Published in final edited form as:

J Exp Psychol Learn Mem Cogn. 2010 November ; 36(6): 1466–1479. doi:10.1037/a0020851.

Method matters: Systematic effects of testing procedure on visual working memory sensitivity

Tal Makovski, Leah M. Watson, Wilma Koutstaal, and Yuhong V. Jiang

Department of Psychology, Center for Cognitive Sciences, University of Minnesota

Abstract

Visual working memory (WM) is traditionally considered a robust form of visual representation that survives changes in object motion, observer's position, and other visual transients. This study presents data that are inconsistent with the traditional view. We show that memory sensitivity is dramatically influenced by small variations in the testing procedure, supporting the idea that representations in visual WM are susceptible to interference from testing. In this study, participants were shown an array of colors to remember. After a short retention interval, memory for one of the items was tested with either a same-different task or a 2-alternative-forced-choice (2AFC) task. Memory sensitivity was much lower in the 2AFC task than in the same-different task. This difference was found regardless of encoding similarity or whether visual WM required a fine memory resolution or a coarse resolution. The 2AFC disadvantage was reduced when participants were informed shortly before testing which item would be probed. The 2AFC disadvantage diminished in perceptual tasks and was not found in tasks probing visual long-term memory. These results support memory models that acknowledge the labile nature of visual WM, and have implications for the format of visual WM and its assessment.

Keywords

Visual working memory; change detection; 2-alternative-forced choice; same-different; visual long-term memory

Visual working memory (WM) allows people to hold visual information in mind for a few seconds after its disappearance (Baddeley, 1986; Hollingworth, Richard, & Luck, 2008). Research in the past decade has tried to clarify capacity limits of visual WM (how much information can be stored?) and its fidelity (how precise is the representation?), with continued debate about the principal characteristics (Jiang, Makovski, & Shim, 2009). Most of this research, however, has assumed that visual WM is a robust storage system whose representation survives changes in object motion, observer's position, and other visual transients. Few studies have addressed the vulnerability of visual WM to subsequent input as well as the corresponding methodological implications. The present study shows that visual WM is not an entirely stable format of representation. We demonstrate that slight variations of testing procedure significantly alter the observed memory sensitivity in visual WM tasks in

Correspondence should be addressed to Tal Makovski, S214 Elliott Hall, 75 East River Road, Minneapolis, MN 55455. tal.makovski@gmail.com.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/xlm.

ways unpredicted by previous research in psychophysics and long-term memory. These findings have significant implications for the format and stability of representation in visual WM.

The stability of visual WM

In his seminal work dissociating visual WM from iconic memory, Phillips (1974) showed that whereas iconic memory was severely impaired by changes in spatial location between a memory and a test array, visual short-term memory for the arrays survived this shift. Later research showed that performance in visual short-term memory tasks was not reduced when the test array was enlarged or shrunk compared with the memory array (Jiang, Olson, & Chun, 2000), suggesting that the memory representation is abstracted from the array's retinal positions. In addition, neurophysiological studies on monkeys have shown that cells in the prefrontal cortex maintain their stimulus selectivity during the delay interval of a working memory task, even when the interval was filled with additional visual input (Miller, Erickson, & Desimone, 1996). These findings have led to the dominant view of visual working memory as a robust memory representation that survives changes in object motion, observer's position, and other visual transients (Potter & Jiang, 2009).

However, recent evidence suggests that, although visual WM is less vulnerable to interference than iconic memory is, its representation is not entirely stable. The presence of additional visual input during the retention interval disrupts memory performance, even when the visual input is task irrelevant (Makovski, Shim, & Jiang, 2006). It appears that memory representation for an array of objects can be maintained for a period of time with high capacity and fidelity (Landman, Spekreijse & Lamme, 2003; Sligte, Scholte & Lamme, 2008) *until* new visual input is presented. The new input produces significant interference on the initial memory representation (Griffin & Nobre, 2003; Makovski, Sussman, & Jiang, 2008; Sligte et al., 2008). These data have prompted new proposals about visual WM that divide this system into two stages (Sligte et al., 2008) or two “modes” (Makovski & Jiang, 2007; Makovski, et al., 2008): a labile stage that holds a relatively large number of items, and a robust stage that holds a small number of items.

