

# The reliability and internal consistency of one-shot and flicker change detection for measuring individual differences in visual working memory capacity

Hrag Pailian · Justin Halberda

Published online: 13 December 2014  
© Psychonomic Society, Inc. 2014

**Abstract** We investigated the psychometric properties of the one-shot change detection task for estimating visual working memory (VWM) storage capacity—and also introduced and tested an alternative flicker change detection task for estimating these limits. In three experiments, we found that the one-shot whole-display task returns estimates of VWM storage capacity ( $K$ ) that are unreliable across set sizes—suggesting that the whole-display task is measuring different things at different set sizes. In two additional experiments, we found that the one-shot single-probe variant shows improvements in the reliability and consistency of  $K$  estimates. In another additional experiment, we found that a one-shot whole-display-with-click task (requiring target localization) also showed improvements in reliability and consistency. The latter results suggest that the one-shot task can return reliable and consistent estimates of VWM storage capacity ( $K$ ), and they highlight the possibility that the requirement to localize the changed target is what engenders this enhancement. Through a final series of four experiments, we introduced and tested an alternative flicker change detection method that also requires the observer to localize the changing target and that generates, from response times, an estimate of VWM storage capacity ( $K$ ). We found that estimates of  $K$  from the flicker task correlated with estimates from the traditional one-shot task and also had high reliability and consistency. We highlight the flicker method’s ability to estimate executive functions as well as VWM storage capacity, and discuss the potential for measuring multiple abilities with the one-shot and flicker tasks.

**Keywords** Working memory · Individual differences · Memory

H. Pailian (✉) · J. Halberda  
Department of Psychological and Brain Sciences, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA  
e-mail: pailian@jhu.edu

In recent decades, research on individual differences in working memory capacity has enjoyed broad appeal and made a significant impact throughout psychology, in part because differences in working memory (WM) are argued to be a locus for broader impacts throughout cognition (Vogel & Awh, 2008). Here, we focus on evaluating the suitability of paradigms intended to measure the storage capacity of visual WM.

Baddeley and Hitch’s (Baddeley, 2000; Baddeley & Hitch, 1974) model of working memory differs from previous theories relating to short-term memory storage by virtue of its having multiple storage and processing constituents. Under this framework, information is stored separately within modality-specific modules, as opposed to within a single multidimensional workspace. These modality-specific sub-components include the phonological loop, which maintains verbal and auditory information, and the visuospatial sketchpad, which maintains visual and spatial information. On the basis of their functions as storage repositories, they are sometimes referred to as *verbal short-term memory* and *visual short-term memory*, respectively. Here, we refer to them as *verbal WM storage* and *visual WM storage*.

The Baddeley and Hitch model also includes two processing units that preside over verbal and visual storage, called the *central executive* and the *episodic buffer*. These units constitute the “working” elements of WM, in that, for instance, the central executive mediates between the sensory modules by directing the input and output of information to and from the verbal and visual stores, as well as governing the manipulation of stored information and the division and direction of attention to tasks. Although the exact functions of the episodic buffer remain to be conclusively defined, it is typically discussed as an integrator of multidimensional information into a single representation, and is believed to connect WM to perception and long-term memory in order to support abstract thought.

Though the proposal of two processing components and two storage modules has played an organizing role for much of the research on WM and attention throughout recent decades (Baddeley, 2012), an emphasis has also been placed on explicating the limits constraining the verbal and visual WM stores. Both verbal and visual WM storage are believed to be capacity-limited, in that only a certain amount of information can be maintained within them at any one time (Baddeley & Hitch, 1974; Miller, 1956). Factors that have been suggested to influence these apparent limitations include (but are not limited to) temporal delay (Peterson & Peterson, 1959), articulatory rehearsal processes (Baddeley, 1986; Baddeley, Thomson, & Buchanan, 1975), attentional focus (Cowan, 2001), memory set size (Luck & Vogel, 1997; Pashler, 1988; Pylyshyn & Storm, 1988), and item complexity (Alvarez & Cavanagh, 2004; Bays & Husain, 2008; Wilken & Ma, 2004). As research studies have identified possible limits in WM, new work has sought to evaluate the stability of individual differences in these limits.

A classic issue for task construction in psychology research involves the challenge of assessing the reliability and consistency of our measures. For example, paradigms aiming to estimate the storage limits of verbal and visual WM should produce estimates that reflect a limit in storage capacity (i.e., memory), as opposed to other factors (i.e., encoding strategies). This goal is admittedly difficult to achieve, because it is nearly impossible to design a paradigm that isolates storage functions separately from executive processes (Conway, Cowan, Bunting, Theriault, & Minkoff, 2002). Most tasks require processes like switching attention, dividing attention, strategically encoding information, and so on. It may be that measuring these other abilities, either within or across tasks, will empower improved purchase on storage limits in WM. Likewise, convergent measures from a wide variety of tasks, as opposed to just one “gold-standard” task, may also be valuable.

To the extent that we believe that visual and verbal WM storage limits reflect invariant and enduring limitations of these modality-specific stores (e.g., expecting to observe a single, unchanging visual WM storage capacity limit that operates throughout changes in visual clutter in a scene, array sizes, display times, etc.), the estimates of these limits should remain constant across such changes; for example, visual WM storage capacity estimates should remain equivalent across changes in array size. Recent results have suggested that some observers (specifically, observers who have low overall visual WM storage capacities) will tend to store different amounts of information, dependent on the array structure (e.g., low-capacity observers may store less information as array size increases; Linke, Vicente-Grabovetsky, Mitchell, & Cusack, 2011; Matsuyoshi, Osaka, & Osaka, 2014). In this way, variability in storage capacity might be a feature of certain memory systems rather than a flaw in measurement. But, even in

such cases, we might expect to find a reasonable correlation of capacity estimates across changes to an array (e.g., across set sizes): For instance, observers with lower capacities at set size 4 may show a further reduced capacity at set size 8. In general, tasks deemed to be ideal assayers of verbal and visual WM storage capacity limits are expected to produce estimates that are reliable across a variety of contexts. Such attributes are operationalized by measures of *reliability*, or the degree to which multiple subpartitionings of a dataset produce estimates that agree with one another, and *internal consistency*, or the degree to which a measure is able to consistently produce similar estimates across periods of time. Failure to attain these goals can prove problematic for a method that aims to support theorizing and generalization, because it calls into question the efficacy and, potentially, the construct validity of the implemented measure (e.g., consider arguments surrounding IQ and “g,” or generalized intelligence). Estimates fluctuating across periods of testing prove particularly detrimental to the individual-differences approach, since a lack of internal consistency constrains the ability to detect meaningful relationships. In such cases, it would also be unclear whether the variance captured by a correlation was attributable to individual differences or rather was simply due to the variability of the measure itself. With such concerns in mind, we turned to evaluating the psychometric properties of tasks designed to measure visual working memory (VWM) storage capacity.

The *one-shot change detection task*, developed by Phillips (1974) and popularized by Luck and Vogel (1997), stands as the primary investigative tool for quantifying limits in VWM storage capacity. This paradigm has numerous advantages over other measures of VWM storage, such as the Corsi block and pattern span tasks, in that it allows for storage functions to be studied over a single, “one-shot” viewing event—where an observer has a single opportunity to view a scene and remember all that he or she can from this single viewing. Many have lauded this “one-shot” approach because it is thought to minimize the contributions of executive processes and the deployment of complex strategies.

In the one-shot task, observers are briefly presented with a memory array that consists of a set of simple stimuli. Set size varies across trials (e.g., one to eight colored squares). After this brief initial encoding period, observers must temporarily hold these items in memory over a blank consolidation interval (e.g., a blank screen lasting 900 ms). Investigations into the durations of these displays suggest that 100 ms is sufficient for the successful encoding of the memory stimuli (Vogel, Woodman, & Luck, 2001), whereas 50 ms/object is appropriate for their subsequent consolidation (Vogel, Woodman, & Luck, 2006). The offset of the blank consolidation display is followed by a target array. In “whole-display” versions of this task, the target array is either identical to the memory array or differs on the basis of a single feature. Alternatively, in “single-probe” versions of this task, only one item reappears at a

previously occupied location, and this item is either identical to the one that had appeared at that location in the memory array or has changed to a new color. In either case, observers are subsequently asked to make a two-alternative forced choice, indicating whether the two arrays/probed items were identical or differed in some respect. Successful completion of this task requires observers to compare the representations of items stored in memory (subject to a VWM storage limit) with the items that are currently attended and available for free viewing. As such, resultant performance accuracies should be, it is reasoned, limited by the maximum number of individual items or amount of visual information that can be stored in VWM. Performance accuracy in this task is used to produce an estimate of VWM storage capacity, “ $K$ ,” according to the models that we will discuss in detail in the experiment sections below.  $K$  values are typically calculated for each set size using Pashler’s (1988) equation, for whole displays, or Cowan’s (2001) equation, for single-probe displays, and are then averaged to compute an overall estimate of an individual’s VWM storage capacity limit.

The one-shot change detection task has been central to the VWM storage literature. For instance, it has stood as the metric of record for debates concerning the units of VWM storage (Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007; Luck & Vogel, 1997). Its  $K$  estimates have been shown to correlate with a variety of cognitive indices, which include the ability to disengage from attentionally capturing stimuli (Fukuda & Vogel, 2011), the magnitude of slowing caused by contingently capturing distractors (B. A. Anderson, Laurent, & Yantis, 2011), the breadth of attention (Gmeindl, Jefferies, & Yantis, 2011), various span tasks (Cowan et al., 2005; Cowan, Fristoe, Elliott, Brunner, & Sauls, 2006), measures of intelligence (Fukuda, Vogel, Mayr, & Awh, 2010), and the ability to filter information (Vogel, McCollough, & Machizawa, 2005). This method has further been used to establish neurophysiological measures of VWM storage capacity (Diamantopoulou, Poom, Klaver, & Talsma, 2011; Vogel & Machizawa, 2004), pinpoint the neural loci of VWM storage (Todd & Marois, 2004; Xu & Chun, 2006), and describe the developmental trajectory of its capacity limits (Cowan, AuBuchon, Gilchrist, Ricker, & Sauls, 2011; Cowan, Morey, AuBuchon, Zwillig, & Gilchrist, 2010; Riggs, McTaggart, Simpson, & Freeman, 2006; Simmering, 2012). In short, it would be difficult to overstate the importance of the one-shot change detection paradigm to the literature on VWM storage over the past 20–30 years.

Despite its ubiquitous use, few attempts have been made to investigate the psychometric properties of the one-shot change detection task (but see Kyllingsbæk & Bundesen, 2009; Makovski, Watson, Koutstaal, & Jiang, 2010). Here, we measured the levels of reliability and internal consistency for this paradigm’s estimates of VWM storage capacity,  $K$ . More specifically, we ran this task at a variety of set sizes and

numbers of trials, and looked at the degree to which these estimates agreed with one another. We indexed the reliability of these estimates by correlating the  $K$  values calculated using different set sizes, and we indexed their internal consistency by performing random split-half correlation analyses on averaged  $K$  estimates and maximum  $K$  estimates. On the basis of the notion that an individual’s VWM storage capacity is invariant, or at least correlated across contexts, we expected that the  $K$  values calculated at different set sizes would correlate with one another, and that what little variance remained would be attributed to random fluctuations. That is, if this estimate truly represents a storage limit that resides in VWM, irrespective of array size, individual differences in  $K$  should not change with the number of items tested. To presage our results, we were surprised to find that correlations of  $K$  estimates across set sizes varied on the basis of the number of trials used and whether the test arrays were presented as whole displays (i.e., the set size of the target array was equivalent to the set size of memory array) or single probes (i.e., only one item was presented). In response to the psychometric shortcomings that we discovered for the one-shot whole-display change detection task, we present a series of experiments demonstrating the psychometric qualities of variations of the one-shot task and of an alternative, *flicker change detection task* that appears to produce a reliable and consistent measure of individual differences in VWM storage capacity.

### Experiment 1a

In Experiment 1a, we investigated the psychometric properties of the whole-display one-shot change detection task, using parameters commonly used throughout the literature. For example, the timing parameters in our study are based on those implemented in previously established experiments (Hyun, Woodman, Vogel, Hollingworth, & Luck, 2009; Luck & Vogel, 1997; Schmidt, Vogel, Woodman, & Luck, 2002; Vogel, Woodman, & Luck, 2001), as are the stimulus properties, such as color, shape, and size. To reflect the variety of the numbers of memory items used throughout the literature, we used set sizes that were under capacity (i.e., two squares), near capacity (i.e., four squares), and over capacity (i.e., eight squares).

Observers saw 240 trials in which a memory array of colored squares was presented, followed by a blank consolidation screen and a test array. Observers had to respond with whether or not they thought that one of the memory squares had changed color.

We calculated estimates of capacity for each set size (i.e., two, four, and eight squares) using Pashler’s (1988) equation. By measuring the degree to which these  $K$  estimates correlated across set sizes, we assessed the reliability with which the one-shot change detection task produces estimates of VWM

storage capacity that are constant or correlated across array sizes. We were further able to determine the internal consistency of this paradigm by performing random-halves reliability analyses over the averaged  $K$  values and maximum  $K$  values. If the one-shot change detection paradigm is a suitable estimator of VWM storage capacity, the  $K$  estimates obtained from each of the set sizes would be expected to correlate with one another (reliability) and have similar scores across trials (internal consistency).

## Method

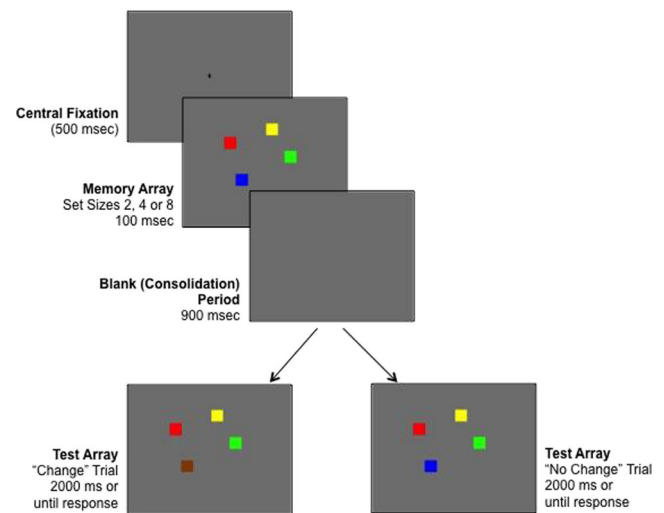
**Observers** Fourteen Johns Hopkins University students with normal or corrected-to-normal vision participated in our study in exchange for course credit.

