in ms (±*SE*). **c** Mean number of alternations required for successful change detection (±*SE*). **d** Estimated capacity assuming efficient (gray lines) and inefficient search (black lines) at each of the 2 set sizes (±*SE*)

sizes, $t(30) = -0.54$, $p = .60$ (see Fig. 2d). Although this null result was consistent with our expectations, readers should not draw strong conclusions from the absence of a significant difference.

Typically, in the one-shot task, capacity (K) estimates are averaged across set sizes (see Pailian & Halberda, 2015, for evidence that estimates do not correlate well with one another across set sizes in the whole-report version of the one-shot task, suggesting that the task measures different things at different set sizes). We adhere to that typical practice and compare estimates of capacity from the flicker task and the one-shot task when averaging across set size (e.g., Anderson, Laurent, & Yantis, 2011; Fukuda & Vogel, 2011). One-shot K estimates ($K_{Average} = 3.1$) were significantly higher than efficient flicker K estimates ($K_{Average} = 2.2$), $t(30) = -4.61$, $p < .001$, and significantly lower than inefficient flicker K estimates ($K_{Average} = 3.8$), $t(30) = 2.84$, $p < .01$.

Differences between one-shot and flicker K estimates might reflect the use nonmemory factors in one task but not the other, such as global attention (Linke, Vicente-Grabovetsky, Mitchell, & Cusack, 2011; Matsuyoshi, Osaka, & Osaka, 2014), filtering efficiency (Vogel, McCollough, & Machizawa, 2005), ensemble processing (Brady & Tenenbaum, 2013), and so forth. Nevertheless, these results demonstrate that the range of estimates from the flicker task roughly converge with those produced by the one-shot task.

Most importantly, overall capacity estimates derived from the flicker task and the one-shot task were positively correlated, $r = .40$, $p = .03$, 95% CI [0.05, 0.66] (see Fig. 3).[3] Spearman–Brown-corrected split-half reliabilities (Brown, 1910; Spearman, 1910) showed that the flicker task provides a highly reliable ($r_{SB} = .88$, $p < .001$) measure of K, and that the one-shot task is less reliable ($r_{SB} = .55$, $p < .01$). The less-than-perfect reliability of these measures sets an upper bound on the correlation we could observe between capacity estimates in the two tasks: The maximum possible correlation would be .56 [(max $r_{A,B} = (r_{ObservedA,ObservedB}) \div \sqrt{(\text{reliability}_A \times \text{reliability}_B)}$—see Nunnally, 1970; Vul, Harris, Winkielman, & Pashler, 2009], so the observed correlation of .40 is relatively strong. Indeed, it suggests an effective relationship of closer to .7 between these tasks, although this estimate has a wide confidence interval due to our small sample size.

The results of Experiment 1 suggest that estimates from the flicker task correlate strongly with the one-shot task and that the flicker task might provide a more reliable estimate of individual differences in working memory than the one-shot task (see also Pailian & Halberda, 2015). Despite the complication of a visual search component to the flicker task, it

[3] Given that inefficient and efficient K estimates differ based on a scaling factor, the correlation between inefficient K estimates and one-shot K estimates is identical to the correlation between efficient K estimates and one-shot K estimates.

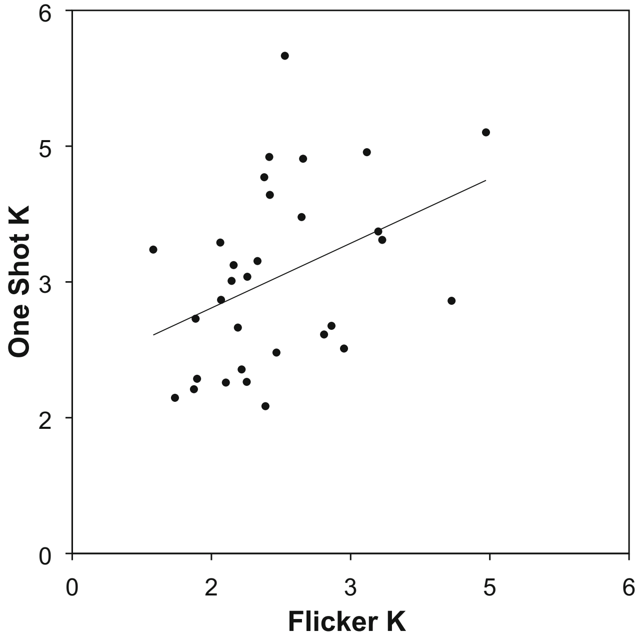

Fig. 3 Correlation between estimates of VWM storage capacity, K, as measured by the one-shot and flicker (efficient search) tasks

appears to provide converging evidence for a capacity estimate in the same range as that estimated by the one-shot task.

The flicker task provides two estimates of memory capacity that differ only by a scaling factor. Consequently, they provide equivalent measures of individual differences in capacity. However, people may differ in which form of search they use. The fact that capacity estimates from the flicker task are correlated with estimates from the one-shot task—which does not involve search—supports the idea that both tasks tap a common underlying capacity.

The remaining experiments examine the stability of VWM estimates from the flicker task under a variety of conditions.

Experiment 2

One of the benefits of the flicker task is the possibility to test how VWM functions with larger arrays of objects. Experiment 2 used displays of 36 items and varied display duration to estimate VWM storage capacity in the flicker task. If the capacity of VWM is limited, then, with a sufficient display time, observers should be able to load VWM to capacity for comparison to the next display. If they do not find a change, they would then need to refill their VWM buffer to capacity with new items. We can use the total change detection time to estimate the number of items being held in VWM during each alternation (cf. Rensink, 2000). Unlike Rensink (2000), we used arrays of grayscale dots with a wide range of shades, making verbal categorization and perceptual grouping more difficult.

Method

Observers

Thirteen undergraduate students from Harvard University with self-reported normal or corrected-to-normal vision were recruited from a psychology subject pool and participated in exchange for course credit.

Displays and procedure

Observers were tested using Macintosh iMac computers with built-in CRT monitors (viewable area: 29.5 cm × 22.5 cm) running on Mac OS9 operating system. The experiments were coded using Micro M-L's Vision Shell C libraries (a commercial package that is no longer available). Viewing distance was unconstrained, but averaged approximately 57 cm, and all visual angle measurements below are based on that distance.

The settings for dot size and change luminance were the same as in Experiment 1. One difference is that in Experiment 2, with 36 total objects present, the items were arranged in a strict 6 × 6 grid (see Fig. 4) such that the centers of adjacent dots were separated by 2.66 degrees. The display duration varied across five blocks of trials for each observer (300 ms; 500 ms; 700 ms; 900 ms; or 1,100 ms), with 20 trials per block. Block order was randomized for each observer.

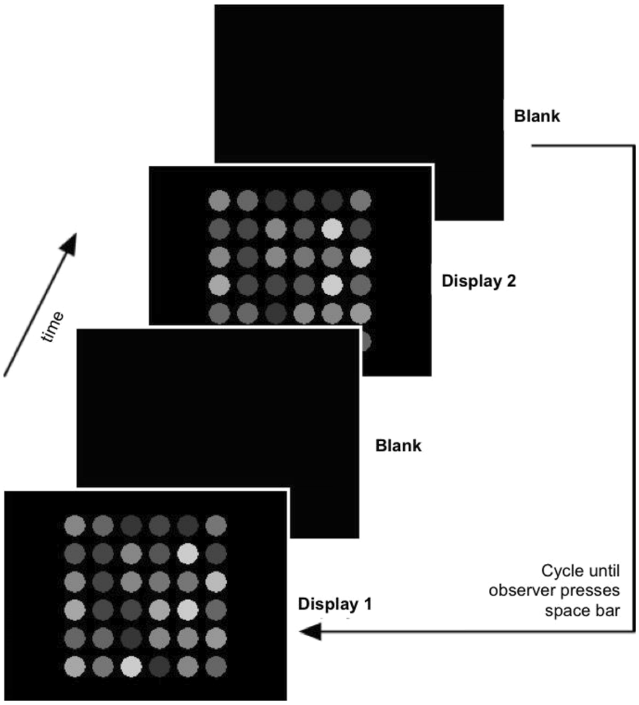

Fig. 4 The trial structure for flicker experiments involving a luminance change among 36 grayscale dots. The cycle of the displays repeated until the observer pressed the space bar to indicate change detection. In this figure, the changing dot is located in the fourth column from the left and in the third row from the bottom

Within a trial, the display duration was the same for the original and the changed array. All displays were separated by a 120-ms blank screen to prevent "pop-out" effects arising from the perception of retinal transients (Blackmore, Brelstaff, Nelson, & Troscianko, 1995; French, 1953; Pashler, 1988; Simons, 1996), and iconic memory was presumably disrupted by the presentation of the subsequent display (Becker, Pashler, & Anstis, 2000). These timings were chosen with the aim that observers must store representations of the grayscale dots in VWM to detect the change. Once the change was detected, observers pressed the space bar to stop the alternation and then used the mouse to click on the changing dot. Response times were recorded from the onset of the first display until the observer pressed the space bar, and click accuracy was recorded. Observers were instructed to respond as quickly as possible while prioritizing accuracy.