These new proposals, however, had limited impact on ongoing research in visual WM. The majority of visual WM studies continue to adopt the traditional view of visual WM as a robust storage system, which shapes the kind of theoretical questions asked (“Is this store limited by slots or resources”, rather than “how is the content of visual WM updated”) and the methodology used to probe memory (standard change detection procedure, rather than an alternative procedure that minimizes interference). The current study further challenges the traditional view. Specifically, we approach the question by investigating the effect of slight variations in testing procedure on visual WM.

Testing visual WM

One of the most common procedures used to probe visual WM is the change detection task (Luck & Vogel, 1997; Pashler, 1988; Phillips, 1974; Rensink, 2002). In this task, participants are shown an array of visual objects to encode. After a short retention interval (typically 1-2 seconds), a test array is presented for participants to make a same-different response. Visual WM studies typically use two versions of the change-detection task. In the whole-display version, the test array is either the same as the memory array or one object is changed. In the single-probe version, either a single object is presented or the entire array is presented but one object is highlighted. Participants must report whether the target object is the same or different from the object previously occupying that location (Jiang, et al., 2000; Luck & Vogel, 1997). Accuracy in reporting the presence or absence of a change is used to estimate visual WM capacity (Cowan, 2001; Pashler, 1988; Phillips, 1974). The two versions of the change-

detection task typically provide comparable estimates of memory capacity (Luck & Vogel, 1997; Makovski, et al., 2008).

Although most studies of visual WM use the same-different task, some studies have used different testing procedures. For instance, Zhang and Luck (2008) probed participants' color memory by asking them to report their response by clicking on a color wheel (see also Wilken & Ma, 2004). In another study, Jiang, Shim and Makovski (2008) used a two-alternative-forced-choice (2AFC) procedure, where two memory probes were presented at the location of a previously encoded object. Participants were asked to select the alternative that was the same as the memorized object. In a third study, Bays and Husain (2008) asked participants to report the direction in which a probe was displaced or rotated in relation to the memorized objects. Other studies have used multiple testing procedures to assess visual WM but have failed to reveal an effect of testing procedure on memory (Gold, Robinson, McMahon, Braun, & Luck, 2006; Wilken & Ma, 2004; Zhang & Luck, 2008).

Are the different testing procedures equal in measuring visual WM? The answer is likely no. For example, Kyllingsbæk and Bundesen (2009) showed that when participants were allowed to respond “don't know” in addition to “same” and “different”, the variability in the capacity estimates of visual WM was reduced. Nonetheless, there are good reasons to believe that visual WM sensitivity is largely independent of the way it is probed. First, the estimated visual WM is comparable whether the estimates are based on a whole-display or a single-probe procedure (Jiang et al., 2000; Luck & Vogel, 1997). Second, signal detection theory suggests that the way sensitivity (i.e., d') is measured is in general independent of the testing procedure, be it a same-different task or a 2AFC task (MacMillan & Creelman, 1991). This idea has been further supported by studies of long-term memory (Bayley, Wixted, Hopkins & Squire, 2008; Khoe, Kroll, Yonelinas, Dobbins, & Knight, 2000), which showed comparable d' in same-different (often termed the *yes/no* task in long-term memory studies) and 2AFC tasks in normal adults. To our knowledge, however, no studies have directly evaluated whether the assumption of invariant memory sensitivity (d') holds true for visual WM.

Different predictions are made by the traditional model of visual WM as a robust storage system and the more recent proposals of visual WM as comprised of an initial, labile stage of representation. If visual WM is a robust storage system, then memory sensitivity should be relatively invariant to the way it is probed. In contrast, if visual WM is vulnerable to interference and if different testing procedures introduce different amounts of interference, then memory sensitivity should differ across different testing procedures.