**Equipment** Observers were tested in a dimly lit room, using a Macintosh iMac computer with an LCD viewing area of  $43.5 \times 27$  cm. The viewing distance was unconstrained, but averaged 60 cm. The experiment was programmed using the Psychophysics Toolbox (Pelli, 1997) and was displayed using MATLAB software.

**Design and procedure** Observers completed 240 trials of a typical whole-display one-shot change detection task, in which they were presented with colored squares (visual angle =  $0.79^\circ \times 0.79^\circ$ ) that were randomly positioned on a homogeneous gray background. Each square was randomly assigned one of ten discrete colors—black, white, red, cyan, yellow, green, blue, orange, brown, or purple. The colors of the squares were constrained, such that no one color could appear twice in the same display.

The beginning of each trial was marked by a central fixation cross that remained on the screen for 500 ms. Observers were subsequently presented with a 100-ms memory array that contained two, four, or eight colored squares, which was followed by a 900-ms blank retention interval. A target array was then displayed for a maximum of 2,000 ms or until a response was made. The stimuli in the target array were identical to those displayed in the memory array, with the exception that, on half of all trials, one square changed color (Fig. 1). Observers were instructed to make a keyboard press to indicate whether the memory and target arrays were the same or different. Performance accuracies were recorded and used to calculate individual differences in VWM storage capacity.

**Calculation of VWM storage capacity,  $K$**  Theoretically, performance accuracy in one-shot change detection is determined by a ratio of the maximum number of individual items stored in VWM,  $K$ , to the total number of items presented,  $N$ , up to a maximum of 1 (Eq. 1a). That is, if an observer's capacity,  $K$ , is equal to or greater than the total number of items in the



**Fig. 1** Schematic of a typical whole-display one-shot change detection trial, used throughout Experiments 1a–1c

display,  $N$ , then the observer will always be correct (proportion correct = 1); if  $K$  is one-half of  $N$ , the proportion correct would equal .5; and so forth:

$$\text{Proportion Correct} = K/N. \quad (1a)$$

This simple equation (1a) can be modified to reflect a biased observer, à la signal detection theory (Eq. 1b):

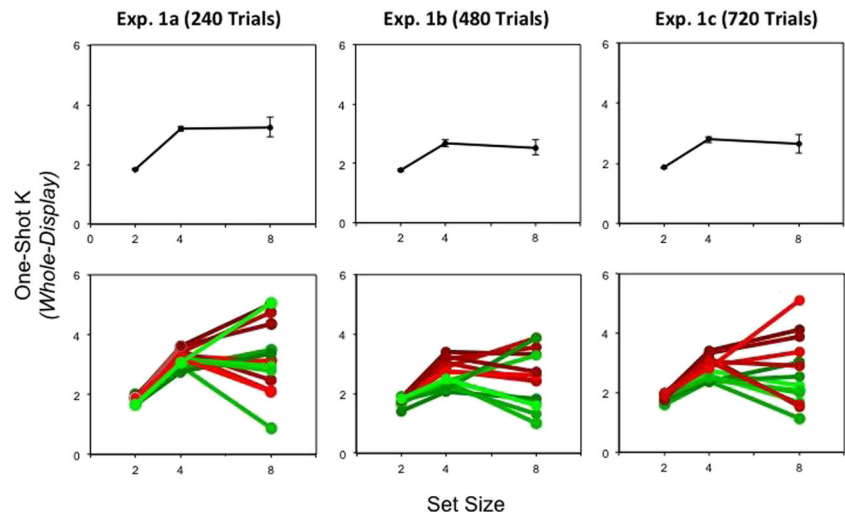
$$K = N(H - F) / \text{div}(1 - F). \quad (1b)$$

Rooted in signal detection theory, Pashler's (1988) Eq. 1b operationalizes performance accuracy as the ratio of corrected hit [ $H = \text{hits}/(\text{hits} + \text{misses})$ ] to false alarm [ $F = \text{false alarms}/(\text{false alarms} + \text{correct rejections})$ ] rates (see Rouder, Morey, Morey, & Cowan, 2011, for a thorough explanation of how response rates are corrected for informed guessing). Throughout all experiments in this series, estimates of VWM storage capacity for the one-shot task were initially calculated for each set size using Eq. 1b. We relied on these estimates to compare across set sizes, and then, as is typical in the literature, we averaged these values across set sizes to produce overall estimates of individual VWM storage capacity.

## Results

The estimated  $K$  values for each set size are illustrated in Fig. 2. We observed a significant effect of set size,  $F(2, 26) = 18.19$ ,  $p = .001$ ,  $\eta_p^2 = .58$ . Post-hoc contrasts revealed no difference between the capacity estimates for set sizes 4 and 8,  $F(1, 26) = 0.03$ ,  $p > .05$ , whereas a significant difference was observed between capacity estimates for set size 2 versus set

**Fig. 2** Estimates of visual working memory (VWM) storage capacity ( $K$ ) as captured using the whole-display one-shot change detection paradigm throughout Experiments 1a–1c: (Top) Group data. (Bottom) Individual differences



sizes 4 and 8,  $F(1, 26) = 55.96, p < .05$ . These results suggest that capacity estimates based on set size 2 significantly underestimate limitations in storage capacity.

Viewing only the images in the top row of Fig. 2, presenting results at the group level, may lead one to infer that capacity estimates are constant across set sizes of 4 and 8, whereas plotting the individual  $K$  values for each observer can reveal a dramatically different picture (Fig. 2, bottom row). A large spreading of  $K$  estimates happens at set size 8, as compared to set size 4. Specifically, the  $K$  estimates for some observers actually go down at set size 8 relative to set size 4, whereas other observers show the opposite of this trend, with constant or increasing capacity between set sizes 4 and 8. In Fig. 2, bottom row, we have colored the results from each observer in order to aid in viewing this variety. Observers with higher capacity estimates at set size 4 are colored in warmer shades, whereas observers with lower capacity estimates at set size 4 are colored in cooler shades. What can be seen is that this nice separation at set size 4 turns into a “mish-mash” of different colors at set size 8. That is, the capacity estimates at set size 4 (i.e., warmer above and cooler below) tell one very little about the capacity estimates at set size 8 (i.e., some with higher capacity at set size 4 remain high, some become lower; some with lower capacity at set size 4 go higher, some remain low). This is an imagistic way of seeing what turns out to be a weak correlation between the performance at set sizes 4 and 8.

This large variety of performance at set size 8 relative to set size 4 is seen in a moderate, *nonsignificant* correlation between the  $K$  estimates calculated at those set sizes ( $r = .31, p = .28$ ). In fact, the correlations calculated for  $K$  estimates failed to reach significance for any set sizes (Table 1). That is, an observer’s  $K$  at one set size tells you basically nothing about what  $K$  will be at another set size. The lack of a correlation between set sizes 4 and 8 can be seen in the scatterplot of  $K$  estimates in Fig. 3.

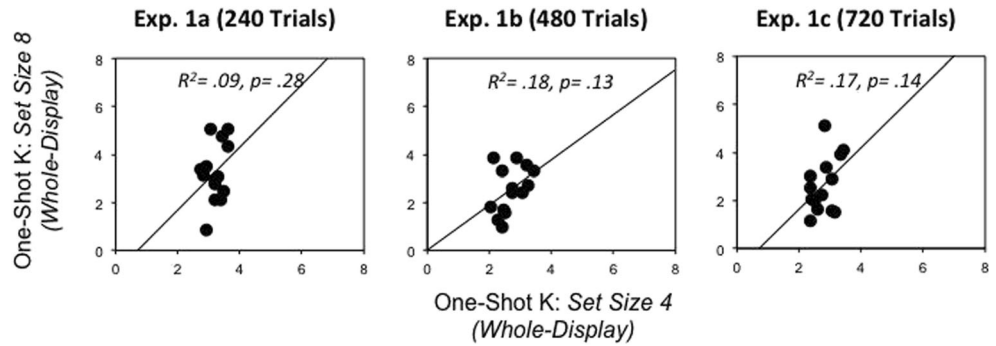
Although a very weak correlation between the  $K$  estimates calculated using set sizes 2 and 4 ( $r = .16, p = .59$ ) and set sizes 2 and 8 ( $r = -.31, p = .28$ ) may stem from capacity being underestimated at the lowest set size, the discrepancy between  $K$  values calculated at set sizes 4 and 8 is quite unexpected, given the widely held assumptions that VWM capacity,  $K$ , is a durable psychological parameter, that it is relatively constant across set sizes, and that it is well measured by the one-shot change detection paradigm.

We assessed the internal consistency of the whole-display one-shot paradigm by performing a random-halves reliability analysis collapsing across set sizes. To this end, we randomly shuffled and then divided each observer’s data, and subsequently computed averaged capacity estimates for each of these halves. We then corrected the resulting correlations using the Spearman (1910)–Brown (1910) prophecy formula to adjust for the shortened number of trials that results from splitting the data into halves. In so doing, we obtained a random-halves reliability coefficient ( $r = .28, p = .005$ ) that indicated low internal consistency. We further did this same process 100 times, calculating the correlation anew with each random shuffle. This provided us with 100 estimates of the random-halves reliability of  $K$  estimates that had been averaged across the three set sizes. A histogram of these 100

**Table 1** Correlations between one-shot (whole-display)  $K$  estimates across all set size pairs, for increasing numbers of trials (Exps. 1a–1c)

	One-Shot $K$ (Whole-Display)		
	Exp. 1a 240 Trials	Exp. 1b 480 Trials	Exp. 1c 720 Trials
$r_{K2,K4}$	.16 ( $p = .59$ )	.49 ( $p = .08$ )	.58 ( $p = .03$ )
$r_{K2,K8}$	-.31 ( $p = .28$ )	.23 ( $p = .43$ )	.29 ( $p = .32$ )
$r_{K4,K8}$	.31 ( $p = .28$ )	.42 ( $p = .13$ )	.41 ( $p = .14$ )

**Fig. 3** Correlations between whole-display one-shot  $K$  estimates across set sizes 4 and 8 throughout Experiments 1a–1c



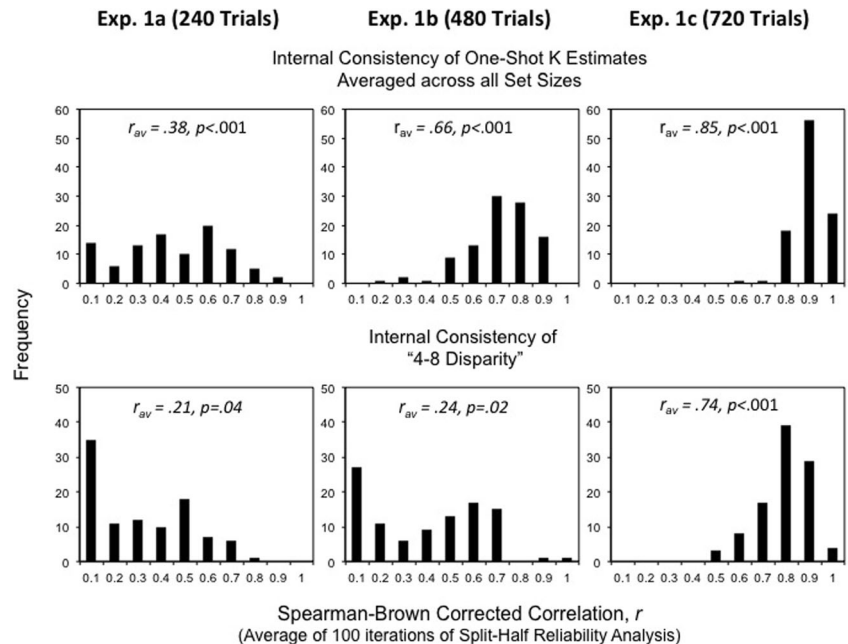
random-halves iterations can be seen in Fig. 4 (top row), and the average  $r$  was low (.38), indicating poor internal consistency for  $K$  after averaging across all set sizes. Such instability may have resulted from the inclusion of performance on set sizes that were far below capacity in the averaged  $K$  estimates. To address this concern, we additionally performed an analysis of internal consistency on the  $K$  estimates averaged solely across set sizes 4 and 8. This approach did not improve the average Spearman–Brown correlation for 100 iterations of a random-halves reliability analysis ( $r_{av} = .39, p < .001$ ). Furthermore, to ensure that problems of instability were not introduced by simply averaging across multiple set sizes, we performed a random-halves reliability analysis on the maximum  $K$  value calculated at any set size for each observer (Cowan et al., 2005). This too failed to produce an improvement in our measure of internal consistency ( $r_{av} = .06, p = .55$ ). Finally, we assessed the stability of the difference between  $K$  values at set sizes 4 and 8 (e.g., do subjects that go down in capacity consistently go down, or is this just random variability?). The difference between the  $K$  estimates at set

sizes 4 and 8 was also determined by 100 iterations of random-halves correlations, and we once again found low internal consistency ( $r_{av} = .21, p = .04$ ) for this difference between set sizes 4 and 8 (Fig. 4, bottom row).

**Experiments 1b and 1c**

The results of Experiment 1a challenge the assumption that the whole-display one-shot change detection task produces reliable estimates of storage capacity across set sizes. Given the wide adoption of this paradigm, this weakness may prove to be a concern. To determine whether these effects resulted from a lack of statistical power, we increased the number of trials used. We doubled (Exp. 1b) and tripled (Exp. 1c) the number of trials used, in an attempt to improve the reliability of the  $K$  estimates produced at different set sizes. We had two new groups of observers complete the one-shot task with these increased numbers of trials. The method and overall design of these experiments were identical to those used in Experiment 1a.