Results and discussion

All trials for which observers selected the wrong item were eliminated from further analyses (5.3% of trials). To eliminate unusually long or short response times that might have resulted from factors unrelated to the task, we removed any trials for which response latency was more than two standard deviations (Pailian & Halberda, 2015; Pailian et al., 2016) above or below an observer's mean for that display duration (mean = 5.4% of trials, range: 3%–8% across blocks). Analyses using the median RT without excluding slow or fast trials produced similar results (see the Supplementary Materials). Latency was measured from onset of the first display until the observer pressed the space bar, so it included both display times and blank screens. Across all display durations, observers detected changes after about 13 s (*SD* = 3.9 s) on average.

To capture changes in performance over time for individual participants in this task, we fit a series of bilinear spline regression models to their data based on the idea that performance would asymptote. Bilinear spline regression provides a means to describe how the relationship between two measures (e.g., response time and display duration) differs based on whether VWM has or has not been loaded to capacity. A given measure might increase as a function of display duration before reaching an asymptote or it might remain constant across a series of initial display durations and increase systematically thereafter. We fit a series of spline models with different knot locations (no knot/simple linear model vs. knot placed at display durations of 500 ms, 700 ms, or 900 ms). The best fit across participants was evaluated based on the sum of squared prediction errors (for modeling results for individual participants, see the Supplementary Materials). Parameter estimates for the best-fitting model in each experiment are summarized in Table 1.

Average response times (see Fig. 5a) were best approximated by a model where response latencies were constant for display durations from 300 ms to 700 ms (slope = 0.89 s of response time / 1 s of display time, p = .84), but increased thereafter (see Fig. 6a). However, this specific bilinear model provided the best fit for only two of the 13 participants. For a majority of participants (7/13), the best-fitting model had constant response latencies from 300 ms to 500 ms and a linear increase thereafter. Across participants, these results suggest that providing more time (up to about 700 ms) allowed observers to store more information, leading to a fairly constant RT to find a change (see Rensink, 2000, 2014, for a similar result).

For the shorter display durations used in Experiment 2 (i.e., 300–900 ms), change detection appears to be process limited; participants did not have enough time during each display to complete all mental operations involved in the task and load VWM to full capacity (although the amount of time needed might vary across individuals). These results differ from those observed by Rensink (2014), who found that display durations as short as 120 ms were sufficient to support memory consolidation during the blank interval. Although it remains an empirical question, the 120-ms asymptote that Rensink observed might reflect contributions from iconic memory, or the amount of time to needed consolidate clusters of items, rather than the time to consolidate each item individually. Our results are consistent with one-shot studies that contain heterogeneous displays (Vogel et al., 2001), which show that more items can be consolidated during the blank duration as time (either display duration or blank interval duration) is extended, with consolidation requiring approximately 50 ms per item.

Interestingly, the constant change detection latencies during this process-limited stage mean that performance was comparable even though shorter display durations provided more opportunities to see the change than longer ones. For example, if change detection took 14 seconds total, observers would have had 32 exposures to the change with a display duration of 300 ms (i.e. = 32 × (300 ms + 120 ms) + 420 ms), but only 22 exposures with display durations of 500 ms and 16 exposures with a display duration of 700 ms.

Figure 5c plots the number of exposures needed for change detection for each display duration. Bilinear spline regressions on the number of change exposures rather than response times (see Fig. 6c) revealed that the group averages were best fit by a model where the number of alternations linearly decreased across durations of 300 ms to 700 ms (slope = −28.0 alternations / 1 s of display time, p < .001), but remained constant thereafter (slope = −0.8 alternations / 1 s of display time, p = .90). This model best fit the data of 5 of the 13 participants, whereas the remaining eight were best fit by a model where the number of alternations decreased across durations of 300 ms to 500 ms, and remained constant thereafter. These results suggest that for display durations ranging from 300 ms up to 700 ms, the overall encoding time rather than the number

Table 1 Summary of estimates produced by the best-fitting bi-linear spline regression for each experiment

		Knot placement	Spline 1		Spline 2		R^2
			Slope	Sig. (p)	Slope	Sig. (p)	
EXP 2	**Response time** (s / 1 s of display time)	700 ms	0.89	.84	11.37	.01	0.15
	Non-search-related response time (s / 1 s of display time)	700 ms	0.95	.44	2.13	.09	0.11
	Number of alternations (alt's / 1 s of display time)	700 ms	−28.00	<.001	−0.78	.90	0.36
	K (items / 1 s of display time)	700 ms	2.77	.001	−0.23	.78	0.20
EXP 3	**Response time** (ms / 1 scaled display time)	Set Size 25	19.97	<.001	33.73	<.001	0.77
	Non-search-related response time (ms / 1 scaled display time))	Set Size 25	7.36	<.001	17.35	<.002	0.47
	Number of alternations (alt's / 1 scaled display time))	Set Size 25	0.46	<.001	0.81	<.001	0.77
	K (items / 1 scaled display time))	Set Size 9	0.04	.71	−0.11	.01	0.12
EXP 4	**Response time** (s / 1 s of display time)	500 ms	−2.80	.08	2.27	<.001	0.23
	Non-search-related response time (s / 1 s of display time)	500 ms	0.31	.73	1.44	<.001	0.34
	Number of alternations (alt's / 1 s of display time)	500 ms	−5.05	<.001	−0.50	.14	0.46
	K (items / 1 s of display time)	700 ms	3.44	<.001	0.23	.78	0.32

of change exposures determined when observers detected the change.

In contrast, for longer durations (700–1,100 ms), response latency increased linearly as a function of display duration (slope = 11.37 s of response time / 1 s of display time, p = .01). However, the number of exposures needed for change detection remained constant at approximately 14 cycles (see Fig. 5c). Thus, for display durations from 700 ms to 1,100 ms, performance appears to be capacity limited. Within a single 700 ms on-time, observers compared the current display with information stored in VWM as well as they could, and if they failed to detect a change, they had adequate time to shift attention and refill VWM to capacity with new information. Apparently, an additional 400 ms of viewing for each display did not allow observers to encode or compare any additional information because the number of cycles required to find the changing target remained constant at approximately 14 cycles. This relationship between display duration and the number of cycles required to detect a change can inform an estimate of VWM capacity in the flicker task (see the Estimating Capacity in the Flicker Task section above).

The finding that response latency and the number of change exposures differ as measures of change detection performance is surprising given that, on first blush, both appear to measure the rate to find a change. In fact, the results of Experiment 2 suggest that these two measures rarely converge when comparing across display durations. For display durations from 300 ms to 700 ms (or 500 ms for some participants), response time to detect a change was constant, while the number of cycles required steadily decreased. In contrast, for display durations of 700 ms to 1,100 ms, the number of cycles required for detection was constant, while response time steadily increased.