Current study: Same-different versus 2AFC

In this study we compared visual WM sensitivity in two procedures: same-different and 2AFC. We selected these procedures because the same-different task was commonly used in visual WM tasks, and the 2AFC task is comparable to it in terms of expected accuracy under chance (50%), task requirements (to remember and compare a memory stimulus to test items), and display sequence. Following studies on long-term memory (e.g., Bayley et al., 2008; Khoe et al., 2000), we used d' as our primary index of memory strength. d' was calculated based on signal detection theory in the same-different task (MacMillan & Creelman, 1991), and was derived from the M-alternative forced-choice conversion table of Hacker and Ratcliff (1979) in the 2AFC task [footnote¹]. Compared with the same-different procedure, the 2AFC procedure is typically considered a high-performance procedure because it is less likely to involve response bias (MacMillan & Creelman, 1991). That is, the same d' corresponds to higher accuracy in the 2AFC procedure than in the same-different procedure. For instance, a d' of 1.5 corresponds to about 86% accuracy in a 2AFC procedure and about 77.5% in a same-different procedure (Macmillan & Creelman, 1991).

Three competing hypotheses were tested. First, if the underlying memory sensitivity in visual WM is independent of the way it is tested (MacMillan & Creelman, 1991), then the same-different and 2AFC procedures should yield comparable d' . Alternatively, the 2AFC procedure may yield higher d' given that more information is provided at the time of testing. Specifically, participants have access to two sources of information in the 2AFC procedure (an old probe that matches previous memory, and a new probe). Participants can make a correct response either when they are certain that probe 1 is the same as their memory, or when they are certain that probe 2 was not presented earlier. In contrast, the same-different procedure provides only one source of information. In addition, from the perspective of long-term memory, it has been suggested that the underlying mnemonic mechanisms may differ between the same-different and 2AFC procedures. Whereas the same-different procedure may be supported primarily by recollection, both familiarity and recollection can contribute to performance in the 2AFC task (Bastin & Van der Linden, 2003; Parkin, Yeomans, & Bindschaedler, 1994; Westerberg et al., 2006). A correct response in the 2AFC task can be generated if participants think that they are more familiar with probe 1 than probe 2. This kind of familiarity comparison is largely circumvented in a same-different task. In turn, one may expect that visual WM sensitivity would be higher when probed with the 2AFC task than the same-different task. This is especially true if the two test probes are similar to each other, in which case the use of relative familiarity should lead to a significant advantage in the 2AFC task compared with the same-different task (Migo, Montaldi, Norman, Quamme, & Mayes, 2009).

A third possibility is that the 2AFC task can result in a disadvantage compared with the same-different task when testing visual WM. This is because whereas only a single testing stimulus needs to be attended and evaluated in the same-different procedure, two testing stimuli are attended and evaluated in the 2AFC procedure. Previous studies have shown that unlike visual long-term memory or visual perception, the representation of information in visual WM is fragile (Sligte et al., 2008). This representation is susceptible to interference from subsequent visual input, including that of test probes (Landman, et al., 2003; Makovski & Jiang, 2007). The interference derives both from the mere presentation of visual input (bottom-up interference) and from the need to attend to it (top-down interference; Makovski et al., 2006; Makovski et al., 2008). The need to attend to more stimuli during testing may result in greater visual WM interference in the 2AFC than in the same-different tasks. Consequently, memory sensitivity may be lower in the 2AFC task than in the same-different task. The following experiments are designed to test these hypotheses.

Experiment 1

This experiment compares memory sensitivity, as measured by d' , in two visual WM testing procedures. Participants performed two blocks of a change detection task on color arrays. In the same-different block, they determined whether a probe item matched a memory item. In the 2AFC block, they selected which one of two probes matched a memory item. The two blocks were otherwise identical. To control for the number of the test items appearing on the test display, we included two versions of the same-different block: In Experiment 1a, the test display consisted of a single item, whereas in Experiment 1b, two identical items appeared in the test. In addition, we equated the magnitude of the memory-to-probe change between the