**Fig. 4** Distribution of internal consistency estimates produced by performing 100 iterations of a random-halves reliability analysis: (Top) Whole-display one-shot  $K$  estimates averaged across all set sizes. (Bottom) Disparity between whole-display one-shot  $K$  estimates calculated at set sizes 4 and 8



## Results

Estimates of VWM storage capacity were calculated using Pashler's (1988) equation. Data collected from one observer in Experiment 1c yielded a negative  $K$  value for set size 8 (the false alarm rate exceeded the hit rate). This individual was excluded from the analysis, since it is not theoretically possible to have a negative storage capacity.

The results of Experiments 1b and 1c are consistent with those observed in the previous experiment. Doubling or tripling the total number of trials did not alleviate the variability in capacity estimates at set sizes 4 and 8. As in Experiment 1a, main effects of set size emerged when observers completed 480 trials (Exp. 1b) of the one-shot task,  $F(2, 26) = 11.14, p = .003, \eta_p^2 = .46$ , and 720 trials of the same task (Exp. 1c),  $F(2, 26) = 8.70, p = .009, \eta_p^2 = .40$  (Fig. 2, top row). No significant differences were observed between the capacity estimates based on set sizes of 4 and 8 for both 480 trials,  $F(1, 26) = 0.38, p > .05$ , and 720 trials,  $F(1, 26) = 0.26, p > .05$ , and the capacity estimates based on set size 2 significantly differed from those for set sizes 4 and 8, for both 480 trials,  $F(1, 26) = 29.93, p < .05$ , and 720 trials,  $F(1, 26) = 26.01, p < .05$  (Fig. 2).

Table 1 displays the correlation coefficients for the capacity estimates obtained in Experiments 1a–1c. Despite the fact that doubling the total number of trials in Experiment 1b did increase the magnitude of the correlation coefficients for all set size pairs, these values still did not reach significance ( $K$  at set sizes 2 and 4,  $r = .49, p = .08$ ;  $K$  at set sizes 2 and 8,  $r = .23, p = .43$ ;  $K$  at set sizes 4 and 8,  $r = .42, p = .13$ ). A fortiori, tripling the number of trials in Experiment 1c also did not improve the relationships between the  $K$  values calculated across set size pairs ( $K$  at set sizes 2 and 4,  $r = .58, p = .03$ ;  $K$  at set sizes 2 and 8,  $r = .29, p = .32$ ;  $K$  at set sizes 4 and 8,  $r = .41, p = .14$ ). For the lack of correlation between set sizes 4 and 8, see Fig. 3.

Increasing the number of trials led to increases in the measures of internal consistency for  $K$  estimates averaged across all set sizes for both 480 trials ( $r_{av} = .66, p < .001$ ) and 720 trials ( $r_{av} = .85, p < .001$ ) (Fig. 4, top row). Similar estimates of internal consistency were obtained when averaging the  $K$  estimates solely across set sizes 4 and 8 (480 trials:  $r_{av} = .65, p < .001$ ; 720 trials:  $r_{av} = .85, p < .001$ ). Relative to these methods, performing 100 iterations of a random-halves reliability analysis on the maximum  $K$  value obtained at any set size yielded a comparable correlation for 720 trials ( $r_{av} = .83, p < .001$ ), but not for 480 trials ( $r_{av} = .44, p < .001$ ).

Once more, to determine whether the disparity between the capacity estimates at set sizes 4 and 8 resulted from random fluctuations or was due to a systematic difference, we performed 100 iterations of a random-halves reliability analysis on the differences between the capacity estimates at set sizes 4 and 8 for each observer. This analysis revealed low levels of

consistency for 480 trials ( $r_{av} = .24, p = .02$ ). In contrast, high levels of consistency were observed for 720 trials ( $r_{av} = .74, p < .001$ ) of the one-shot task, suggesting that observers were systematically performing differently between these set sizes (Fig. 4, bottom row). This means that, although the one-shot paradigm appears to measure something different at set sizes 4 and 8 (as demonstrated by the lack of a significant correlation between the  $K$  estimates at these set sizes; Figs. 2 and 3), each observer does something consistent *within* each of these set sizes, at least when tested for 720 trials (as demonstrated by the high internal consistency for the difference between set sizes 4 and 8; Fig. 4, bottom row).

Taken together, these results suggest that the whole-display one-shot paradigm is measuring different things (i.e., different psychological factors) at different set sizes, but also that the one-shot paradigm has good psychometric properties for measuring overall executive function and memory abilities once  $K$  estimates have been averaged from various set sizes (we do note, however, that the better internal consistency of the average is to be expected any time a variety of distinct psychological factors are averaged together).

## Experiment 2a

The whole-display one-shot task used throughout Experiments 1a–1c was based on most early experiments investigating storage limits in VWM. However, more recent studies have opted to measure individual differences in storage capacity using a single-probe change detection procedure with memory set sizes that are near or above the typical four-item capacity limit (four, six, and eight items).  $K$  estimates in these single-probe change detection tasks are calculated using Cowan's (2001) equation. In Experiments 2a and 2b, we assessed the reliability and internal consistency of  $K$  estimates calculated using this design.

### Method

**Observers** Fourteen Johns Hopkins University students with normal or corrected-to-normal vision participated in our study in exchange for course credit.

**Design and procedure** Observers completed 240 trials of a single-probe change detection task. Each trial began with a central fixation cross that was presented for 500 ms, after which observers were presented with a memory array of four, six, or eight colored squares (visual angle of each square =  $0.79^\circ \times 0.79^\circ$ ; the colors were chosen at random without replacement from ten discrete color possibilities). After a 900-ms blank retention interval, one square appeared at a previously occupied location. On half of all trials, the color

of this target item was identical to that of the item that had appeared at that location in the memory array. On the remaining half of trials, the target item appeared in a new color (Fig. 5). Observers were instructed to make a keyboard press to indicate whether the target item appeared in the same or a different color, relative to the memory array. Performance accuracies were recorded and used to calculate individual differences in VWM storage capacity.

**Calculation of VWM storage capacity  $K$**  Capacity estimates obtained using the single-probe task were calculated using Cowan's (2001) equation (Eq. 2). This method is based on the same underlying rationale as Pashler's (1988) equation:

$$K = N(H - F) \quad (2)$$

Overall change detection accuracy (operationalized as the hit and false alarm rates) is taken to represent the ratio of the maximum number of items that can be stored in VWM relative to the total number of items presented. Cowan (2001) and Pashler's (1988) equations are identical, with the exception that Cowan's equation lacks the "(1 - F)" denominator. This difference emerges from the methodological difference that observers either evaluate the status of a single item or all items presented at test. This change affects the nature of the guessing parameter (guessing in a single-probe task is uninformed by capacity and set size, whereas guessing in a whole-display task is heavily influenced by these factors; see Rouder et al., 2011, for a thorough explanation). In Experiment 2, we relied on capacity estimates calculated using Cowan's (2001)

equation to compare across set sizes, and then averaged these values across set sizes to produce overall estimates of individual VWM storage capacity.

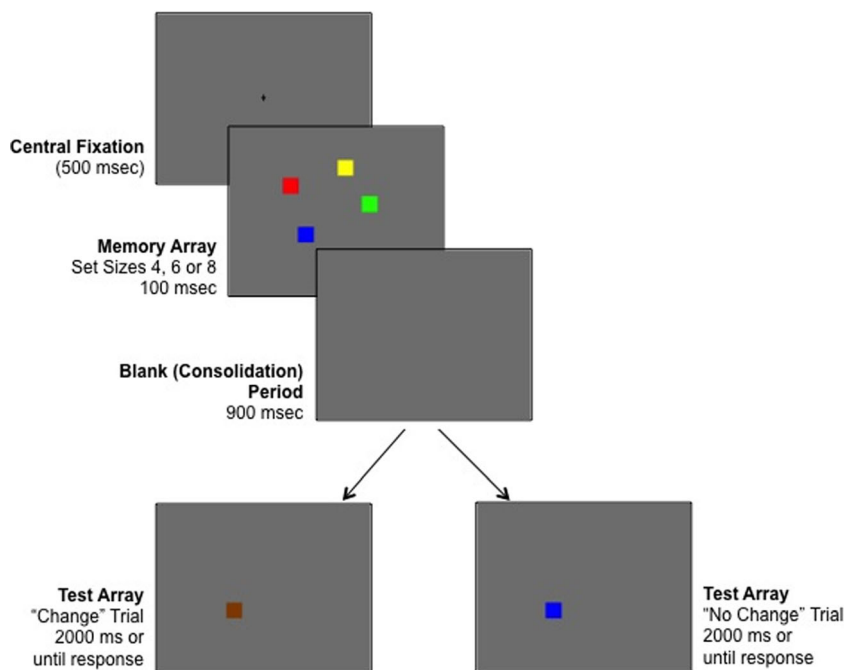
## Results

$K$  estimates calculated for each set size are illustrated in Fig. 6. A within-subjects ANOVA on these values yielded a significant effect of set size,  $F(2, 26) = 7.90, p = .009, \eta_p^2 = .38$  (Fig. 6, top row). Post-hoc contrasts failed to reveal a significant difference between the capacity estimates at set sizes 4 and 8,  $F(1, 26) = 0.58, p > .05$ . However, capacity estimates calculated at set size 6 were significantly lower than those calculated at set sizes 4 and 8,  $F(1, 26) = 14.09, p < .05$ .

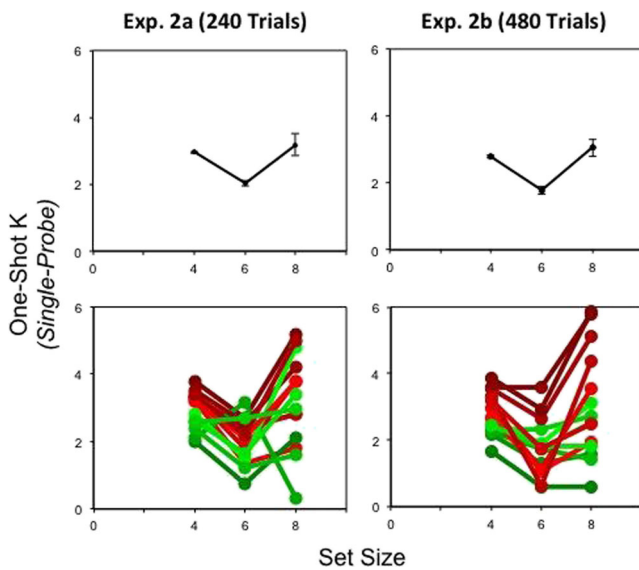
This pattern is further reflected in the individual  $K$  values for each observer (Fig. 6, bottom row). Correlations among the  $K$  values (Table 2) reveal a strong relationship between the estimates calculated at set sizes 4 and 8 ( $r = .69, p = .007$ ). In contrast, capacity estimates based on set size 6 produced weak and nonsignificant relationships with those calculated at set size 4 ( $r = .34, p = .24$ ) and set size 8 ( $r = .09, p = .76$ ) (Fig. 7).

The averaged  $K$  estimates proved to be highly internally consistent (Fig. 8), as demonstrated by the average of 100 iterations of a Spearman–Brown-corrected random split-half analysis ( $r_{av} = .79, p < .001$ ). Estimates of internal consistency obtained by performing 100 iterations of a Spearman–Brown reliability analysis on the maximum  $K$  value observed at any set size produced a relatively lower correlation ( $r_{av} = .53, p < .001$ ). And, an analysis performed on the difference between the  $K$  values estimated at set sizes 4 and 6 suggests weak internal consistency in this difference ( $r_{av} = .24, p < .02$ ). In

**Fig. 5** Schematic of a typical single-probe one-shot change detection trial, used throughout Experiments 2a and 2b







**Fig. 6** Estimates of VWM storage capacity (*K*) as captured using the single-probe one-shot change detection paradigm throughout Experiments 2a and 2b: (Top) Group data. (Bottom) Individual differences

contrast, the difference between the *K* values estimated at set sizes 6 and 8 proved to be moderately internally consistent ( $r_{av} = .62, p < .001$ ).

### Experiment 2b

The results of Experiment 2a demonstrated the strengths of using a single-probe one-shot change detection task, as compared to a whole-display version, in estimating individual differences in VWM storage capacity. Specifically, 240 trials of a single-probe task produced averaged *K* estimates that were highly internally consistent, as well as estimates that were highly reliable between set sizes 4 and 8. However, weak relationships were observed between the *K* estimates calculated at set sizes 4 and 8 with those calculated at set size 6. Inspired by Experiment 1, we sought to determine whether increasing the number of trials used in a single-probe task would improve the strength of relationships involving the *K* values calculated at set size 6. In Experiment 2b, a new group of 15 participants completed 480 trials of a single-probe

**Table 2** Correlations between one-shot (single-probe) *K* estimates across all set size pairs, for increasing numbers of trials (Exps. 2a–2b)

	One-Shot <i>K</i> (Single-Probe)	
	Exp. 2a 240 Trials	Exp. 2b 480 Trials
$r_{K4,K6}$	.34 ( $p = .24$ )	.58 ( $p = .03$ )
$r_{K4,K8}$	.69 ( $p = .007$ )	.91 ( $p < .001$ )
$r_{K6,K8}$	.09 ( $p = .76$ )	.67 ( $p = .009$ )

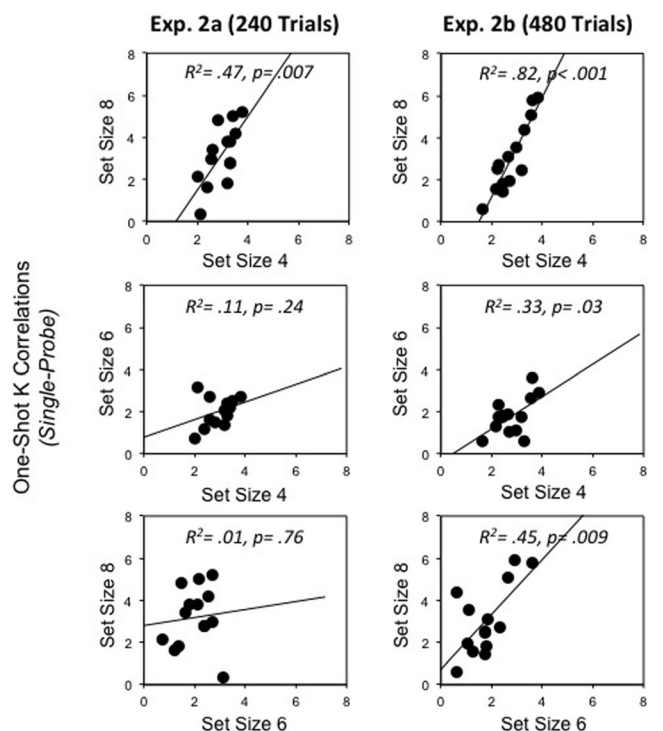
change detection task that was identical to that used in Experiment 2a. Once more, we measured the reliability among the *K* estimates calculated at each set size, as well as the internal consistency of the averaged *K* estimates.