Perhaps the number of cycles and the total response time are only comparable in a flicker task when the display duration gives observers just enough time to compare, shift attention, and load VWM to capacity, but no more. For our displays, that critical duration averaged approximately 700 ms (see Rensink, 2000, for a similar estimate). Alternatively, this critical point might reflect the time when participants switch strategies to minimize the trade-off between effort (how many items they should try to remember) versus reward (how quickly they finish the experiment). Namely, when display durations are short enough (e.g., 300–700 ms), participants may choose not to load VWM to capacity, as this would be relatively effortful. However, such a strategy would prove less optimal for longer durations (e.g., 700–1,100 ms), as it would significantly extend the duration of the experiment. As such, response times observed in the current experiment may not have reached asymptote until 700 ms because individuals chose not to load VWM to capacity rather than because they were unable to do so; further evidence would be needed to distinguish between these possibilities.

Figure 5b displays the average of the three fastest detections separately for each display duration. Bilinear spline regressions (see Fig. 6b) revealed that these values remained constant across display durations of 300–700 ms (0.95 s of response time / 1 s of display time, p = .44), as well as across durations of 700–1,100 ms (2.13 s of response time / 1 s of display time, p = .09). Overall, the fastest response times increased as a function of display time because no change could be detected during the first display presentation regardless of how long it was visible. In fact, the fastest detection time across durations of 700–1,100 ms increases by approximately 200 ms for each increase of 200 ms in display time, suggesting that the time required for non-search-related factors remained

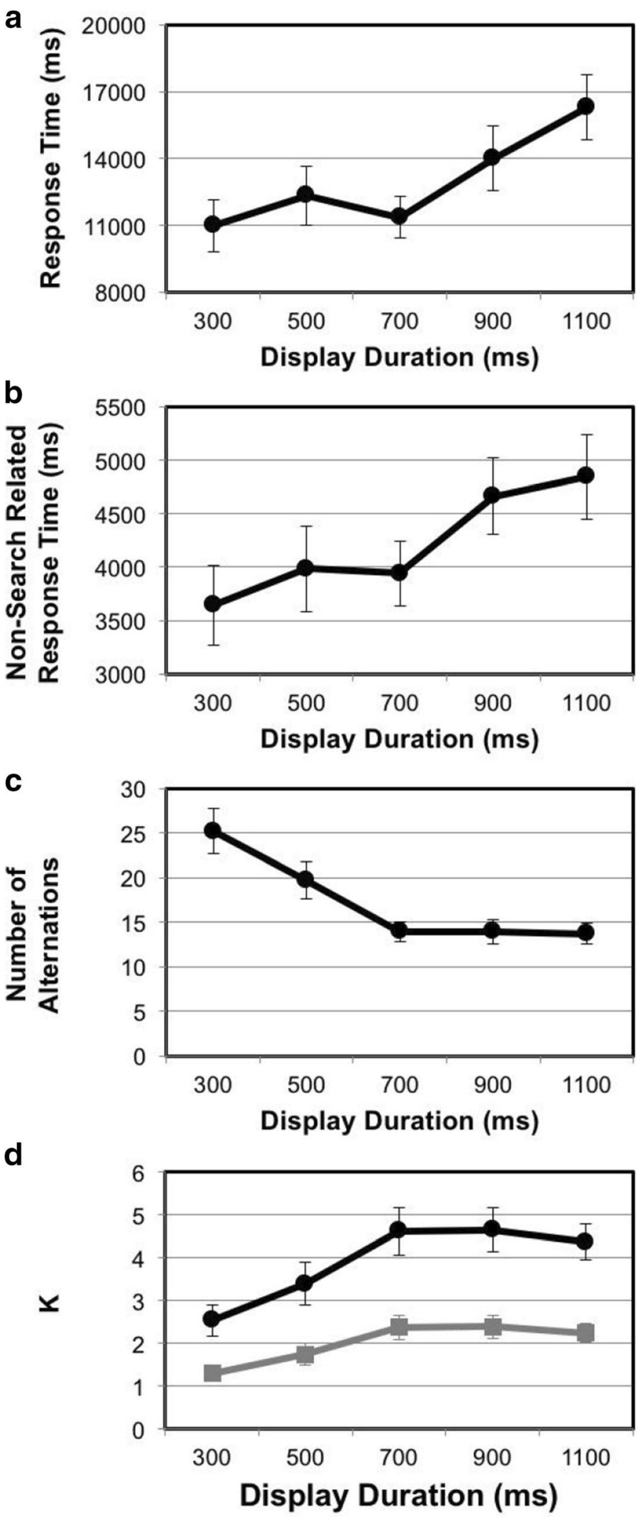

Fig. 5 Results of Experiment 1 (array size = 36, display duration = varied, ISI = 120 ms). **a** Mean response time (RT) in ms (±*SE*) for the five display durations. **b** Non-search-related activity as operationalized by the average of three fastest change detection RT's in ms (±*SE*). **c** Mean number of alternations required for successful change detection (±*SE*). **d** Estimated capacity assuming efficient (gray lines) and inefficient search (black lines) at each of the 5 display durations (±*SE*)

roughly constant across display times once the additional on-time is removed.

Bilinear spline regressions (see Fig. 6d) revealed that average K estimates increased for display durations up to 700 ms (slope = 2.77 items stored / 1 s of display time, p = .001), and did not increase more with longer display times (slope = −0.23 items stored / 1 s of display time, p = .78). This pattern of results best described the data produced by 4 of the 13 participants. Consistent with the RT analysis, performance was, on average, display-time limited until 700 ms but capacity limited with longer display times (see Fig. 5d). Specifically, capacity limits asymptoted at approximately 2.4 items assuming efficient search and 4.6 items assuming inefficient search. A range of 2.4 (inefficient) to 4.6 (efficient) items agrees passably well with the typical three-item capacity estimates derived from the one-shot task, and with the results of Experiment 1. The estimate in Experiment 2 was derived from continuously presenting a changing item among 36 items, far greater than is possible in the one-shot task.

In Experiment 2b (see the Supplementary Materials), we extended display durations up to 1,900 ms, and we increased the ISI to 900 ms. These times are comparable to the longest (memory-test) portion of a typical one-shot change detection task and the ISI length of most one-shot tasks (e.g., Luck & Vogel, 1997). When we originally conducted this follow-up experiment, in 2002, we predicted that we would observer constant capacity across these display times (from 1,100 ms to 1,900 ms). However, we now recognize that the study was underpowered to provide evidence for constancy across display times (evidence for no difference). The results were consistent with Experiment 2 (e.g., a response time slope = 8.2 seconds of increased response time per 1 second increase in display time), and a statistical test did not reject the null result of no change in capacity across display times, but the absence of statistical significance is not evidence for the absence of a difference. The original study was underpowered to provide meaningful evidence using equivalence testing or Bayes factors. Consequently, we describe it here as a suggestive pattern, with details presented in supplementary information as Experiment 2b. Collectively, these studies indicate that ISI is not a major factor, that capacity is reached by 700 ms of display on-time, and that capacity estimates may be stable with times as long as 1,900 ms.

That estimated average capacity asymptoted at a display duration of 700 ms is consistent with evidence from other studies (Rensink, 2000). This value makes some sense when we consider the processing involved: If it takes approximately 150 ms to load information from three items into VWM (Vogel, Woodman, & Luck, 2006; Woodman & Vogel, 2005), then optimal performance in the flicker task would require observers to use the remaining 550 ms within each display (i.e. 700 ms − 150 ms) to (a) compare the stored information to the currently visible display, (b) decide that