¹We did not use A' or Cowan's K because these other measures require one to recode data into misses and false alarms. In 2AFC tasks, a "miss" is also a "false alarm" of the alternative choice, so the recoding of data into misses and false alarms is debatable and often requires a correction. In contrast, the M-AFC table of Hacker & Ratcliff (1979) allows us to convert percent correct in a 2AFC task into d' . Nonetheless, to ensure that our results do not depend on the choice of d' as the dependent measure, we have conducted analyses on all experiments using an alternative d' measure. In these analyses, we artificially coded the left button press as "signal+noise" (a correct left button response is coded as a hit, and if the correct response should have been left and the subject pressed right, that's coded as a miss) and the right button press as "noise". This allowed us to calculate hits and false alarms which were then used to estimate d' as $(Z(\text{hit}) - Z(\text{false alarm})/\sqrt{2})$. Importantly the results were consistent with results from the d' measure reported below.

two procedures: The foil used in the 2AFC task was the same as the “different” probe used in the same-different task. We used foils that were highly similar to the memory item, because the impact of familiarity-based judgment may be greater when the two test probes are more similar (Migo et al., 2009).

A secondary interest of this experiment is the generality of the findings for visual WM tasks involving items with high or low similarity. Recent studies have shown that unlike verbal WM, increased similarity in the to-be-encoded items results in better performance in visual WM tasks (Johnson, Spencer, Luck, & Schoner, 2009; Lin & Luck, 2009; Mate & Baques, 2009). Lin and Luck (2009) suggested that the representation of highly similar visual items may be more stable than that of dissimilar items. Because different testing procedures may be differentially affected by the stability of WM representation, we deemed it important to evaluate the potential interaction between testing procedure and encoding similarity. Therefore, in both the same-different and 2AFC tasks, we tested memory items that were either highly similar to one another (items were drawn from the same general color category; *high-encoding-similarity* condition), or dissimilar to one another (items were drawn from different color categories; *low-encoding-similarity* condition).

Method

Participants—Participants in all experiments were students from the University of Minnesota. They were 18 to 35 years old and had normal color vision and normal or corrected-to-normal visual acuity. Participants received course credits or \$10/hour for their time. There were 32 participants in Experiment 1 with a mean age of 19.6 years. Half completed Experiment 1a and the other half Experiment 1b.

Equipment—In all experiments, participants were tested individually in a dimly lit room. They sat 40 cm away from a 19" CRT monitor (resolution 1280 × 1024 pixels; 75 Hz). Viewing distance was stabilized with a chinrest. The experiments were programmed with Psychophysical toolbox (Brainard, 1997; Pelli, 1997) implemented in MATLAB.

Materials and Trial sequence—Each trial began with a white fixation circle (0.4° in diameter) presented against a black background. Half a second later the memory display was presented for 500 ms. After a blank retention interval of 1000 ms, the test display was presented until participants made a response. Participants were asked to respond as accurately as possible and feedback in the form of a word (“correct” or “wrong”) was presented after each response and lasted 400 ms. To reduce verbal recoding, participants were required to repeat a three-letter word out aloud throughout a trial. The three-letter word was specified at the beginning of each block of 20 trials. Figure 1 (left) illustrates the trial sequence.

The memory display contained three unique color squares (1.6°×1.6°), placed equidistantly on an imaginary circle (radius = 3.2°) centered at fixation. The colors were drawn from three categories (red, green, and blue) and there were 8 different shades in each category (Figure 1, right). In the *low-encoding-similarity* condition, the three memory colors were drawn from different categories and the exact shade was randomly chosen. In the *high-encoding-similarity* condition, all three memory items were pseudo-randomly selected, without replacement, from a single color category.

Two-alternative-forced-choice (2AFC)—The test display in the 2AFC procedure consisted of two color squares presented side-by-side. The squares were centered at one of the previous memory locations; their center-to-center distance was 1.84°. One test probe was identical to the memory color at that location, and the other (foil) was a similar color drawn from the same color category but four steps away in the similarity continuum (see Figure 1,

right). Participants were asked to determine whether the left or right item was the same as the memory item by pressing the “s” (left) or “d” (right) keys.