### Results

Estimates of VWM storage capacity were calculated using Cowan’s (2001) equation. The data collected in Experiment 2b from one observer yielded a negative *K* value for set size 8, and the data from another observer yielded negative *K* values at all set sizes. Both individuals were excluded from the analysis.

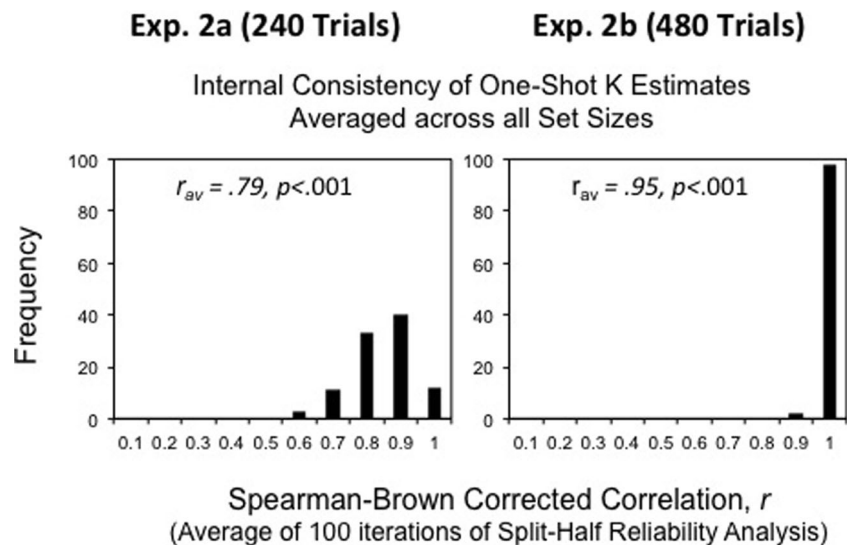
Consistent with the results of Experiment 2a, main effects of set size were observed for 480 trials,  $F(2, 26) = 11.09, p = .001, \eta_p^2 = .46$ , of a single-probe one-shot change detection task (Fig. 6). No significant differences were observed between the capacity estimates calculated at set sizes 4 and 8 for 480 trials,  $F(1, 26) = 0.78, p > .05$ , though the set size 6 estimates did differ significantly from those calculated at set sizes 4 and 8,  $F(1, 26) = 24.17, p < .05$ .

In Experiment 2b, 480 trials of a single-probe one-shot task improved relationships (Fig. 7), as exemplified by correlations between the *K* estimates calculated at set sizes 4 and 8 ( $r = .91, p < .001$ ), 4 and 6 ( $r = .58, p = .03$ ), and 6 and 8 ( $r = .67, p = .009$ ). The resulting average of 100 iterations of a Spearman–Brown corrected random split-half reliability analysis performed on the averaged *K* estimates (Fig. 8) indicated



**Fig. 7** Correlations between single-probe one-shot *K* estimates across all combinations of set sizes in Experiments 2a and 2b

**Fig. 8** Distribution of internal consistency estimates produced by performing 100 iterations of a random-halves reliability analysis on single-probe one-shot  $K$  estimates averaged across all set sizes



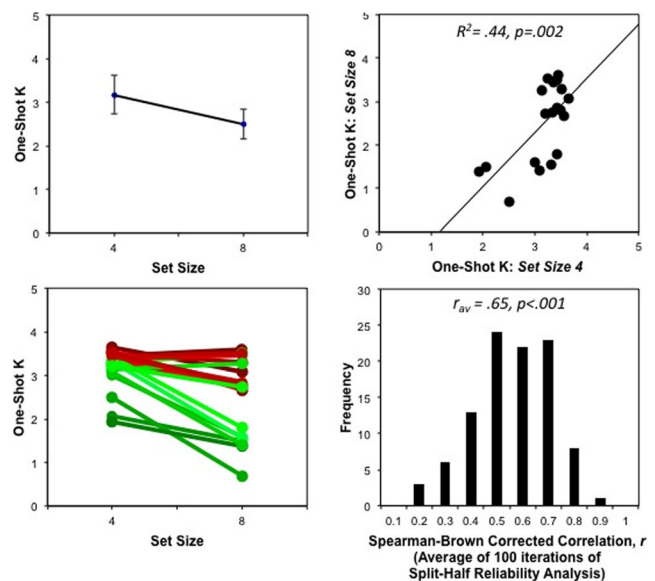
excellent internal consistency ( $r_{av} = .95, p < .001$ ), and this consistency was also seen in a Spearman–Brown reliability analysis on the maximum  $K$  value observed at any set size ( $r_{av} = .92, p < .001$ ).

#### Discussion

In Experiment 1, our assessment of the psychometric properties of the one-shot task with whole-display presentation revealed a lack of agreement between the capacity estimates calculated at different set sizes. Such incongruences may have resulted from observers performing the task differently at various set sizes. For example, the optimal strategy for detecting a change in a supracapacity display (e.g., a set size 8 display) may be different from the one suited to lower set sizes (e.g., a set size 4 display). One relevant strategy for larger arrays would be if observers relied somewhat on an ensemble “gist” representation of the scene in addition to representations of the single items. Observers might differ in their ability to implement this strategy—for instance, some might be able to do so successfully, whereas other might fall apart when faced with an overabundance of information.

Across Experiments 2a–2b, we found a clear improvement in the observed relationships between the capacity estimates calculated at set sizes 4 and 8, with somewhat reduced improvements for set size 6. Whereas whole-display designs (Exps. 1a–1c) demonstrated a weak relationship, single-probe designs (Exps. 2a–2b) produced a strong relationship. We would suggest that the superiority of the single-probe task may lie within its design. Specifically, the single-probe task demands that observers base their change detection response on an individual item. This requirement may lead observers to adopt a consistent strategy for both smaller and larger set sizes. One possible strategy is that observers, rather than using gist, might load *individual* items into VWM in order to

compare them against the single item presented at test. Under such a story, the requirement to focus on individual items is the basis for the improved reliability of the single-probe design. A question arises whether it would be possible to engender this type of focusing even within a whole-display design. In Experiment 3, we investigated whether we could increase the reliability of capacity estimates across set sizes in a whole-display design by requiring observers to report the single item that changed.



**Fig. 9** Performance in a one-shot change detection task with an added target localization criterion: (Top left) Group  $K$  estimates across set sizes. (Bottom left) Individual differences in performance in  $K$  estimates across set sizes. (Top right) Correlations between capacity estimates calculated at set sizes 4 and 8. (Bottom right) Distribution of internal consistency estimates produced by performing 100 iterations of a random-halves reliability analysis for difference between the capacity estimates calculated at set sizes 4 and 8

### Experiment 3

In Experiment 3, we investigated whether the correlation between capacity estimates calculated at set sizes 4 and 8 in a whole-display task could be improved by pushing observers to use a more local strategy like the one in the single-probe task. We had 19 observers complete 640 trials of a whole-display one-shot change detection task that was identical to that used in Experiments 1a–1c, with the following exceptions. First, in Experiment 3, the displays consisted only of either four or eight items. Second, when observers responded that they had perceived a change between the memory and test displays, they were required to localize the change by using the mouse cursor to click on the suspected changing target. As in Experiment 1, we investigated the correlation between the one-shot  $K$  values for set sizes 4 and 8—and we looked for an improved correlation relative to what had been seen in Experiment 1, in which no localization of the changing target was required.

### Results

Estimates of VWM storage capacity were calculated using Pashler's (1988) equation.

The estimated  $K$  values for each set size are shown in Fig. 9 for the group (top left), and for each individual subject (bottom left). As predicted, we found a significant correlation between the estimates of  $K$  at set sizes 4 and 8,  $r = .67$ ,  $p = .002$ . This correlation can be seen graphically in Fig. 9 (top right). Notice that this is a dramatic improvement over the nonsignificant correlations that we saw between set sizes 4 and 8 in Experiment 1, where no clicking on a changed target was required (e.g.,  $R^2 = .17$  in Exp. 1c for 720 trials). This suggests that requiring observers to localize the changing target by clicking on the suspected target whenever they report that they believe a change has occurred leads to a more consistent strategy across set sizes 4 and 8, resulting in a much higher correlation in one-shot  $K$  estimates across these set sizes.

As in Experiment 1, we also investigated the changes in  $K$  between set sizes 4 and 8 and also the internal consistency of the  $K$  estimates. The averaged  $K$  estimates proved to be highly internally consistent, as demonstrated by the average of 100 iterations of a Spearman–Brown-corrected random split-half analysis ( $r_{av} = .90$ ,  $p < .001$ ). A similar analysis on the maximum  $K$  value observed at either set size produced a lower but still strong relationship ( $r_{av} = .74$ ,  $p < .001$ ). Finally, any difference between  $K$  at set size 4 and 8 was stable, as revealed by 100 iterations of Spearman–Brown-corrected random-halves correlations, for which we found internal consistency ( $r_{av} = .65$ ; Fig. 9, bottom right). This stability can also be seen in Fig. 9 (bottom left), where it is clear that observers with higher capacity at set size 4 (warmer tones) also tend to be the individuals who have higher capacity at set size 8 (i.e., warmer

above cooler tones). These results suggest that the one-shot paradigm can return improved correlations across set sizes when observers are required to click on the suspected changing target whenever they report that a change has occurred.

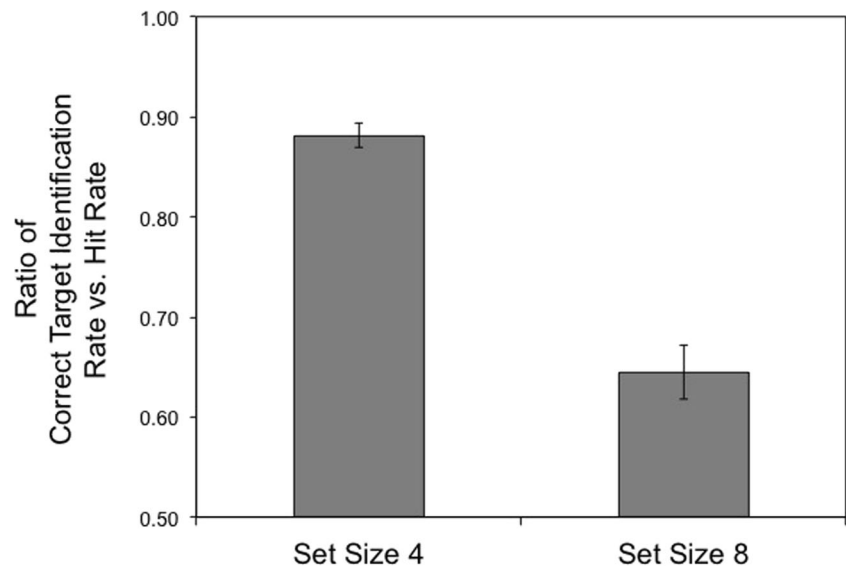
Finally, we investigated how accurately observers were able to localize the change. We determined the proportion of trials on which observers correctly clicked on the changed target, and we compared this to the proportion of hit trials—on which observers correctly reported that a change had occurred. A two-way (set size [4 vs. 8] and response measure [proportions of hits vs. correct target identifications]) within-subjects analysis of variance (ANOVA) yielded a main effect of set size,  $F(1, 18) = 432.71$ ,  $p < .001$ . Furthermore, the ANOVA produced a main effect of response measure,  $F(1, 18) = 75.11$ ,  $p < .001$ , and a significant interaction of set size and response measure,  $F(1, 18) = 432.71$ ,  $p < .001$ . This suggests that observers did not always have knowledge of which item had changed when they accurately responded that a change had occurred, and that they were less likely to know which item was the target when faced with the larger set size. However, given that the comparison of correct identifications across set sizes might be unfair, due to varying hit rates (i.e., a low proportion of set size 8 hit trials would lessen the number of correct target identification trials), we compared the ratios of correct target identification trials to the proportions of hit trials for both set sizes (Fig. 10). A one-way (set size: 4 vs. 8) within-subjects ANOVA revealed a significant difference in this ratio,  $F(1, 18) = 123.45$ ,  $p < .001$ , suggesting that observers were indeed less able to localize the target item when detecting a change in the eight-item displays. Such differences may provide an important estimate of VWM performance beyond the percent correct at each set size.

Taken together, these results suggest that the reliability of the  $K$  estimates obtained using a whole-display one-shot task can be improved by requiring observers to localize the change. It may be that doing so leads observers to adopt a strategy involving more local, item-based representations, rather than more global, gist-like representations. However, it appears that observers are still able to use a more global representation of the display at higher set sizes, since they were able to notice that a change had occurred without having knowledge of the specific item that had changed in a feature.