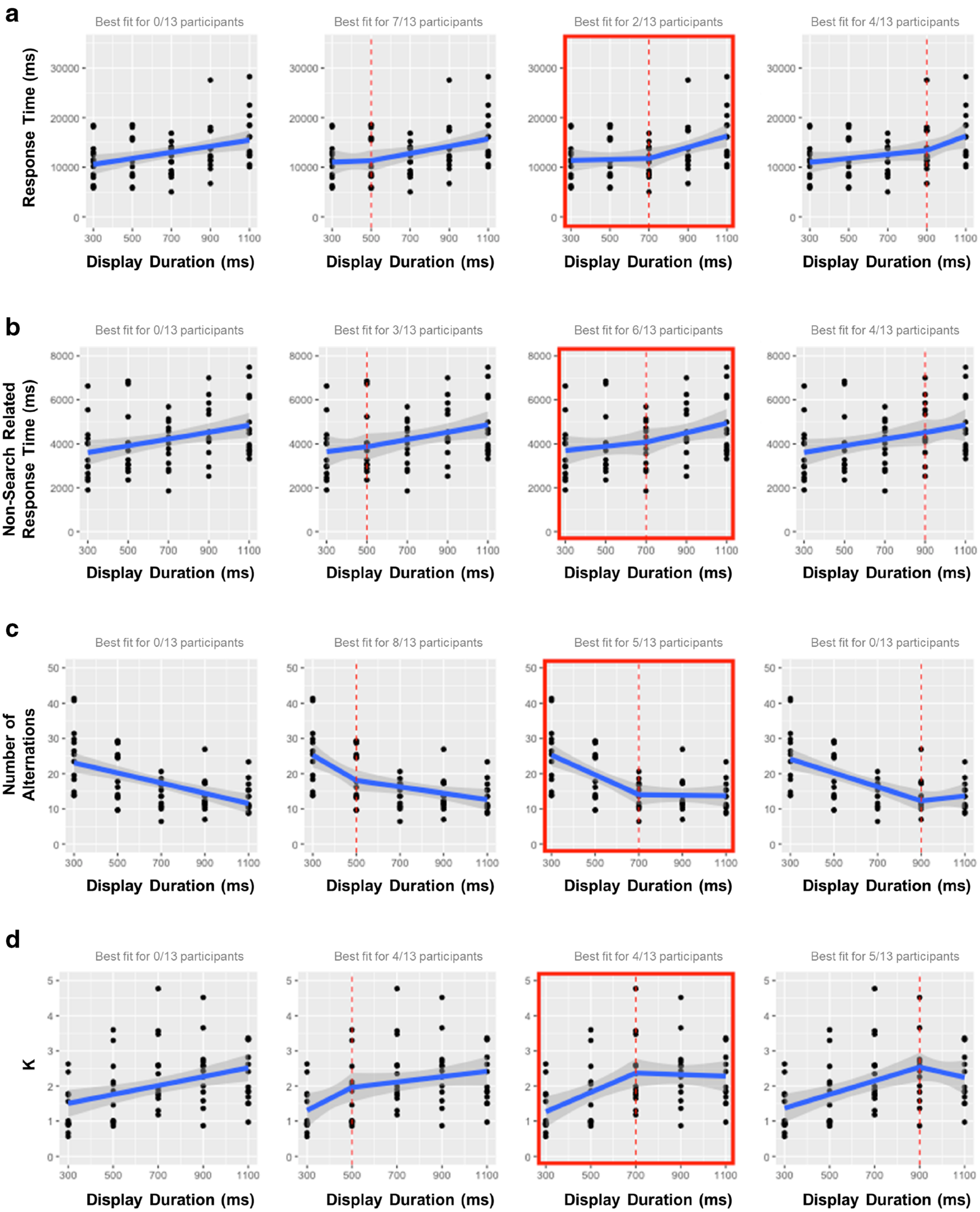

Fig. 6 Results of bi-linear spline regression models (no knot vs. knots placed at display durations of 500 ms, 700, ms, 900 ms) fit to the group averages observed in Experiment 1 for (**a**) response times, (**b**) non-search-related activity, (**c**) number of alternations required for successful change detection, (**d**) K estimates (efficient). The line of best fit for each model and its standard error are respectively illustrated by the blue line and gray surround. Dotted vertical lines mark the knot location. For each given measure, the graph of the best fitting model is bounded by a red outline. (Color figure online)

no change occurred, (c) empty the contents of VWM, (d) shift attention to three new items, and (e) load information from these items into VWM. That we found no change in performance with display times of 700 ms to 1,100 ms (and perhaps as high as 1,900 ms) suggests that 700 ms is enough time for most observers to complete all of these operations satisfactorily.

Experiment 3

Experiment 2 produced an asymptotic capacity estimate (between 2.4 and 4.6 items), similar to estimates derived from one-shot change detection tasks, despite substantial differences in the measures, stimuli, and methods. In Experiment 3, we explore effects of array size. A feature of flicker change detection (e.g., compared with one-shot) is its ability to present subjects with search arrays containing many more than the typical two, four, six, or eight objects used in one-shot tasks. However, concerns remain about the validity of estimates derived from the flicker task. With larger numbers of objects in each display, crowding might disrupt encoding. Larger displays also might require saccades during each display on-time to encode the stimuli. And, the use of a shorter blank-screen duration than is typical for one-shot tasks might have disrupted the consolidation of information in VWM (Vogel et al., 2006).

In Experiment 3, we explore the capacity of VWM with an ISI of 900 ms (the length typically used in one-shot experiments) and a display duration of 700 ms (the estimated knot point from Experiment 2) while examining whether the number of objects in the array affects estimates of capacity. In Experiment 3, we used array sizes of 4, 9, 16, 25, or 36 items. If Flicker change detection capacity is limited to around three objects per display on-time, then manipulating the number of items in the array should not affect the capacity estimate.

Method

The materials and procedures for Experiment 3 were identical to those of Experiment 2, except that (a) computers were iMac OSX machines running custom software written in MATLAB Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) with an LCD viewing area of 43.5 cm × 27 cm, (b) all trials were presented with a display duration of 700 ms and an ISI of 900 ms, and (c) observers completed five randomly ordered blocks of 20 trials each, one with each array set size (4, 9, 16, 25, or 36 items). The dots in each display were presented in a square grid, centered on the display, with constant interdot distances (2.66 degrees of visual angle) and diameter (1.6 degrees of visual angle). Observers were 13 Johns Hopkins undergraduates with normal or corrected-to-normal vision who received course credit for participating.

Results and discussion

All trials for which observers selected the wrong item (11.1% of trials) and all trials with response times more than two standard deviations above or below an observer's mean for that block (mean = 4.78% of trials, range: 3.00%–6.59% across blocks) were eliminated from further analyses. Across all array set sizes, observers detected changes after about 12.6 s (*SD* = 3.3 s) on average.

Given that the temporal parameters in Experiment 3 remained constant across all conditions (in contrast to Experiment 2), the number of alternations required to detect a change is redundant with response latency. Consequently, we only analyzed response latencies (see Fig. 7a, c).

Given that set size increased exponentially rather than linearly across blocks, we scaled response times using a square root transformation and conducted bilinear spline regressions on these scaled values (see Fig. 8a–c). The results of this analysis (see Fig. 8a) suggested that average scaled response times increased across displays containing 4–25 items (slope = 19.97 ms of response time / 1 scaled display item, $p < .001$), with a higher increase at the highest set size of 36 items (slope = 33.73 ms of response time / 1 scaled display item, $p < .001$). This pattern was observed in 7 of the 13 individuals.

Response times for the three fastest detections ranged from approximately 2,700 ms at Set Size 4 to 8,800 ms at Set Size 36 (see Fig. 7b). Bilinear spline regressions performed on scaled values of this measure (see Fig. 8b) demonstrated an increase across displays containing 4–25 items (slope = 7.36 ms of response time / 1 scaled display item, $p < .001$). These values experienced a further increase at Set Size 36 displays (slope = 17.35 ms of response time / 1 scaled display item, $p < .002$). This model best approximated the data of 6 of the 11 participants. This increase in RTs for the three fastest detections implies that observers take significantly longer to initiate storage of items as the total number of items in the display increases (e.g., they may spend some number of alternations gathering general info about the array, or they may not be as highly motivated to search larger arrays), and, perhaps, that they spend more flashes double-checking after they have successfully identified the target item (e.g., storing the target location for a mouse click may become more difficult as the number of items increases). It is also possible that search becomes less efficient with increasing set size and that subjects rarely find the change among the first items stored. Measuring the contributions of these subcomponents of active search and non-search-related RT will require future work to determine how search planning and execution vary with array size, number of trials, display times, ISI's, and so on.