Same-different—The same-different task of Experiment 1a consisted of a single color square at the location of one of the memory items. Participants were asked to determine whether the single probe was the same or different from the memory item previously displayed at that location, by pressing the “s” (same) or “d” (different) keys. On half of the trials, the probe was the same as the memory color at that location. On the other half of the trials, the probe was a color from the same category as the memory item but four steps away in the similarity continuum.

Experiment 1b equated the number of probe items presented during testing for the same-different and 2AFC procedures. Similar to the 2AFC task, two test items were presented on the probe display in the same-different task (1.84° center-to-center distance). However, the two items were identical to each other and participants were again asked to judge whether the probes were the same or different as the item previously presented there.

Procedure and Design—Participants were tested in two blocks of trials, one involving the 2AFC task and the other the same-different task. The order of the tasks was counterbalanced across participants. Each block consisted of 20 practice trials and 180 experimental trials. The experimental trials were divided randomly and evenly into two encoding conditions (*high-encoding-similarity* and *low-encoding-similarity*), three probe categories (red, blue, or green), and two responses (same/different, or left/right). Experiments 1a and 1b were the same except for the use of two identical probes in the same-different task of Experiment 1b.

Data analysis—We used d' as the primary dependent variable because it is a bias-free estimate of memory sensitivity. In the same-different procedure, d' was calculated as the difference between $z(\text{hit})$ and $z(\text{false alarm})$ (Green & Swets, 1966). In the few cases where hit rate or correct rejection rate reached 100% they were treated as 99% (e.g., Bayley et al., 2008). In the 2AFC procedure, d' was calculated by converting accuracy into d' using Hacker and Ratcliff's (1979) M-AFC conversion table (MacMillan & Creelman, 1991; Migo et al., 2009). We list accuracy data in the Appendix.

Results

Figure 2 plots the d' results averaged across participants as a function of testing procedure and encoding similarity. We conducted a mixed-factor ANOVA using experiment (1a or 1b) as a between-subject factor, and testing procedure (2AFC or same-different) and encoding similarity (low- or high- encoding similarity) as within-subject factors. This analysis revealed a significant main effect of testing procedure, as d' was significantly higher in the same-different procedure ($M = 1.54$) than the 2AFC procedure ($M = 0.96$), $F(1, 30) = 91.80$, $p < .01$, $\eta_p^2 = .75$, Cohen's $d = 1.6$. Testing procedure did not interact with encoding similarity, $F < 1$, or with experimental version, $F(1, 30) = 1.36$, $p > .25$. The main effect of experimental version was not significant, $F < 1$.

Consistent with recent findings (Johnson et al., 2009; Lin & Luck, 2009; Mate & Baques, 2009), memory sensitivity was higher when the encoding items were similar than dissimilar, $F(1, 30) = 13.92$, $p < .01$, $\eta_p^2 = .32$. However, there was a marginally significant interaction between experiment and encoding similarity, $F(1, 30) = 3.95$, $p = .057$, $\eta_p^2 = .12$, as the effect of encoding similarity was stronger in Experiment 1b than 1a. The three-way interaction was not significant, $F(1, 30) = 3.38$, $p = .076$, $\eta_p^2 = .10$. A recent study has suggested that the encoding similarity effect might be the result of a homogeneity signal, computed on a trial-by-trial basis that alters subjects' decision criterion (Viswanathan, Perl, Visscher, Kahana &

Sekuler, 2010). Partly in line with this argument we found that not only was d' increased in the *high-encoding-similarity* condition, but response bias also differed, reflecting a greater tendency to respond “different” in the condition with greater encoding similarity (0.22 vs. 0.05, $F(1, 30) = 9.69$, $p < .01$, $\eta_p^2 = .24$).

Given that each participant was tested in both the same-different and 2AFC tasks, there might have been strategy shifts from one task to the next. To ensure that the above results were not produced by carryover effects across blocks, we focused on the first block of data, at which point participants were ignorant of the next task. We compared data from the 16 participants who did the same-different task as their first task, with data from the other 16 participants who did the 2AFC task as their first task. This analysis confirmed that d' was significantly lower in the 2AFC task ($d' = 0.85$) than in the same-different task ($d' = 1.54$), $F(1, 30) = 30.58$, $p < .01$, $\eta_p^2 = .51$.