### Discussion

The results of Experiments 1a–1c suggested that the whole-display one-shot change detection task might not dependably capture individual differences in VWM storage capacity, because the estimates of  $K$  that it generates appear to measure different psychological factors at different set sizes. The results of Experiments 2a–2b (single-probe) suggested that this implementation of a single-probe one-shot task greatly improves the reliability and consistency of the estimates of

**Fig. 10** The higher ratio of target identification rate to hit rate for set size 4 than for set size 8 indicates observers' ability to detect change at a larger set size without knowledge of which specific item has changed in a feature



individual differences in VWM storage capacity, with the result being clear and consistent relationships in capacity across set sizes. The results of Experiment 3 (whole-display with a click on the changed item) also demonstrated that the whole-display one-shot task can return reliable and consistent estimates of VWM storage capacity. In this case, observers were required to localize the single changed item by clicking on it anytime they reported that a change had occurred. We hypothesized that both Experiments 2a–2b (single-probe) and 3 (whole-display with click) led observers to adopt a more local storage mode—wherein they tended to store individual items rather than a global gist. Of course, relational and global information are important and required to succeed at any change task (e.g., spatial layout for items, luminance/contrast relations across all items). For this reason, our suggestion is only that a shift toward a more local strategy (as well as a more consistent strategy across set sizes) may be important for the increases in reliability and consistency that we observed in Experiments 2a–2b and 3. In Experiments 4 and 5, we investigated an alternative approach (flicker change detection) for enhancing the consistency and reliability of estimates of VWM storage capacity.

#### Experiment 4a

When faced with a complex visual scene, individuals are free to search through objects in the scene, load items into and out of VWM, switch attention elsewhere, and use the information they gain to make decisions. These operations may be repeated as many times as desired until a visual target of interest is identified. And every complex scene has many more than three or four visual objects to search through. A flicker change

detection method can approximate some of these aspects of the real-world use of VWM.

Traditionally used in the change blindness literature, the flicker paradigm requires observers to localize a change that occurs between two alternating displays. These displays differ by a single feature and are separated by a blank interval that prevents attentional capture via visual transients (Rensink, O'Regan, & Clark, 2000). While the observer engages in an active visual search to find the changing item, this sequence of events is repeated until the target stimulus is identified. Response times are recorded and can be used as the dependent variable for estimating VWM storage capacity ( $K$ ). Although this is, admittedly, still far from a real-world context, a simplified flicker paradigm may allow one to study storage processes within the context of a dynamic visual search, and it may allow for investigations into how these processes operate over extended periods of viewing. It may also be that the requirement to localize the single changing item could encourage observers to store items in VWM as individuals and to search the array at a consistent rate dependent on their VWM storage capacity.

Rensink (2000) was the first to appropriate the flicker paradigm to estimate VWM storage capacity. In a series of experiments, he instructed observers to search for either a single orientation change that occurred among a set of vertical and horizontal black bars, or a single polarity change that occurred among black and white vertically oriented bars. Display durations were varied in these experiments, because Rensink believed that an increase in stimulus display time would allow observers to load more information into VWM up to a certain point. The limit at which performance reached asymptote was taken to represent an individual's VWM storage capacity. Using this method, Rensink estimated the limits for orientation and polarity to be 5.5 and 8–9 items,

respectively. It may be that these estimates do not converge with the three- to four-item capacity limits seen in the one-shot tasks because of the simplified items used (i.e., polarity and categorical orientation). These initial explorations have yet to be extended. Expanding on these ideas, it may become possible to estimate an observer’s average VWM capacity ( $K$ ) from an average response time (RT) in the flicker task using stimuli more similar to those in a traditional one-shot task. Here, we assessed the reliability and internal consistency of capacity estimates produced by the flicker paradigm. We used a flicker task that was designed to closely resemble the one-shot task used in Experiments 1a–1c (same set sizes, stimulus array properties, etc.). Similarly, we calculated flicker estimates of VWM storage capacity at each set size and determined the extent to which this paradigm produced  $K$  values that were reliable and internally consistent throughout periods of testing.

**Method**

*Observers* Seventeen Johns Hopkins University students with normal or corrected-to-normal vision took part in exchange for course credit.

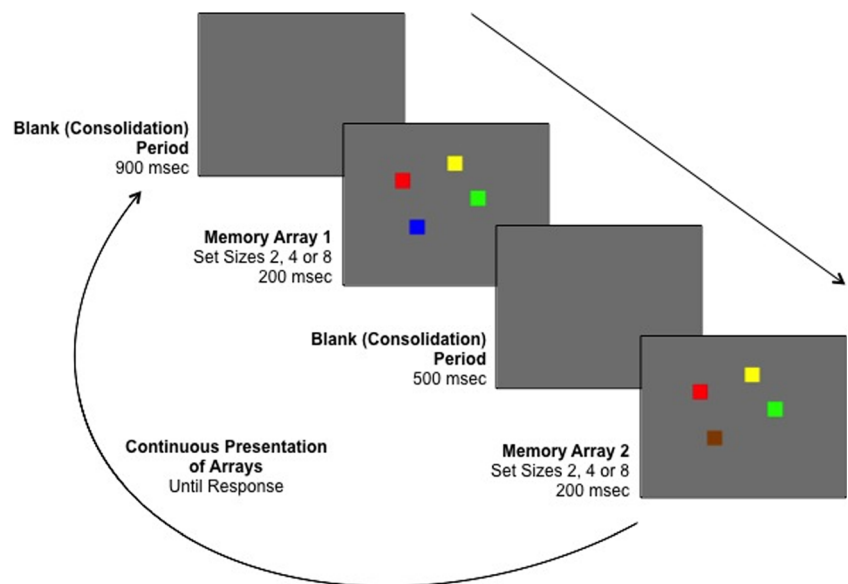
*Design/procedure* Observers completed 240 trials of a flicker change detection task that was designed to be comparable to the one-shot tasks used in Experiments 1a–1c. For this reason, the stimuli used in the flicker task consisted of colored squares that were randomly distributed across a homogeneous gray background. Once again, the color assignments for the squares were random without replacement and were selected randomly from one of ten discrete color possibilities (black, white, red, cyan, yellow, green, blue, orange, brown, or purple).

A schematic of a flicker trial is presented in Fig. 11. Each trial began with the presentation of a central fixation cross that was displayed for 500 ms, followed by a 200-ms memory array that consisted of two, four, or eight colored squares (random without replacement). This display was succeeded by a 500-ms blank interval and a 200-ms target array that was almost identical to the memory array. The masking effects of the subsequent display screen combined with the 500-ms blank interval are sufficient to minimize the contributions of iconic or sensory memory to flicker change detection performance (Halberda, Simons, Pailian, & Wetherhold, 2014; Rensink, O’Regan, & Clark, 1997).

Unlike the one-shot change detection paradigm, the memory and target arrays used in the flicker task *always* differed by a constant color change in one item. This sequence of displays continuously looped until observers pressed the spacebar, indicating that they had found the changing item (i.e., localized the change). This keypress caused the most recent display to be displayed on the screen, and observers were instructed to click on the square that they believed had been changing color. This requirement of having to localize the change presumably deters the use of global gist strategies that could be used in the one-shot whole-display change detection task—in which one need only report *whether* there was a change, not where the change was occurring. Response times (RTs) were recorded from the onset of the target display until the keypress, and were used to calculate individual VWM storage capacities ( $K$ ) and non-search-related activity ( $RT_{nsrch}$ ).

*Data analysis: The flicker task* As with visual search for a complex target, RT was the raw measure from a flicker trial. RTs would be affected by the number of items that the observer searched through before finding the changing target. If observers searched through fewer items, their RTs would be

**Fig. 11** Schematic of a typical flicker change detection trial, used throughout Experiments 4a–4c



faster, and if they had to search through more items, RTs would be slower (assuming a nonparallel search). Consistent with Rensink (2000), we used several simplifying assumptions for transforming the RTs into an estimate of capacity.

As a first assumption, we relied on the average RT to estimate average  $K$  (i.e., the first term in our model was  $RT_{avg}$ ). It could also be possible to rely on the full RT distributions to fit a  $K$  likelihood at the single-trial level, but such an approach is beyond the scope of the present article—here, we attempted a more straightforward extension that was inspired by the approach of Rensink (2000).

Because the flicker task involves visual search, we next specified a model of search. For instance, how many items, on average, does the observer search through before finding the target? Several search efficiencies have been discussed in the visual search literature. One is efficient search—in which the observer visits items in order to inspect whether they are the target, and if the observer finds that an item is not the target, it is marked in memory so that the observer never visits that item again. This search can be called “random without replacement” and is efficient in that it protects against revisiting previously visited items (Cave & Wolfe, 1990; Koch & Ullman, 1985; Treisman, 1993; Treisman & Gelade, 1980; Wolfe, 1994). For the flicker task, one can imagine an example of perfectly efficient search in which the subject loads objects A and B into VWM; waits during the blank screen; on the next flash, notices that A and B did not change; shifts attention and loads objects C and D into VWM; waits during the blank screen; and so forth. This subject looks at each object for the minimum amount of required time and does not repeat objects.

Previous modeling work on visual search has revealed that, if search is efficient (random without replacement), then a subject will have to search on average  $[(N + 1)/2]$  items before finding the target (Johnson & Kotz, 1977; Treisman & Gelade, 1980). Across all trials, sometimes the subject will be lucky and the target will be one of the first items attended; sometimes the observer will be unlucky and the target will be one of the last items attended. The intuitive understanding here is that across all trials averaged together, the subject’s average search time will be the time that it takes to search roughly half of the display.

A second aspect of the visual search model for the flicker task is that the search rate is constrained by the display duration and blank duration. This is because the subject loads objects during the display duration and holds them over the blank duration. If the blank is longer, then search times will be longer. For this reason, (display duration + blank duration) appears as a term in the search model to compute the search speed.

A subject might instead adopt many other search strategies. For example, one might adopt an inefficient search, or

“amnesic search” (Horowitz & Wolfe, 1998, 2003), in which information is not stored about which objects have already been visited during the search. Or, an observer might adopt a mixed strategy of spending an additional cycle on certain items, or repeatedly visiting the visually salient items. Because target identity and target position vary randomly from trial to trial in the flicker task, all of these various search strategies can be understood to affect the average number of visits made before the target is found (here counting each visit, even a visit to a previously seen item, as a new visit). These are important distinctions, and future work on the flicker task can benefit from a serious focus on comparing the computational predictions of various models and using empirical evidence—for example, from eyetracking—to adjudicate among the models. Here, we sought to determine a lower bound for VWM storage capacity estimates ( $K$ ) by adopting a maximally efficient model for visual search. We believe that this efficient model was close to true for the subjects in these experiments, and we felt it best to assume a single model for all observers. However, admittedly, questions of visual search are a major open area for understanding how the flicker task can be used to estimate VWM capacity ( $K$ ).

Our term for VWM storage capacity ( $K$ ) was the major interest for the present experiments, and it served to “chunk” the rate at which an observer searched the array. During active visual search, a subject’s performance in the flicker task can be affected by VWM storage capacity ( $K$ ). For example, if an observer has a VWM storage capacity of two items, then he or she might hold visual information from two items in VWM over the blank interval in order to compare them to items on the next flash. If a different observer has a VWM storage capacity of only one item, then this subject might hold just one item during the blank in order to compare it to this same item on the next flash. In this way, a subject with a larger VWM storage capacity would be able to search through the items more quickly to find the changing target. This difference in VWM storage capacity would thereby show up as an RT difference on the flicker task. In this way, the average RT on the flicker task can be used to estimate the average VWM storage capacity for each subject (i.e., the next term in our formula is a relationship between  $K$  and search rate).

As a final, nuanced, consideration, we note the possibility that non-search-related activity could distort search times in the flicker task. RTs in the flicker task might be influenced by non-search-related factors such as the executive control factors that determine how long it takes to initiate search, to direct attention to the stimulus array, to load a first set of items into memory (because no change detection would be possible until individual items are loaded into memory), to double check the changing item to ensure correct localization of the changing target, and to perform a motor response. For many researchers, these executive control factors, rather than VWM storage capacity, may be the factors of interest for empirical research.

Thus, it would be wrong to call the non-search-related components of the flicker paradigm nuisance parameters, or noise. Rather, we think it important to consider both the storage (i.e.,  $K$ ) and nonsearch (i.e.,  $RT_{\text{nsrch}}$ ) aspects of change detection performance, and we will look in detail at these two components in our analyses.

For the purposes of estimating VWM storage capacity ( $K$ ) from RTs in the flicker task, we will adopt an approach of subtracting non-search-related activity ( $RT_{\text{nsrch}}$ ) from the average response time ( $RT_{\text{avg}}$ ) in order to gain an estimate of active visual search time ( $RT_{\text{nsrch}}$ ). There is no straightforwardly accepted approach for estimating the time spent on non-search-related activity ( $RT_{\text{nsrch}}$ ). Elsewhere, we have adopted the approach of using the average of each subject's three fastest correct RTs to estimate non-search-related activity in the flicker task (Halberda et al., 2014); we have found that the results are the same in spirit if any number in the range of one to five trials is used, and we chose three as the middle of this range. The logic of this approach is that the subject will occasionally find the changing target among one of the very first items visited during the active search. On these trials, the majority of the RTs would be the time involved in non-search-related activity—with only very little time spent on actively searching. Thus, the average of the fastest correct RTs should provide an estimate of the duration of how long, on average, the subject spends on non-search-related activity. This approach is not perfect, and we believe it should be viewed as exploratory; however, if we did not subtract non-search-related activity from the average RT, our estimates of VWM capacity would be too low, because they would include both search and non-search-related processes. By subtracting this estimate (the average of the three fastest correct RTs) from the average RT, we can derive an estimate of the amount of time, on average, a subject spends actively searching the display (i.e., this subtraction appears in our formula as  $RT_{\text{avg}} - RT_{\text{nsrch}}$ ).

Before continuing, we note that the approach that we have described remains basically unchanged if one considers VWM to be limited by information rather than by the number of items that it can hold (for debate, see Alvarez & Cavanagh, 2004; Awh et al., 2007; Brady, Konkle, & Alvarez, 2009; Luck & Vogel, 1997; Vogel et al., 2001). Our interest in using RTs in the flicker task to estimate VWM storage capacity is neutral with respect to this important debate. Throughout, we will rely on language couched in terms of an object-limited VWM, but one could translate our claims by discussing units of information rather than objects, and by making adjustments to the maximally efficient search model to allow for decisions based on partial information.

Admittedly, our reasoning has required many assumptions. The core of the idea is that the average RT in the flicker task can generate an estimate of average VWM storage capacity. Many details related to search efficiency and non-search-

related activity are open for future research using the flicker paradigm, and the need for such assumptions could rightly be taken as a current limitation for using the flicker paradigm to estimate VWM storage capacity ( $K$ ). Here, we present a first foray into the consistency and reliability of estimates of VWM storage capacity ( $K$ ) derived from performance on the flicker task.