A more complete understanding of efficiency in this task might benefit from an analysis of the RT distributions. For instance, Rensink (2000) relied on the proportion of detections occurring within epochs to estimate capacity, and future work

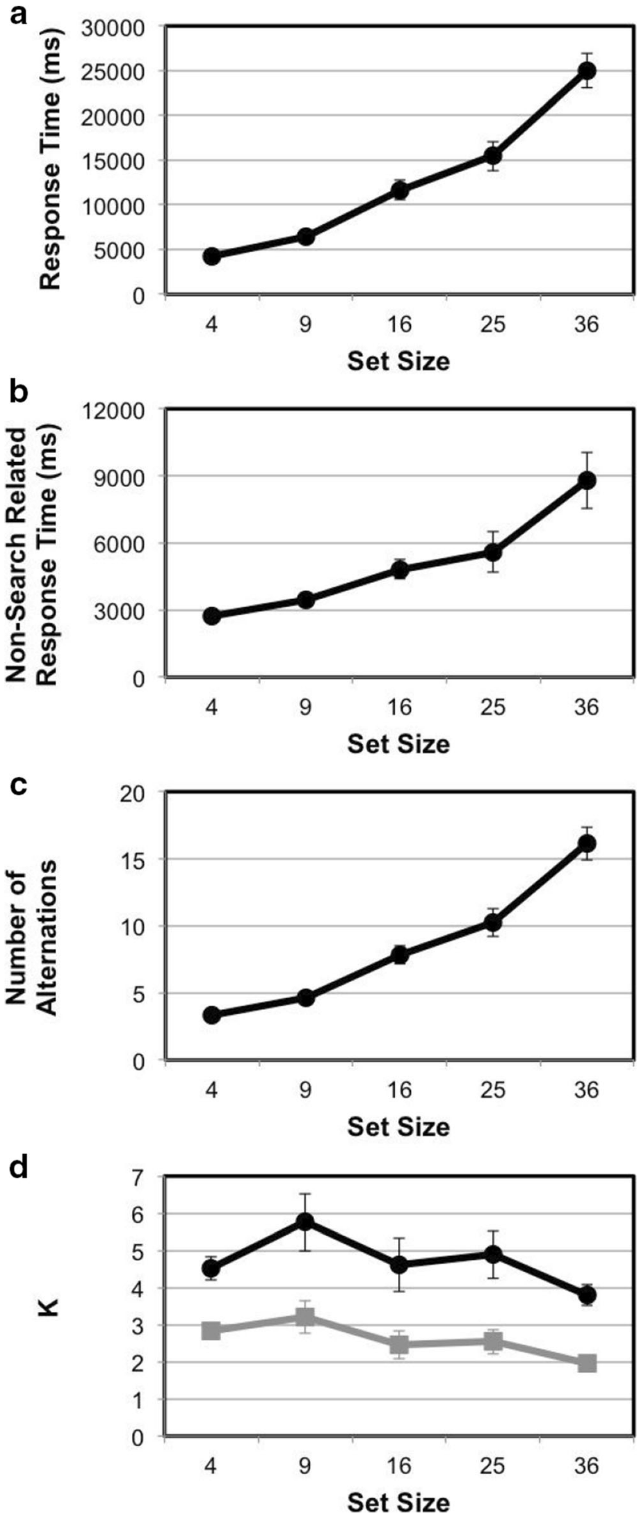

Fig. 7 Results of Experiment 2 (array size = varied, display duration = 700 ms, ISI = 120 ms). **a** Mean response time (RT) in ms (±*SE*) for the five set sizes. **b** Non-search-related activity as operationalized by the average of three fastest change detection RT's in ms (±*SE*). **c** Mean number of alternations required for successful change detection (±*SE*). **d** Estimated capacity assuming efficient (gray lines) and inefficient search (black lines) at each of the 5 set sizes (±*SE*)

could generalize that approach to full RT distributions. Such studies could also estimate non-search-related activity from the shape of early detection events. We estimated nonsearch activity using the three fastest RTs and we were able to calculate capacity even with increases in the time spent on non-search-related activities, but future work could refine our modeling using a more complete evaluation of RT distributions.

Bilinear spline regressions performed on scaled K values (see Fig. 8c) suggest that capacity remains relatively constant across displays containing four and nine items (slope = .04 items stored / 1 scaled display item, *p* = .71), but experience a slight decrease across higher set sizes (slope = −.11 items stored / 1 scaled display item, *p* = .01). This was the case for 4 of the 13 participants. This significant difference in capacity with increasing array size violates our expectation of no change, but the magnitude of the effect is still small enough to be explained by subtle, rather than dramatic, changes in the underlying processes (i.e., a decrease of approximately one item held in VWM as array size increases from four to 36 items). These differences could result from an adjustment of search strategy where participants stored fewer items at larger array sizes, either because search was more difficult or for fear of missing the target and having to search through the entire display once more. Another possibility is that the likelihood of revisiting an already visited item may increase with increasing array size, leading to the appearance of a difference in VWM capacity.

Overall, the flicker task provides fairly stable estimates of VWM capacity with array sizes of four to 36 items. The average raw capacity estimates across all set sizes was 2.6 items assuming efficient search (*SD* = 0.9) and 4.7 items (*SD* = 1.6) assuming inefficient search (see Fig. 7d), with estimated capacity decreasing slightly as a function of array set size.

Experiment 4

The results of Experiments 2 and 3 suggest that display durations of 700 ms and ISI's of 900 ms are sufficient for observers to fully load VWM to capacity and to compare stored representations with items on the screen. In Experiment 4, we examine whether these findings generalize from luminance changes to color changes. We chose color changes based on their widespread use in one-shot change detection studies (for review, see Brady, Konkle, & Alvarez, 2011).

In Experiment 4, observers completed a flicker task in which one colored square in an array alternated between two colors (see Fig. 9). ISI's were held constant at 900 ms, and display durations varied between 300ms and 1,100 ms by 200 ms intervals, similar to Experiment 2. If the temporal parameters identified in the previous experiments are suitable for a variety of feature dimensions, capacity estimates (K)