Was the disadvantage in the 2AFC task a result of slower responses (and thus potentially longer retention interval) in this procedure? To address this issue, we compared RT data for all trials with correct responses in the two procedures. This analysis showed that RT was comparable between the 2AFC ($M = 975$ ms) and the same-different tasks ($M = 971$ ms), $F < 1$. Thus, a speed-accuracy tradeoff cannot account for the differential effects of testing procedure on visual WM sensitivity observed in this experiment. We will not present RT data in subsequent experiments because we also did not observe consistent speed-accuracy tradeoffs in subsequent experiments, and because the tasks were unspeeded.

Discussion

Experiment 1 showed that visual WM sensitivity was much lower when measured in a 2AFC procedure than a same-different procedure. These results are inconsistent with the commonly accepted notion of invariant sensitivity for different testing procedures in psychophysics (e.g., Bayley et al., 2008; Khoe et al., 2000; MacMillan & Creelman, 1991, but see Yeshurun, Carrasco & Maloney, 2008). In addition, the results are the opposite of what has been reported in some studies of long-term memory, where sensitivity can be higher in the 2AFC task than the same-different task, especially when the former is accomplished on the basis of familiarity judgment (Migo et al., 2009).

The impaired memory sensitivity associated with the 2AFC task is consistent with the third hypothesis laid out above. Specifically, because representation in visual WM is unstable, it is highly vulnerable to interference from testing stimuli (Landman, et al., 2003; Makovski & Jiang, 2007; Makovski et al., 2008; Sligte et al., 2008). Since there are more testing stimuli to attend to and evaluate in the 2AFC procedure, there is more interference during testing in the 2AFC task than in the same-different task. Interference from the testing array reduces memory sensitivity measured in 2AFC. This interpretation also fits with the notion that visual WM closely interacts with perception. In a recent computational model, Johnson et al. (2009) postulated that visual WM and perception share common inhibitory fields. Activation in visual WM affects the way external input is perceived, and vice versa. While Experiment 1 was not aimed to test this model, its results are in agreement with the model's assumption that visual WM interacts closely with the delivery of new visual input (see also Soto, Wriglesworth, Bahrami-Balani & Humphreys, 2010).

How much of the 2AFC impairment was due to bottom-up stimulus differences, as opposed to top-down task differences between 2AFC and same-different procedures? In Experiment 1a the same-different task involved one probe item, whereas the 2AFC task involved two probe items. This difference in visual properties was reduced in Experiment 1b where we presented two identical probes in the same-different task. Results were similar in the two experiments, suggesting that low-level differences in stimulus properties could not account for our data. To

further reduce stimulus differences during testing, in a follow-up experiment we presented participants with two different probes in both the 2AFC and the same-different tasks. In the same-different task, we asked participants to determine whether the left probe (or the right probe, for half of the participants) was the same or different from the memory item. In the 2AFC task they determined which probe was the same as the memory item. In this procedure, the visual input was identical for the two testing procedures, but the tasks were different. We replicated results from Experiment 1 in that d' was significantly higher in the same-different task ($M = 1.29$) than in the 2AFC task ($M = 0.87$), $p < .01$. This confirms that task differences, rather than stimulus differences, underlie the disadvantage of the 2AFC task compared with the same-different task.

Finally, Experiment 1 showed higher d' for the *high-encoding-similarity* condition than the *low-encoding-similarity* condition, replicating recent findings (Johnson et al., 2009; Lin & Luck, 2009; Mate & Baques, 2009). This difference may partly arise from relational encoding of memory items that makes it easier to chunk similar items than dissimilar ones. It was suggested that the advantage in the *high-similarity* condition was due to more stable representation of highly similar items (Lin & Luck, 2009), in which case one might expect to see a reduced effect of testing procedure in the *high-encoding-similarity* condition compared with the *low-similarity-condition*. However, we did not find an interaction between encoding similarity and testing procedure. This may be attributed to the relatively small effect of encoding similarity on memory sensitivity (Lin & Luck, 2009), making it difficult to observe higher-order interaction effects. In the experiment that showed a bigger effect of encoding similarity (Experiment 1a), there was a numerical trend towards a reduced difference between the 2AFC and same-different tasks in the *high-encoding-similarity* condition. Future studies that show stronger effects of encoding similarity are needed to examine whether visual WM representation is more stable for more similar memory items.