We can now write a formula that will translate the average response time on the flicker task ( $RT_{\text{avg}}$ ) into an estimate of average VWM storage capacity ( $K$ ):

$$RT_{\text{avg}} - RT_{\text{nsrch}} = (\text{display duration} + \text{blank duration}) * \frac{[(N + 1)/2]}{K} \quad (3a)$$

On the left side, we have our estimate of the amount of time, on average, a subject spends actively searching the display ( $RT_{\text{avg}} - RT_{\text{nsrch}}$ ). On the right side, we have a specification of how the search RT emerges from the search rate (display duration + blank duration) and the number of items, on average, visited before finding the target  $[(N + 1)/2]$ . The number of items searched is divided by the subject's VWM storage capacity ( $K$ ). Intuitively, the VWM storage capacity ( $K$ ) is serving to “chunk” the total number of items that have to be searched.

A rearrangement of Eq. 3a yields a formula for estimating VWM storage capacity ( $K$ ):

$$K = \frac{(\text{display duration} + \text{blank duration}) * [(N + 1)/2]}{RT_{\text{avg}} - RT_{\text{nsrch}}} \quad (3b)$$

In the following series of experiments, we used the flicker task to estimate individual differences in VWM storage capacity ( $K$ ) and non-search-related activity ( $RT_{\text{nsrch}}$ ) using Eq. 3b. As we did for the one-shot paradigm in Experiments 1a–1c, we investigated the psychometric properties of this paradigm, to evaluate its suitability for providing estimates of VWM storage capacity.

## Results

Given that a color change always occurred during each trial, observers were instructed to respond only when they knew which square was changing color. For this reason, trials in which observers incorrectly identified the target stimulus were excluded from the analysis. Moreover, observers whose overall incorrect identification rates exceeded 5% were excluded altogether (three observers).

Because the duration of non-search-related activity might change with set size, the three fastest RTs at each set size were averaged to produce a value for  $RT_{\text{nsrch}}$  at each set size—that

is, the amount of time an observer takes to build a gist of the array and initiate search by storing a first set of memory items, and so forth.

Because of floor effects in RTs, flicker estimates based on set size 2 would likely provide inflated  $K$  values, but we included this set size in our tasks in order to increase the comparability of psychometric properties between our flicker method and the one-shot change detection method of Experiments 1a–1c. We recommend that future flicker studies rely on set sizes larger than four.

RTs were trimmed above and below two standard deviations of the mean so as to minimize the effects of outlier RTs (4.9% of trials).

Values for  $K$  from the one-shot paradigm tend to span between one and six items (Perez & Vogel, 2012). Figure 12 illustrates the  $K$  estimates obtained for the flicker paradigm in Experiment 4 using Eq. 3b. Here, we observed a significant effect of set size,  $F(2, 26) = 6.76, p = .016, \eta_p^2 = .34$ . As in the results obtained throughout Experiments 1a–1c, we found no significant differences in  $K$  estimates between set sizes 4 and 8,  $F(1, 26) = 6.09, p > .05$ , whereas  $K$  estimates calculated at set size 2 did significantly differ from those at the other set sizes,  $F(1, 26) = 6.93, p < .05$ . These results suggest that estimates of capacity produced by the flicker task were inflated when observers were presented with two squares as compared to when they were presented with the larger set sizes, and perhaps that observers reduced the number of items they stored as set size increased (Fig. 12).

To determine the reliability of the flicker  $K$  estimates, we correlated the  $K$  estimates for all possible combinations of set sizes (Table 3). Capacity estimates based on set sizes 2 and 4 showed a strong correlation ( $r = .82, p < .001$ ), as did capacity estimates based on set sizes 2 and 8 ( $r = .55, p = .04$ ) and 4 and 8 ( $r = .82, p < .001$ ). Furthermore, the average of 100 iterations of a Spearman–Brown-corrected random-halves reliability analysis (Fig. 13) yielded an extremely high correlation

( $r_{av} = .98, p < .001$ ). One can also view the consistency of the  $K$  estimates across set sizes in Fig. 12 (bottom row), where it is clear that observers with higher capacity at set size 4 (reddish tones) also tended to be the individuals who had higher capacity at set size 8 (i.e., warmer above cooler tones across set sizes).

As further evidence of the correlation between VWM storage capacity estimates ( $K$ ) at set sizes 4 and 8, the correlation between these estimates can be seen in Fig. 14.

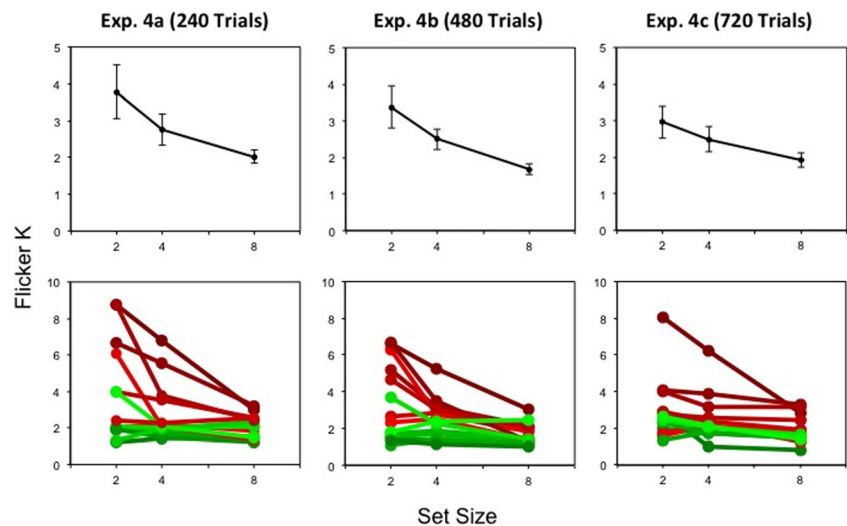
As a further check, we also investigated the consistency of the non-search-related response times ( $RT_{nsrch}$ ) across set sizes. These too showed high correlations across set sizes (Table 4), suggesting that individual differences in this parameter remained stable across changes in set size, which can be seen graphically in Fig. 15.

Taken together, these results suggest that the flicker paradigm produces both reliable and internally consistent estimates of VWM storage capacity ( $K$ ) and non-search-related activity ( $RT_{nsrch}$ ).

#### Experiment 4b and 4c

In Experiment 4a, we ran 240 trials of the flicker task, in an attempt to investigate the psychometric properties of this paradigm. We observed strong correlations between the  $K$  values estimated at different set sizes and consistency in the averaged  $K$  estimates across trials. Given the effects that increasing the number of trials had on the correlations in Experiments 1a–1c, we were interested to determine whether doubling or tripling the number of trials in our flicker task would further strengthen the reliability of this measure. As such, in Experiment 4b, we ran a new group of 15 observers on 480 trials of a flicker task that was identical to that used in

**Fig. 12** Estimates of VWM storage capacity ( $K$ ), as captured using the flicker change detection paradigm throughout Experiments 4a–4c: (Top) Group data. (Bottom) Individual differences





**Table 3** Correlations between flicker *K* estimates across all set size pairs, for increasing numbers of trials (Exps. 4a–4c)

	Flicker <i>K</i>		
	Exp. 4a 240 Trials	Exp. 4b 480 Trials	Exp. 4c 720 Trials
$r_{K2,K4}$	.82 ( $p < .001$ )	.83 ( $p < .001$ )	.92 ( $p < .001$ )
$r_{K2,K8}$	.55 ( $p = .04$ )	.51 ( $p = .07$ )	.64 ( $p = .013$ )
$r_{K4,K8}$	.82 ( $p < .001$ )	.80 ( $p = .001$ )	.77 ( $p = .001$ )

Experiment 4a, and in Experiment 4c a separate group of 16 observers on 720 trials of this same task.

**Results**

In Experiment 4b, one observer was excluded from the analysis due to incorrectly identifying a distractor on more than 5 % of the trials, and two observers were excluded in Experiment 4c on the basis of the same criterion. The three fastest RTs for each observer at each set size were once again used to estimate non-search-related activity ( $RT_{nsrch}$ ), and the remaining RTs were once again trimmed to less than two *SDs* from the mean for Experiments 4b (4.4% of correct trials) and 4c (4.6% of correct trials), and these were used to calculate *K* values based on Eq. 3b.

The capacity estimates obtained in Experiment 4b (480 trials) showed a pattern similar to those from Experiment 4a (Fig. 12). The main effect of set size,  $F(2, 26) = 10.11, p = .006, \eta_p^2 = .44$ , was driven by the *K* values calculated at set size 8 being lower than those estimated at set sizes 2 and 4,  $F(1, 26) = 13.90, p < .05$ , whereas the *K* values calculated at set sizes 2 and 4 did not differ,  $F(1, 26) = 5.78, p > .05$ . Capacity estimates across set sizes were still highly reliable, as is demonstrated by the correlations between the *K* values calculated at set sizes 2 and 4 ( $r = .83, p < .001$ ), 2 and 8 ( $r = .51, p = .07$ ), and 4 and 8 ( $r = .80, p = .001$ ). This consistency can also be viewed graphically in Fig. 12 (bottom row). As further evidence of the correlation between VWM

storage capacity estimates (*K*) at set sizes 4 and 8, the correlation between these estimates can be seen in Fig. 14. As in Experiment 4a, the estimates for non-search-related activity ( $RT_{nsrch}$ ) were consistent and stable across set sizes (Fig. 15 and Table 4).

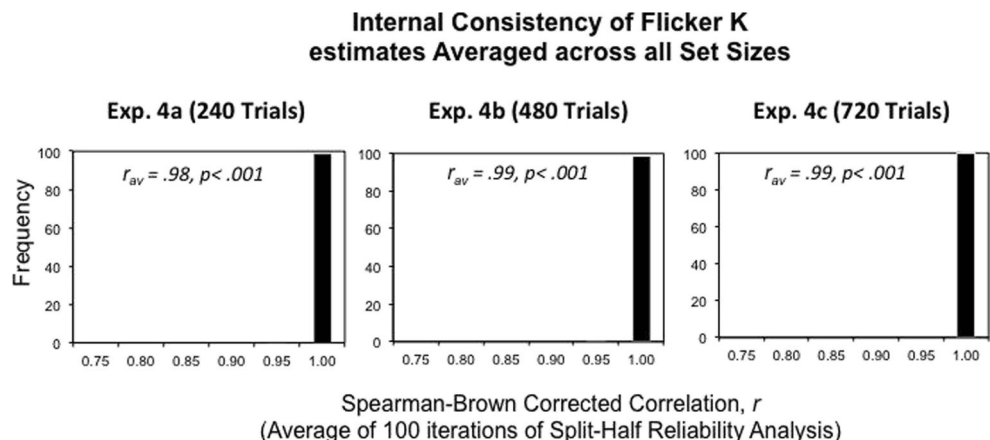
The pattern of results observed in Experiment 4c (720 trials) also replicated that of Experiment 4a. The ANOVA revealed a main of effect of set size,  $F(2, 26) = 7.95, p = .009, \eta_p^2 = .38$ . This effect was driven by a difference in the estimates calculated at set size 2 versus set sizes 4 and 8,  $F(1, 26) = 8.72, p < .05$ , whereas the *K* estimates at set sizes 4 and 8 did not differ,  $F(1, 26) = 6.62, p > .05$ . The *K* values across all set sizes were highly reliable, as demonstrated by correlations between the *K* estimates computed at set sizes 2 and 4 ( $r = .92, p < .001$ ), 2 and 8 ( $r = .64, p = .013$ ), and 4 and 8 ( $r = .77, p = .001$ ). This consistency can also be viewed graphically in Fig. 12 (bottom row). As further evidence of correlation between the VWM storage capacity estimates (*K*) at set sizes 4 and 8, the correlation between these estimates can be seen in Fig. 14. As in Experiment 4a, the estimates for non-search-related activity ( $RT_{nsrch}$ ) were consistent and stable across set sizes (Fig. 15 and Table 4).

In contrast to the results of Experiments 1a–1c, increasing the number of trials had little to no effect on the magnitude of correlations between the *K* values estimated at different set sizes (Table 3), due to a ceiling effect. The correlation between capacity estimates produced by the flicker task remained quite high, and remained relatively constant across the numbers of trials used. Such invariance was further reflected in the internal consistency of this paradigm, since the averages of 100 iterations of a Spearman–Brown-corrected random split-half analysis remained very strong for both 480 and 720 trials (for both,  $r_{avs} = .99, ps < .001$ ) of the flicker task (Fig. 13).

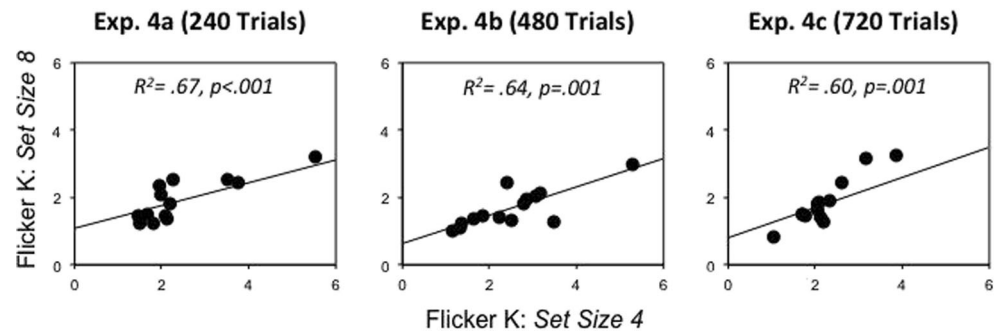
**Discussion**

We found that the flicker paradigm produced estimates of VWM storage capacity that were consistent across set sizes

**Fig. 13** Distribution of internal consistency estimates produced by performing 100 iterations of a random-halves reliability analysis on flicker *K* estimates averaged across all set sizes



**Fig. 14** Correlations between flicker  $K$  estimates across set sizes 4 and 8 throughout Experiments 4a–4c



and numbers of trials used. To highlight the robustness of this improvement relative to the one-shot paradigm, consider that at only 240 trials (Exp. 4a), the flicker paradigm produced  $K$  values whose correlation coefficients and estimates of internal consistency were already higher than those produced by 720 trials of a whole-display one-shot task (Exp. 1c). Furthermore, whereas the strength of these relationships for the one-shot task increased as a function of the number of trials used, reliability estimates for the flicker paradigm remained relatively unchanged across all trial sizes. These results suggest that the flicker paradigm is a reliable and internally consistent alternative for measuring individual differences in VWM storage capacity.