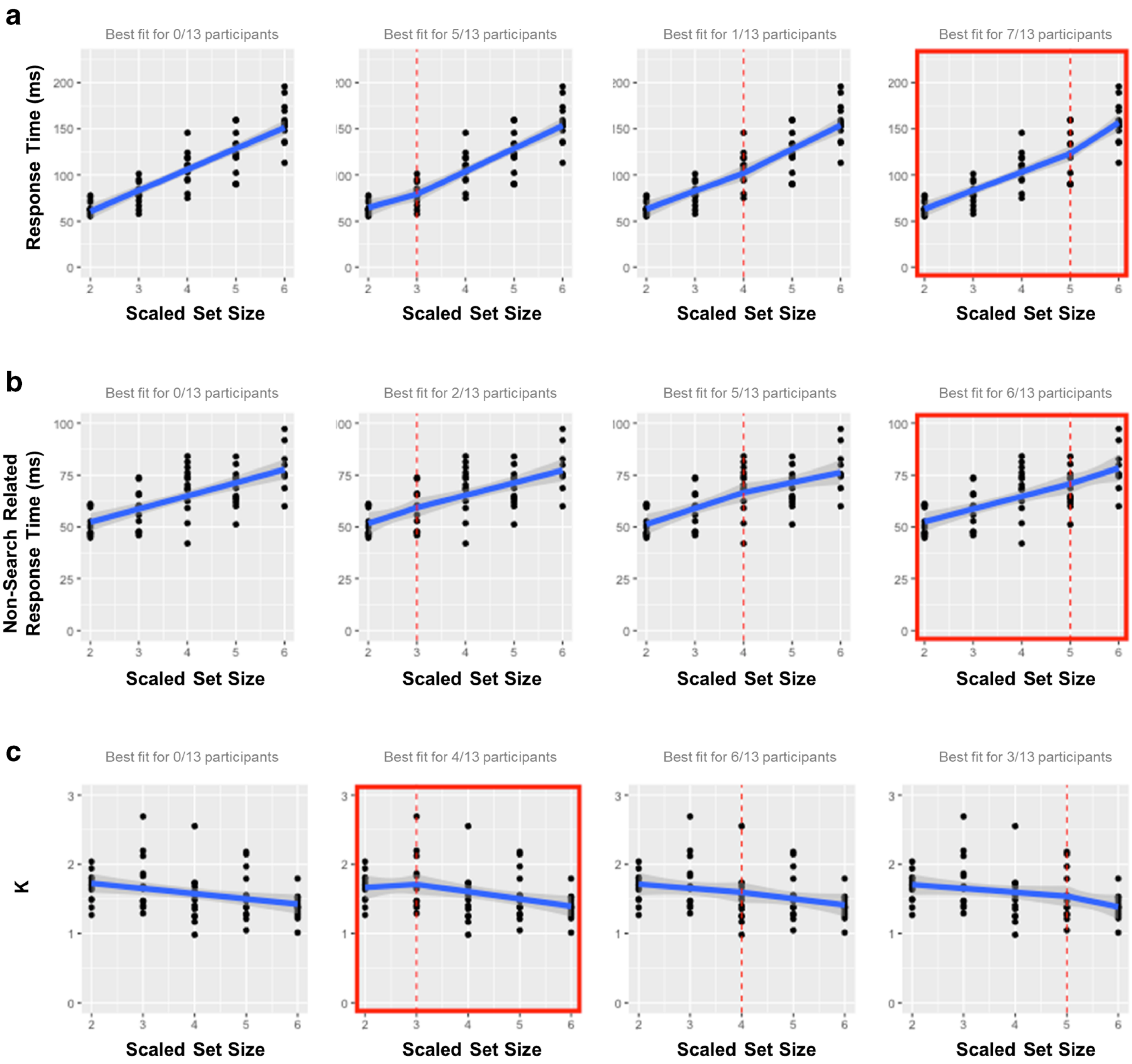

Fig. 8 Results of bilinear spline regression models (no knot vs. knots placed at scaled display set sizes of nine, 16, and 25 items) fit to the group averages observed in Experiment 2 for (**a**) response times, (**b**) Non-search-related activity, (**c**) K estimates (efficient). The line of best fit for each model and its standard error are respectively illustrated by the blue line and gray surround. Dotted vertical lines mark the knot location. For each given measure, the graph of the best fitting model is bounded by a red outline. (Color figure online)

calculated at various display durations in this color flicker task should asymptote at 700 ms of on-time.

Method

Observers

Thirteen undergraduate students from Harvard University with self-reported normal or corrected-to-normal vision participated in exchange for course credit. Data from two additional participants were excluded from analyses due to low target identification accuracy rates (<90%; cf. Pailian & Halberda, 2015).

Displays and procedure

Observers were tested on Macintosh iMac computers running OS10 with LCD monitors (viewable area: 43.5 cm × 27 cm). Viewing distance was unconstrained, but averaged approximately 60 cm. The experiment was programmed using MATLAB Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

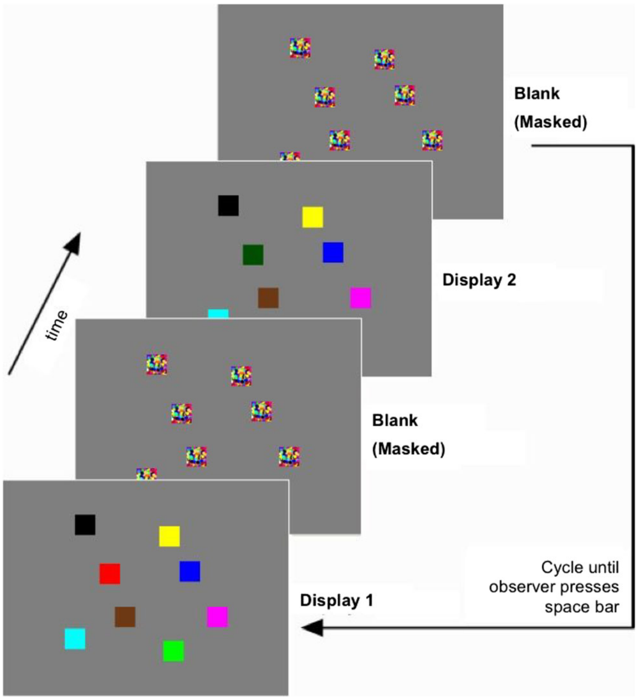

Fig. 9 The trial structure for flicker experiments involving color changes among colored squares. The cycle of the displays repeated until the observer pressed the space bar to indicate change detection. In this figure, the target is alternating between red and forest green. (Color figure online)

Except as noted, the methods were similar to those in previous experiments. Experiment 4 varied display durations, similar to Experiment 2. The durations of displays were constant within each trial, but display durations varied across five blocks of 20 trials (300 ms; 500 ms; 700 ms; 900 ms; or 1,100 ms), with block order randomized for each observer.

Stimuli consisted of eight colored squares (0.79° × 0.79°) that were randomly positioned within an invisible square (7.86° × 7.86°) that was located at the center of the screen. The colors of these squares were chosen randomly without replacement from a set of nine discrete colors (black, red, cyan, yellow, lime, blue, magenta, brown, forest green) and were presented against a homogeneous gray background.

At the beginning of each trial, two digits appeared at the center of the screen. Observers were instructed to rehearse these digits aloud throughout the trial as a way to interfere with the verbal encoding of color identities (Dixon & Shedden, 1993; Vogel et al., 2001). Verbal rehearsal was monitored by an experimenter who was seated behind the participant for the duration of the experiment.

After the digits, observers viewed the first display of colored squares, followed by a display in which the items were covered by colorful masks for 900 ms to eliminate afterimages (Coltheart, 1980). This mask display was then replaced by a second display that was identical to the first, except that one square had changed color. This color change was constrained, such that no one color could appear twice within the same display. After another 900-ms mask display, the sequence repeated until observers detected the change.

Once the change was detected, observers pressed the space bar to indicate they had noticed the change and subsequently made a mouse click to identify the perceived target. Once more, response times were recorded from the onset of the first display until the observer pressed the space bar, and click accuracy was recorded. Observers were instructed to respond as quickly but as accurately as possible.

Results and discussion

All trials for which observers selected the wrong item (3.62% of trials) and all trials with response times more than two standard deviations above an observer's mean for that block (mean = 3.92% of trials, range: 2.00%–6.32% across blocks) were eliminated from further analyses. Group results of all measures of interest are illustrated in Fig. 10a–d.

Across all display durations, observers detected changes after about 5.16 s (*SD* = 0.66 s) on average. Bilinear spline regressions (see Fig. 11a) revealed that average response times remained relatively stable across display durations of 300 ms and 500 ms (slope = −2.80 s / 1 s of display time, *p* = .08). With a larger sample, this small difference might be statistically significant, so the nonsignificant difference from a zero slope should be interpreted with caution. Response times increased linearly thereafter, from 500 ms to 1,100 ms (slope = 2.27 s / 1 s of display time, *p* < .001). This pattern was observed for 7 of the 13 individuals. A similar analysis performed on the number of alternations required (see Fig. 11c) suggested that these values decrease across display durations of 300–500 ms (slope = −5.0 alternations / 1 s of display time, *p* < .001), but remain steady thereafter (slope < −0.5 alternations / 1 s of display time, *p* = .14). This model best fit the data of 7 of the 13 individuals. As such, more exposures were required to find the target at the shortest duration. These results replicated the basic pattern seen in Experiment 2.

Differences in response time and number of alternations between display durations of 300 ms and longer may reflect insufficient exposure to adequately encode stimuli. Indeed, bilinear regression analyses performed on response times for the three fastest detections (see Fig. 11b) demonstrate that these values remained relatively constant across shorter display durations of 300 ms and 500 ms (slope = 0.31 s / 1 s of display time, *p* =.73), but increased from 500 ms onwards (slope = 1.44 s / 1 s of display time, *p* < .001). This pattern was observed for 3 of the 13 participants.