## Experiment 5

The one-shot paradigm has supported tremendous progress in our study of VWM. If the flicker paradigm is to be a valuable alternative to the one-shot paradigm (an alternative method that may have strong reliability and internal consistency), then it is important to determine whether the flicker estimate of VWM storage capacity ( $K$ ) measures the same psychological factor that the one-shot estimate of VWM storage capacity ( $K$ ) measures. In Experiment 5, to validate the flicker paradigm estimate of  $K$ , we had a new group of observers complete both the flicker and the whole-display one-shot change detection

**Table 4** Correlations between flicker non-search-related activity estimates (nonstorage executive control) across all set size pairs, for increasing numbers of trials (Exps. 4a–4c)

	Flicker $RT_{\text{non-search}}$		
	Exp. 4a 240 Trials	Exp. 4b 480 Trials	Exp. 4c 720 Trials
$r_{SS2,SS4}$	.64 ( $p = .015$ )	.90 ( $p < .001$ )	.61 ( $p = .021$ )
$r_{SS2,SS8}$	.53 ( $p = .053$ )	.82 ( $p < .001$ )	.68 ( $p = .008$ )
$r_{SS4,SS8}$	.68 ( $p < .007$ )	.92 ( $p < .001$ )	.71 ( $p = .004$ )

tasks, and we investigated the extent to which the estimates of storage capacity ( $K$ ) from these paradigms would correlate with each other.

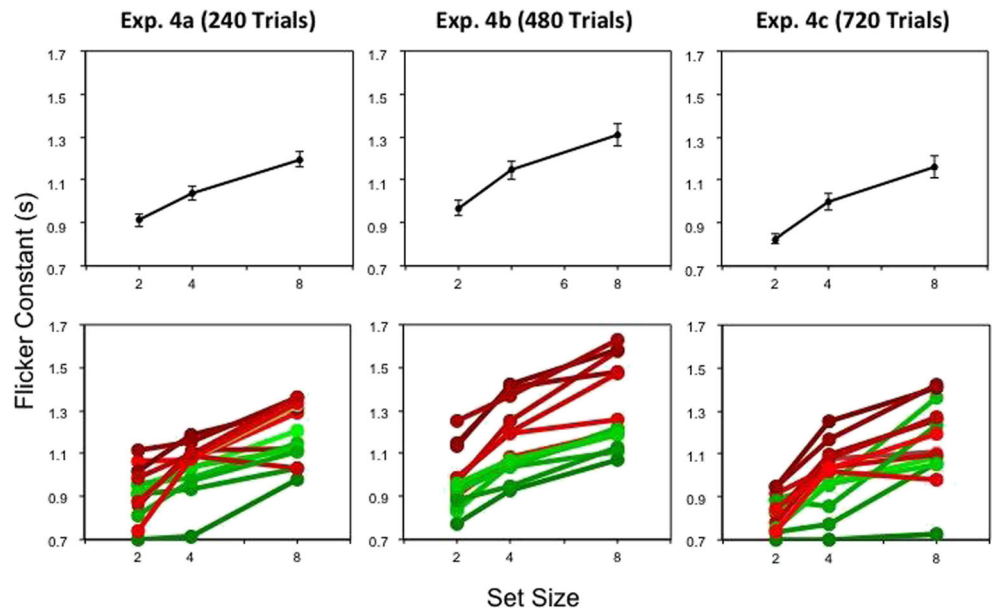
In Experiment 5, we had an additional goal. As we discussed in Experiments 4a–4c, the flicker paradigm generates estimates of both storage-related factors ( $K$ ) and non-search-related activity ( $RT_{\text{nrch}}$ ; e.g., executive functions, visual search, and decision criteria). The one-shot paradigm provides only one estimate, and traditionally authors have assumed that the only relevant factor measured by the one-shot paradigm is VWM storage capacity ( $K$ ). It may be that  $K$  from the one-shot conflates storage- and non-search-related activity into its single estimate of  $K$ . To explore this possibility, in Experiment 5 we included analyses that controlled for the flicker non-search-related activity while looking at the correlation between flicker  $K$  and one-shot  $K$ . To explore whether the one-shot  $K$  estimate also includes information about non-search-related activity, we also included analyses that first controlled for variations in flicker  $K$  before asking whether the one-shot  $K$  correlated with flicker non-search-related activity. If one-shot  $K$  includes information about non-search-related activity (e.g., executive functions, gist construction, and decision making), we should find that, even after controlling for flicker  $K$ , the one-shot estimate of  $K$  would correlate with flicker non-search-related activity. If true, this would begin to reveal something of the multiple factors that may be conflated within the one-shot  $K$  estimate, and a trade-off of factor loadings across set sizes might by one route for describing why the whole-display one-shot  $K$  does not appear to correlate across set sizes.

## Method

**Observers** Twenty-six Johns Hopkins University students with normal or corrected-to-normal vision took part in exchange for course credit.

**Design and procedure** The observers completed 480 trials of a flicker change detection task and 480 trials of whole-display one-shot change detection, in counterbalanced order. The designs of these tasks were identical to those used in

**Fig. 15** Estimates of executive control ( $RT_{nsrch}$ ), as captured using the flicker change detection paradigm throughout Experiments 4a–4c: (Top) Group data. (Bottom) Individual differences

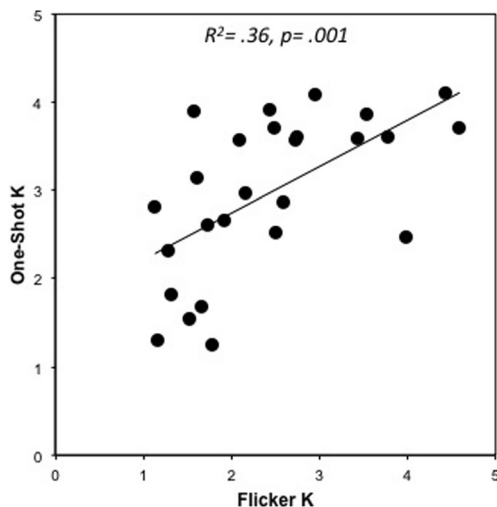


Experiments 1 and 4, with the exception that the displays consisted of four, six, or eight items.

**Results**

$K$  values for each paradigm were calculated for each set size and then averaged to produce an overall estimate of  $K$  for each task. Estimates of non-search-related activity were calculated for each set size, as described above, and were then averaged to produce an overall value for flicker non-search-related activity.

As a first estimate of the agreement between flicker  $K$  and one-shot  $K$ , we performed a linear regression of the  $K$  values from these two tasks. This returned a significant linear relationship that can be seen graphically in Fig. 16.



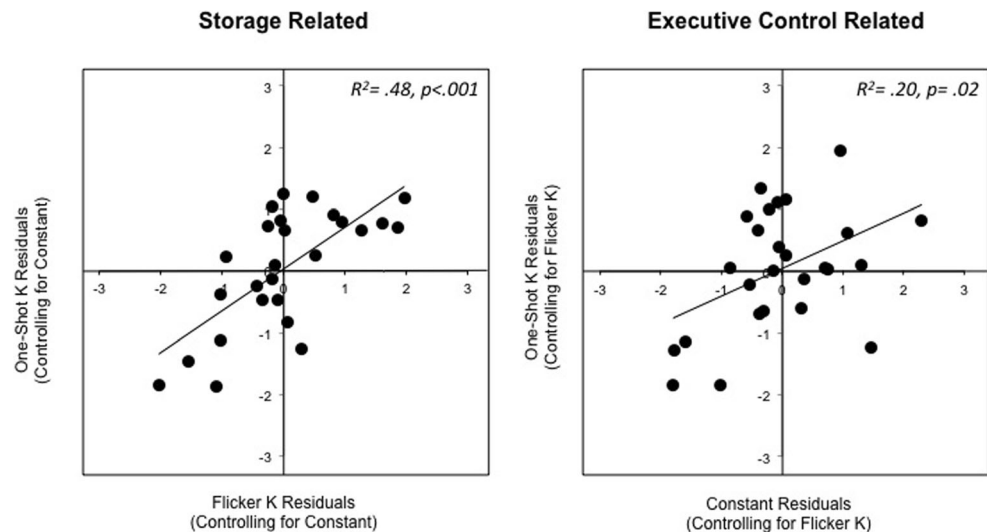
**Fig. 16** Overall correlation between flicker and one-shot estimates of storage capacity ( $K$ )

To explore this relationship in greater detail, we asked whether flicker  $K$  estimates would correlate with one-shot  $K$  estimates after removing the contribution of non-search-related activity ( $RT_{nsrch}$ ). We used a linear regression to partial out this variable from both the one-shot  $K$  and flicker  $K$  estimates and compared the residuals via linear regression. Even after controlling for non-search-related activity in this way, the correlation between the residuals for flicker  $K$  and one-shot  $K$  (Fig. 17, left side) remained significant, suggesting that these estimates are related by more than just executive functions (e.g., search efficiency; D. E. Anderson, Vogel, & Awh, 2013) and perceptual and decision factors ( $r = .70, p < .001$ ).

To ask whether the one-shot estimate of  $K$  also includes residual information about non-search-related activity hidden within its estimate, we performed the complementary analysis, in which we tested for a correlation between the one-shot  $K$  residuals (controlled for flicker  $K$ ) and the flicker non-search-related activity (controlled for flicker  $K$ ). This approach was an attempt to control for VWM storage, as estimated by the flicker  $K$ , and to ask whether the remaining variability (in one-shot  $K$ ) could be related to non-search-related activity (as estimated by flicker  $RT_{nsrch}$ ). Consistent with the suggestion that the one-shot paradigm’s estimate of VWM storage capacity ( $K$ ) includes residual information about non-search-related activity, we found that the one-shot  $K$  residuals (controlled for flicker  $K$ ) were significantly related to the flicker non-search-related activity (controlled for flicker  $K$ ),  $r = .45, p = .02$ . This relationship can be seen graphically in Fig. 17, right side.

Together, these results suggest that the flicker paradigm’s estimate of VWM storage capacity ( $K$ ) measures the same thing that the one-shot composite estimate of  $K$  is estimating,

**Fig. 17** Partial correlations between estimates of performance in the flicker and one-shot change detections tasks: (Left) Partial correlation between flicker and one-shot  $K$  estimates, controlling for the flicker non-search-related activity—a measure of the storage-related abilities involved in both tasks. (Right) Partial correlation between the flicker non-search-related activity and one-shot  $K$  estimates, controlling for flicker  $K$  estimates—a measure of the executive-control-related abilities involved in both tasks



and suggests that the one-shot's estimate of  $K$  may also include residual information about non-search-related activity hidden within.

## General discussion

We assessed the psychometric properties of the hallmark method for estimating individual differences in VWM storage capacity, the one-shot change detection paradigm, and a novel approach, the flicker paradigm. The one-shot change detection task has been popular within the VWM literature, since its “one-shot” nature has been thought to isolate storage abilities by preventing the use of complex strategies. Ironically, we found that presenting whole displays only once while varying set size might have an effect opposite from the intended one—that is, although performance can be highly restricted at lower set sizes, additional non-storage-related cognitive processes such as ensemble gist representation may come into play at higher set sizes. Thus, the one-shot paradigm may be measuring different things at different set sizes. The low reliability of capacity across set sizes that we found for the whole-display one-shot task may connect with recent discoveries that low-capacity subjects may show a drop in capacity as the array set size increases (Linke et al., 2011; Matsuyoshi et al., 2014). Such drops might reduce the reliability and consistency of capacity estimates from the one-shot task. In support of this intuition, we found no significant correlation between the one-shot  $K$  estimates at set sizes 4 and 8, even after observers participated in more than 700 trials. On the basis of these shortcomings, we have, with humility, called the stability and reliability of the one-shot whole-display paradigm into question.

We also found that some variations of the one-shot paradigm do show improved reliability and consistency. Noteworthy, these variations may require the observer to

localize the changing item, either by including only a single probe item in the change display (Exps. 2a–2b) or by requiring the observer to click on the item that is believed to have changed (Exp. 3). Determining how variations of the one-shot task change subjects' strategies and the encoding of information into VWM, and how such changes affect the reliability and consistency of estimates of VWM storage capacity, could be areas for continued research.

We should also note the importance and value of the traditional one-shot whole-display paradigm. Even in Experiment 1, we found acceptable levels of internal consistency when the one-shot task's estimates of  $K$  were collapsed across set sizes. This is the measure from the one-shot task that is most typically used in the literature. This stability does mean that previous results using this estimate are likely to be reliable.

We also found that an alternative method, the flicker paradigm, provides high internal consistency and that its estimate of VWM storage capacity ( $K$ ) correlates well across set sizes. We also found that flicker  $K$  correlated well with one-shot  $K$ , suggesting that the flicker paradigm may be a viable alternative for measuring individual differences in VWM storage capacity.

One avenue that may be promising for future work will be looking at multiple abilities within a task rather than computing a single metric. The estimates of  $K$  derived from the one-shot whole-display task may be measuring both storage-related and non-storage-related psychological factors. The flicker paradigm offers two separate estimates of these factors (i.e., storage  $K$  and the non-search-related RT). Each of these was found to correlate with one-shot  $K$  when controlling for the other (Exp. 5). We believe that this approach of exploring various factors within the composite measures generated by these two methods could prove to be productive for developing further models of VWM and task performance.

For example, consider performance on the one-shot whole-display task. The variability in performance at set size 8 in the

one-shot task may reflect individual differences in the ability to effectively organize large amounts of information during encoding. This suggestion has been promoted by Vogel, McCollough, and Machizawa (2005), who demonstrated that  $K$  estimates correlate with the ability to filter irrelevant information. In their study, observers were asked to judge whether an orientation change occurred in a set of relevant colored bars that were presented amidst a set of irrelevant colored bars (distractors). They found that the performance exhibited by high- $K$  individuals (with  $K$  averaged across all set sizes) for remembering two relevant bars presented among two distractors was comparable to performance for remembering two relevant bars presented in the absence of distractors. In contrast, low- $K$  individuals were unable to filter out the irrelevant information, and, as such, one-shot change detection performance for two relevant bars presented among two distractors was comparable to performance when four relevant bars were presented in the absence of distractors. Fukuda and Vogel (2011) have further built on this finding by demonstrating that high- $K$  individuals (averaged across all set sizes), in comparison to low- $K$  individuals, are able to more quickly disengage from distracting stimuli that capture attention. The latter study lends further support to the notion that capacity estimates produced by the one-shot task may be indicative of executive control abilities (e.g., widening the scope of attention or filtering of information), as well as control of what information gains access into VWM. This argument may provide an alternative interpretation of the findings reported by Matsuyoshi, Osaka, and Osaka (2014), who claimed that VWM storage capacity is inherently unstable. In their study, young adults were separated into high- and low- $K$  categories on the basis of their performance on set size 12 trials in a one-shot task. Whereas low- $K$  individuals demonstrated a decrease in capacity estimates as a function of set size, high- $K$  individuals exhibited the opposite trend. However, it is important to note that the division of individuals into high- and low- $K$  categories solely on the basis of performance on trials with an overload of information may not reflect VWM storage capacity per se, but rather, differences in encoding ability. Some individuals may be well equipped to organize information in an efficient manner, whereas others may experience catastrophic failure in doing so, or have chosen to be conservative in the amount of information that they encode (Linke et al., 2011). An account that highlights VWM use rather than VWM capacity limits is further supported by the observation that increasing encoding time for the objects allowed low- $K$  young adults to recover from the deficit observed at higher set sizes, whereas no added benefit was observed for high- $K$  individuals. Similar questions arise across the lifespan for both young children (Cowan et al., 2011; Cowan et al., 2010; Pailian, Libertus, Feigenson, & Halberda, 2014; Riggs et al., 2006; Simmering, 2012) and older observers (Matsuyoshi et al., 2014). For instance, differences between young and old adults observed by

Matsuyoshi et al. may reflect age-related impairments in attentional control and encoding ability, rather than differences in VWM capacity per se. Considering performance on small arrays that are lower than the typical four-item limit may be helpful toward estimating lapses in these executive control abilities (i.e., focus and maintenance of attention; Rouder et al., 2008). Testing the reliability of estimates of capacity and exploring a variety of tasks to estimate both storage- and non-storage-related individual differences will be important for answering these questions.

Additionally, one-shot  $K$  estimates may reflect differences in the level at which visual information is represented. According to hierarchical models of VWM (Brady & Alvarez, 2011; Brady, Konkle, & Alvarez, 2011; Halberda, Sires, & Feigenson, 2006), information can be represented at multiple levels. Observers can represent items within a display as individuals, or they can combine/collapse over all individuals to represent them as an ensemble. The latter form of representation extracts statistical information about the overall group at the expense of knowledge about individuals. The use of such “ensemble” representation has been shown to enhance visual cognition, since it allows observers to overcome limitations in encoding by computing general scene statistics, such as mean color, mean location, mean size, and so on (for a review, see Alvarez, 2011). Brady and Tenenbaum (2013) have suggested that representing displays at the ensemble level may prove advantageous in performing one-shot change detection tasks. This may be especially true for trials consisting of larger set sizes in a whole-display task. Instead of loading individual items into VWM and comparing them to stimuli presented on the screen, observers may rely on detecting a change in general gist across the memory and test arrays. Providing only a single-item probe (Exps. 2a–2b) or asking observers to click on the suspected changing target (Exp. 3) may increase item-based versus gist-based encoding. In future work, using a whole-display with clicks could be valuable for investigating hierarchically organized memory representations. More generally, identifying the additional non-storage-related factors that come into play during the one-shot task appears to be a valuable avenue for future research.

Measuring multiple abilities seems possible within the flicker task. For example, we had some success in the present experiments with trying to estimate non-search-related activity and controlling for it or correlating it with other measures (e.g., Exp. 5). The non-search-related activity term in Eq. 3b ( $RT_{\text{nsrch}}$ ) is only a first, coarse, example of an approach that could lead to more refined models of task performance that would take into account additional factors. Such approaches might allow scientists to better investigate the interaction of VWM storage and executive control processes within a single paradigm.

Despite the potential that the flicker paradigm holds for investigating issues in VWM, much work will be needed prior to hailing it as a complete and refined paradigm. For example, the present method of measuring the non-search-related activity lumps various abilities and executive control processes into a single term. Future attempts should focus on incorporating other executive control factors that may contribute to RTs and determining a more elegant way of measuring these components. Eyetracking experiments may prove helpful, in this regard. Determining the amount of time that observers spend gathering a gist before initiating search can be used to generate a more principled equation that could better model performance in the flicker task. Eyetracking experiments may also prove helpful in understanding how observers search during the flicker task. Is search truly random without replacement, or are observers frequently returning to previously searched locations? How long do observers remain fixated on nontargets during their search? Studies of these behaviors may aid in creating more refined models of visual search, VWM storage, and non-search-related activity in the flicker task.

Though we have demonstrated a strong relationship between estimates of storage capacity measured by the flicker and one-shot tasks (Exp. 5), further work would prove helpful in establishing the construct validity of the flicker paradigm. For example, demonstrating the stability of individual estimates over longer periods of time, such as over the course of a year, would demonstrate the task's ability to capture an invariant capacity limit. Comparisons between flicker  $K$  values and measures of verbal WM storage capacity could be used to demonstrate the domain specificity of VWM storage capacity that flicker aims to capture. Nonetheless, the good psychometric properties of the flicker task suggest that this paradigm may be a very powerful tool for conducting research examining individual differences. In this vein, we have used the flicker task to demonstrate the heritability of VWM storage capacity in monozygotic and dizygotic twins (Wilmer et al., 2012) and to track the trajectory of VWM storage capacity across development (Pailian et al., 2014).

The flicker paradigm capitalizes on the dynamic nature of VWM and active visual search. In real-world contexts, the use of VWM includes both storage- and non-storage-related processes. The flicker paradigm is one approach for investigating the variety of these psychological factors, and it holds much promise for allowing researchers to answer novel questions.

## References

- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15, 122–131. doi:10.1016/j.tics.2011.01.003
- 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, 106–111. doi:10.1111/j.0963-7214.2004.01502006.x
- Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences*, 108, 10367–10371. doi:10.1073/pnas.1104047108
- Anderson, D. E., Vogel, E. K., & Awh, E. (2013). A common discrete resource for visual working memory and visual search. *Psychological Science*, 24, 929–938. doi:10.1177/0956797612464380
- Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, 18, 622–628. doi:10.1111/j.1467-9280.2007.01949.x
- Baddeley, A. (1986). *Working memory*. Oxford, UK: Oxford University Press, Clarendon Press.
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4, 417–423. doi:10.1016/S1364-6613(00)01538-2
- Baddeley, A. (2012). Working memory: Theories, models, and controversy. *Annual Review of Psychology*, 63, 1–29. doi:10.1146/annurev-psych-120710-100422
- Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press.
- Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14, 575–589. doi:10.1016/S0022-5371(75)80045-4
- Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, 321, 851–854. doi:10.1126/science.1158023
- Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22, 384–392. doi:10.1177/0956797610397956
- Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual short-term memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, 138, 487–502. doi:10.1037/a0016797
- 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. doi:10.1167/11.5.4
- 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, 85–109. doi:10.1037/a0030779
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322. doi:10.1111/j.2044-8295.1910.tb00207.x
- Cave, K. R., & Wolfe, J. M. (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology*, 22, 225–271. doi:10.1016/0010-0285(90)90017-X
- Conway, A. R. A., Cowan, N., Bunting, M. F., Theriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, 30, 163–183. doi:10.1016/S0160-2896(01)00096-4
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87–114. doi:10.1017/S0140525X01003922. disc. 114–185.
- Cowan, N., AuBuchon, A. M., Gilchrist, A. L., Ricker, T. J., & Saults, J. S. (2011). Age differences in visual working memory capacity: Not based on encoding limitations. *Developmental Science*, 14, 1066–1074.

- 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*, 42–100. doi:10.1016/j.cogpsych.2004.12.001
- Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition*, *34*, 1754–1768.
- Cowan, N., Morey, C. C., AuBuchon, A. M., Zwilling, C. E., & Gilchrist, A. L. (2010). Seven-year-olds allocate attention like adults unless working memory is overloaded. *Developmental Science*, *13*, 120–133. doi:10.1111/j.1467-7687.2009.00864.x
- Diamantopoulou, S., Poom, L., Klaver, P., & Talsma, D. (2011). Visual working memory capacity and stimulus categories: A behavioral and electrophysiological investigation. *Experimental Brain Research*, *209*, 501–513. doi:10.1007/s00221-011-2536-z
- Fukuda, K., & Vogel, E. K. (2011). Individual differences in recovery time from attentional capture. *Psychological Science*, *22*, 361–368.
- 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*, 673–679. doi:10.3758/17.5.673
- Gmeindl, L., Jefferies, L. N., & Yantis, S. (2011). *Individual differences in the ability to restrict the breadth of attention are correlated with visuospatial working memory capacity*. Poster presented at the 2011 Conference of the Vision Sciences Society, Naples, Florida.
- Halberda, J., Simons, S. J., Pailian, H. P., & Wetherhold, J. (2014). *The flicker paradigm provides converging evidence for a three-item limit of visual working memory*. Manuscript submitted for publication.
- Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapped sets can be enumerated in parallel. *Psychological Science*, *17*, 572–576.
- Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, *357*, 575–577. doi:10.1038/29068
- Horowitz, T. S., & Wolfe, J. M. (2003). Memory for rejected distractors in visual search? *Visual Cognition*, *10*, 257–298. doi:10.1080/13506280143000005
- Hyun, J.-S., Woodman, G. F., Vogel, E. 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*, 1140–1160. doi:10.1037/a0015019
- Johnson, N. L., & Kotz, S. (1977). *Urn models and their applications*. New York, NY: Wiley.
- Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.
- Kyllingsbæk, S., & Bundesen, C. (2009). Changing change detection: Improving the reliability of measures of visual short-term memory capacity. *Psychonomic Bulletin & Review*, *16*, 1000–1010. doi:10.3758/PBR.16.6.1000
- 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.
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281. doi:10.1038/36846
- 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*, 1466–1479. doi:10.1037/a0020851
- Matsuyoshi, D., Osaka, M., & Osaka, N. (2014). Age and individual differences in visual working memory deficit induced by overload. *Frontiers in Psychology*, *5*(384), 1–7. doi:10.3389/fpsyg.2014.00384
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97. doi:10.1037/h0043158
- Pailian, H., Libertus, M. E., Feigenson, L., & Halberda, J. (2014). *Visual working memory storage capacity increases between ages 3 and 8 years controlling for gains in attentional and executive control in a visual search paradigm*. Manuscript submitted for publication.
- Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*, 369–378. doi:10.3758/BF03210419
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442. doi:10.1163/156856897X00366
- Perez, V. B., & Vogel, E. K. (2012). What ERPs can tell us about working memory. In S. J. Luck & E. S. Kappenman (Eds.), *Oxford handbook of event-related potential components* (pp. 361–372). New York, NY: Oxford University Press.
- Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193–198.
- Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, *16*, 283–290. doi:10.3758/BF03203943
- Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 179–197. doi:10.1163/156856888X00122
- Rensink, R. A. (2000). Visual search for change: A probe into the nature of attentional processing. *Visual Cognition*, *7*, 345–376.
- 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. doi:10.1111/j.1467-9280.1997.tb00427.x
- Rensink, R. A., O'Regan, J. K., & Clark, J. J. (2000). On the failure to detect changes in scenes across brief interruptions. *Visual Cognition*, *7*, 127–145.
- Riggs, K. J., McTaggart, J., Simpson, A., & Freeman, R. P. J. (2006). Changes in the capacity of visual working memory in 5- to 10-year-olds. *Journal of Experimental Child Psychology*, *95*, 18–26.
- Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*, 5975–5979. doi:10.1073/pnas.0711295105
- Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review*, *18*, 324–330. doi:10.3758/s13423-011-0055-3
- Schmidt, B. K., Vogel, E. K., Woodman, G. F., & Luck, S. J. (2002). Voluntary and automatic attentional control of visual working memory. *Perception & Psychophysics*, *64*, 754–763. doi:10.3758/BF03194742
- Simmering, V. R. (2012). The development of visual working memory capacity during early childhood. *Journal of Experimental Child Psychology*, *111*, 695–707.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295. doi:10.1111/j.2044-8295.1910.tb00206.x
- Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*(6984), 751–754.
- Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective, & Behavioral Neuroscience*, *5*, 144–155. doi:10.3758/CABN.5.2.144

- Treisman, A. (1993). The perception of features and objects. In A. D. Baddeley (Ed.), *Attention: Selection, awareness, and control—A tribute to Donald Broadbent* (pp. 5–35). Oxford, UK: Oxford University Press, Clarendon Press.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136. doi:10.1016/0010-0285(80)90005-5
- 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*, 173–176.
- Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*, 748–751. doi:10.1038/nature02447
- Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures of individual differences in controlling access to working memory. *Nature*, *438*, 500–503. doi:10.1038/nature04171
- 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. doi:10.1037/0096-1523.27.1.92
- 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*, 1436–1451. doi:10.1037/0096-1523.32.6.1436
- Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11:1120–1135. doi:10.1167/4.12.11
- Wilmer, J. B., Germine, L., Ly, R., Hartshorne, J. K., Kwok, H., Pailian, H., . . . Halberda, J. (2012, May). *The heritability and specificity of change detection ability*. Paper presented at the 2012 Conference of the Vision Sciences Society, Naples, Florida.
- Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*, 202–238. doi:10.3758/BF03200774
- Xu, Y., & Chun, M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*, 91–95. doi:10.1038/nature04262