We used Equations 4a and 4b to quantify individual differences in VWM storage capacity (K) across each of the display durations (see Fig. 10d). Bilinear spline regressions performed on averaged K values (see Fig. 11d) revealed that estimates of storage capacity increased across durations of 300–700 ms (slope = 3.44 items stored / 1 s of display time, *p* < .001),

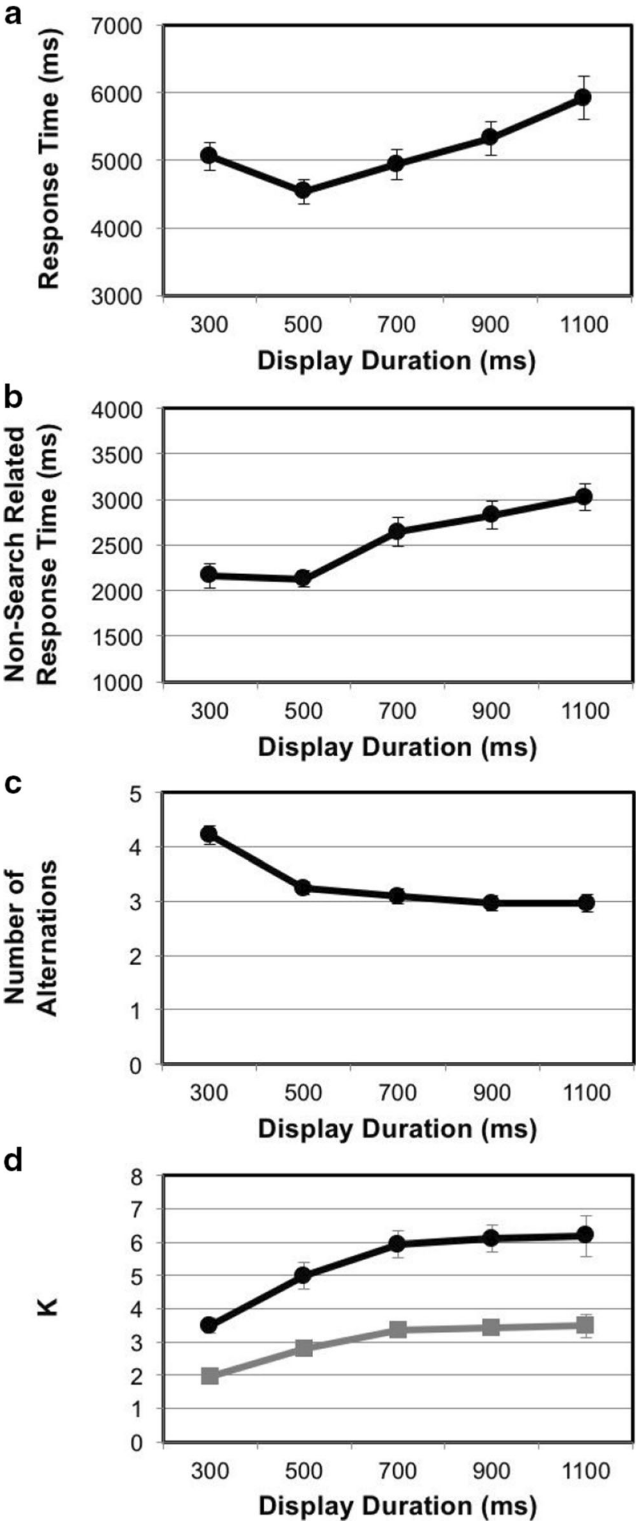

Fig. 10 Results of Experiment 3 (array size = 8, display duration = varied, ISI = 900 ms). **a** Mean response time (RT) in ms (±SE) for the five display durations. **b** Non-search-related activity as operationalized by the average of three fastest change detection RT's in ms (±*SE*). **c** Mean number of alternations required for successful change detection (±*SE*). **d** Estimated capacity assuming efficient (gray lines) and inefficient search (black lines) at each of the 5 display durations (±*SE*)

but remained constant thereafter (slope = 0.23 items stored / 1 s of display time, $p = .78$). This model best approximated the data of 5 of the 11 participants. These results are consistent with those observed in Experiment 2 with luminance changes among grayscale dots. Namely, display durations of 700 ms on-time and 900 ms ISI's were sufficient for observers to fully load VWM with color information and make subsequent comparisons. Moreover, the average of K estimates calculated at asymptotic display durations (700–1,100 ms) observed here ($K_{\text{efficient}}$ = 3.42 items, $K_{\text{inefficient}}$ = 6.07 items) are roughly comparable to those observed in Experiments 2 and 3 (they might be slightly larger because storing categorical colors can be easier than storing items that subtly differ in grayscale; Alvarez & Cavanagh, 2004).

General discussion

The capacity estimate for VWM based on the flicker task was approximately three items, both for visually simple, hard to categorize stimuli (Experiments 1–3) and for colored squares commonly used in one-shot tasks (Experiment 4). This estimate is in line with that derived from the more typical one-shot change detection task (Experiment 1), and the capacity estimates for individual participants using these two methods are positively correlated (Experiment 1). Perhaps surprisingly, the flicker task appeared to provide a more internally reliable measure of capacity than did the one-shot task (see also Pailian & Halberda, 2015).

The flicker task yielded remarkably consistent capacity estimates for display on times ranging from 700 ms to 1,100 ms, ISI's of 120 ms and 900 ms, and set sizes ranging from four to 36 items. We found performance to be process limited (i.e., by response time) with display durations shorter than 700 ms and memory limited (i.e., by number of exposures to the change) for display durations longer than 700 ms (see also Rensink, 2000). Consequently, total response time and the number of changes viewed are distinct measures of change detection performance that agree only when just enough (and no more) on-time is provided (e.g., 700 ms) to allow comparison and loading of new items into memory. Most studies in the change blindness literature treat these measures as equivalent, but they might differentially tap the operation of VWM.

Perhaps most importantly, unlike the one-shot task, the flicker task allows us to assess the capacity of VWM with as many as 36 items in the array (in the present study) over the course of prolonged searching. Both larger arrays and more extended viewing might better tap the ways in which VWM is used in daily life. Thus, the flicker task can complement existing work using the one-shot task, and our results suggest that the three-item limit of VWM is robust to variations in the number of items in a scene, types of items, total viewing time for a scene, ISI duration, and display duration.

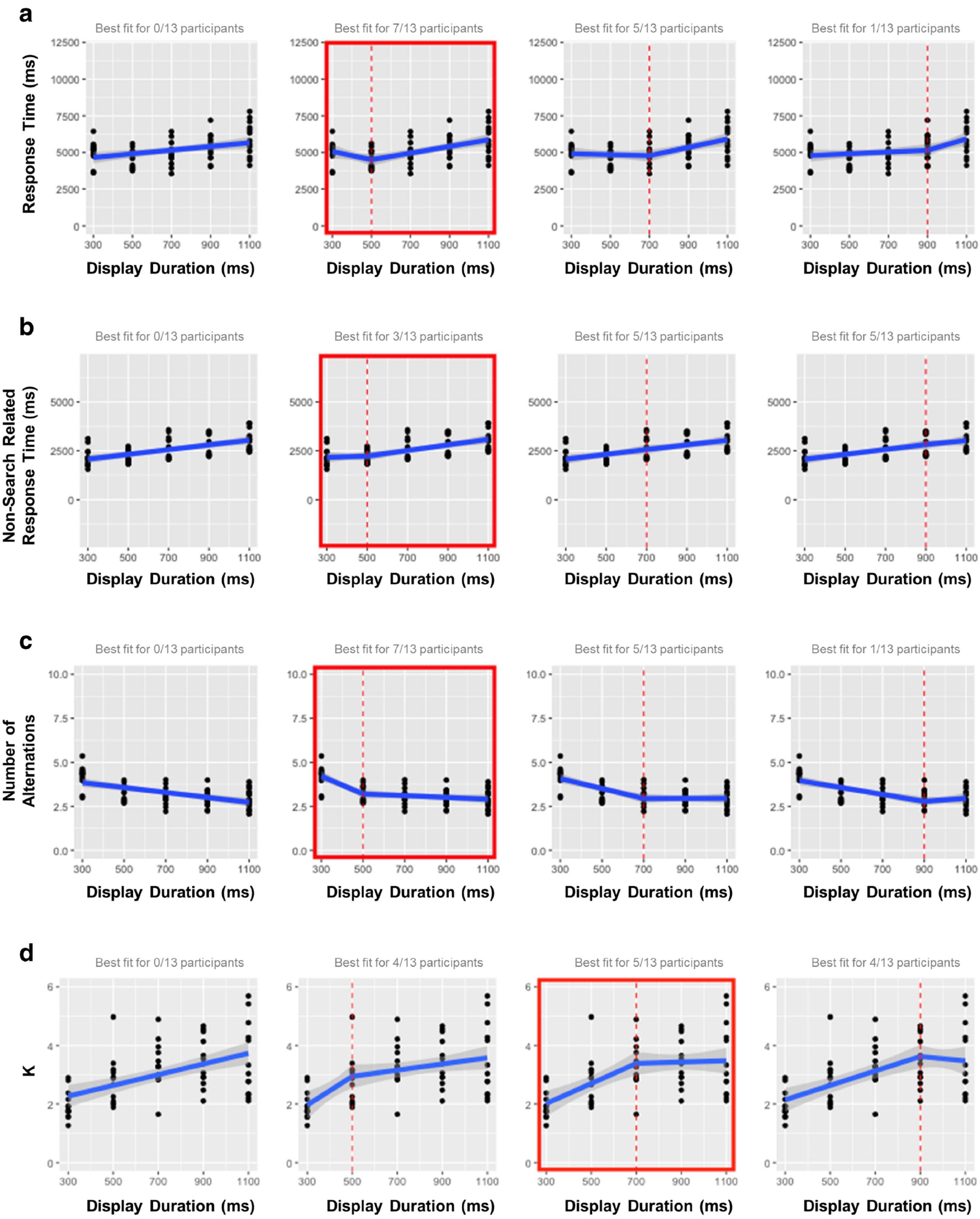

Fig. 11 Results of bilinear spline regression models (no knot vs. knots placed at display durations of 500 ms, 700, ms, 900 ms) fit to the group averages observed in Experiment 3 for (**a**) response times, (**b**) non-search-related activity, (**c**) number of alternations required for successful change detection, (**d**) K estimates (efficient). The line of best fit for each model and its standard error are respectively illustrated by the blue line and gray surround. Dotted vertical lines mark the knot location. For each given measure, the graph of the best fitting model is bounded by a red outline. (Color figure online)

Despite superior reliability and roughly comparable overall capacity estimates, the flicker task does have drawbacks as a measure of VWM, compared with the one-shot task: (1) The duration of non-search-related processes can vary across observers and array types, (2) the time required to fill VWM to capacity is not constant across individuals (the model that best fit the group data did not produce the best fit for each individual subject), and (3) it is difficult to determine how often people reexamine previously checked items (i.e., the extent to which search occurs with replacement). Of course, individual differences in these processes are potentially interesting and measurable in their own right, and the flicker task allows for such exploration in future studies.

To address these concerns, we report a range of estimates intended to be interpreted within the specific context and parameters used in this task. More broadly, we argue for a converging estimate of the underlying construct of working memory capacity based on multiple tasks. Most studies of VWM have relied exclusively on the one-shot task to measure capacity, but no one task is likely to uniquely and exclusively measure an underlying latent construct. That is, performance on any individual task may provide a measure of the underlying construct, but it also includes task-specific factors that presumably are not related to that construct. By using multiple tasks that each presumably require VWM but vary in their task-specific components, we can better estimate the capacity of working memory as an underlying construct.

As VWM plays a crucial role in bridging between the present and the recent past for visual processing, understanding how VWM unfolds over time may be crucial for understanding its limitations and functioning. We suggest that the flicker task may be a useful assay of this active and unfolding functioning of VWM

Constraints on generality

Some previous work used the flicker task to measure developmental change in VWM capacity (Pailian, Libertus, Feigenson & Halberda, 2016), but relatively few studies have tested whether estimates of capacity are reliable in populations other than undergraduate students at highly selective universities. Given previous work linking performance on working memory tasks to levels of education, socioeconomic status, and age (Wilmer et al., 2012), the capacity estimates observed here may not be representative of all groups of individuals. Capacity might also vary with the types of items, item heterogeneity, change salience, item complexity, and item arrangement, and further work is needed to determine whether capacity estimates are stable across such variability.

Open practices statement The data and materials for all experiments are not currently available online, and no experiments were preregistered.

Author contributions J.H., D.J.S., H.P., and J.W. developed the ideas and studies; H.P., J.H., and J.W. analyzed data; H.P., J.H., and D.J.S. wrote the paper. Experiment 1 was coded and conducted by H.P. in J.H.'s laboratory at Johns Hopkins University (2014–2015). Experiments 2a–b were coded by J.W .and conducted by J.W. and J.H. in D.J.S.'s laboratory at Harvard University (2001–2002). Experiment 3 was coded by paid research assistants and conducted in J.H.'s laboratory at the Johns Hopkins University (2009). Experiment 4 was coded and conducted by H.P. in Dr. George Alvarez's laboratory at Harvard University (2015–2016).

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.

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences, 108*(25), 10367–10371.

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

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

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851–854.

Bays, P. M., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*(10), 1–11.

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

Blackmore, S. J., Brelstaff, G., Nelson, K., & Troscianko, T. (1995). Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. *Perception (London)*, 24, 1075-1075.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and towards structured representations. *Journal of Vision, 11*(5), 4, 1–34.

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

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

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

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

Dixon, P., & Shedden, J. M. (1993). On the nature of the span of apprehension. *Psychological Research, 55*(1), 29–39.

French, R. S. (1953). The discrimination of dot patterns as a function of number and average separation of dots. *Journal of Experimental Psychology, 46,* 1–9.

Fukuda, K., & Vogel, E. K. (2011). Individual differences in recovery time from attentional capture. *Psychological Science, 22*(3), 361–368.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*(6693), 575–577.

Horowitz, T., & Wolfe, J. (2003). Memory for rejected distractors in visual search?. *Visual Cognition, 10*(3), 257–298.

Johnson, N. L., & Kotz, S. (1977). *Urn models and their application: An approach to modern discrete probability theory* (Vol. 77). New York, NY: Wiley.

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*, 1476–1486.

Lleras, A., Rensink, R. A., & Enns, J. T. (2005). Rapid resumption of interrupted visual search: New insights on the interaction between memory and vision. *Psychological Science, 16*, 684–688.

Lleras, A., Rensink, R. A., & Enns, J. T. (2007). Consequences of display changes during interrupted visual search: Rapid resumption is target specific. *Perception & Psychophysics*, *69*(6), 980–993.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390(6657), 279–281.

Matsuyoshi, D., Osaka, M., & Osaka, N. (2014). Age and individual differences in visual working memory deficit induced by overload. *Frontiers in Psychology, 5,* 1–7.

Nunnally, J. C., Jr. (1970). Introduction to psychological measurement. New York, NY: McGraw-Hill.

Pailian, H., & Halberda, J. (2015). The reliability and internal consistency of one-shot and flicker change detection for measuring individual differences in visual working memory capacity. *Memory & Cognition, 43*(3), 397–420.

Pailian, H., Libertus, M. E., Feigenson, L., & Halberda, J. (2016). Visual working memory capacity increases between ages 3 and 8 years, controlling for gains in attention, perception, and executive control. *Attention, Perception, & Psychophysics, 78*(6), 1556–1573.

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

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

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

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179–197.

Rensink, R. A. (2000). Visual search for change: A probe into the nature of attentional processing. *Visual Cognition, 7,* 345–376.

Rensink, R. A. (2014). Limits to the usability of iconic memory. *Frontiers in Psychology, 5,* 971.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8,* 368–373.

Simons, D. J. (1996). In sight, out of mind: When object representations fail. *Psychological Science, 7*(5), 301–305.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*(11), 1–29. https://doi.org/10.1037/h0093759

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures of individual differences in controlling access to working memory. *Nature, 438*(24), 500–503.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 92–114.

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,* 1436–1451.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274–290.

Wilmer, J. B., Germine, L., Ly, R., Hartshorne, J.K., Kwok, H., Pailian, H., … Halberda, J. (2012, May 11–16). *The heritability and specificity of change detection ability.* Talk presented at VSS, the Vision Sciences Society, Naples, Florida.

Woodman, G. F., & Vogel, E. K. (2005). Fractionating working memory: Consolidation and maintenance are independent processes. *Psychological Science, 16*(2), 106–113.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 11, 1120–1135.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235.