Experiment 2

We designed Experiment 2 for two reasons. First, because results from Experiment 1 were unexpected based on prior psychophysical and long-term memory results (Bayley et al., 2008; Khoe et al., 2000; MacMillan & Creelman, 1991; Migo et al., 2009), we deemed it necessary to establish the robustness and generality of the finding. One special property of the design of Experiment 1 was that the foils used during testing were highly similar to the memory item. Perhaps the high similarity between the two probes presented at the time of testing had prompted greater attention to the probes in the 2AFC task, leading to the interference observed in Experiment 1. If this is the case, then the use of less similar foils may eliminate differences between 2AFC and same-different procedures. Alternatively, visual WM representation may be volatile and disrupted whenever there are two rather than one testing probes to attend to. If this is true, then increasing the dissimilarity between the two probes used in the 2AFC task may not be beneficial.

We therefore replicated Experiment 1a but used foils that were drawn from a different color category than the memory item. This is an important modification because it has been suggested that the ability to detect small, within-category changes may tap into a different characteristic of visual WM (its resolution) than the ability to detect large, between-category changes (its capacity; Barton, Ester & Awh, 2009). We only tested participants using items with low encoding similarity (i.e., all memory items were drawn from different color categories) because the task is trivially easy in the *high-encoding-similarity* condition. In addition, memory load was increased to five for half of the trials to compensate for the reduction in task difficulty.

The second purpose of this experiment was to further evaluate the interference hypothesis by manipulating the robustness of visual WM representation. Previous studies have shown that

whereas visual WM for multiple items is fragile, memory representation for a single item is more robust (Makovski & Jiang, 2007). In particular, if participants are given a cue during the retention interval about which item will be tested, they can focus on the memory of that item and refresh its representation (Griffin & Nobre, 2003; Landman et al., 2003; Lepsien, Griffin, Devlin, & Nobre, 2005; Lepsien & Nobre, 2007; Makovski & Jiang, 2007). This leads to higher performance and greater resistance to interference (Makovski et al., 2008; Matsukura, Luck, & Vecera, 2007; Sligte et al., 2008). The cue is presented after the items have disappeared, therefore it is known as a *retro-cue* (i.e., its impact is retrospective; Griffin & Nobre, 2003). It follows from the interference hypothesis that when a retro-cue is provided before the delivery of the probe items, the impact of the probes should be reduced. In turn, differences between the same-different and 2AFC tasks should be reduced in the *retro-cue* condition compared with the standard *no-cue* condition.

Method

Participants—Sixteen participants (mean age 20.3 years) took part in Experiment 2.

Stimuli, Procedure, & Design—The general method was similar to Experiment 1a except for the following changes. Colors on the memory array were randomly selected, without replacement, from nine distinct colors (orange, red, green, blue, white, yellow, purple, brown, and azure). There were 3 or 5 color squares placed equidistantly on an imaginary circle (radius = 6.0°). The foil used during testing was a color from one of the nine categories and it was not presented on the memory display.

Half of the trials involved the use of a *retro-cue*. On these trials, after the 1000 ms retention interval, a central white arrow (1.6° in length) was presented for 100 ms, followed by a 400 ms blank interval. The test display was then presented (Figure 3 left). Participants were told that the item retrospectively designated by the central arrow would be the one tested. In the *no-cue* condition, the test display appeared immediately after the 1000 ms retention interval (Figure 3 right). Based on the results of previous studies (Griffin & Nobre, 2003; Landman et al., 2003; Makovski et al., 2008; Sligte et al., 2008), we predicted that the *retro-cue* condition would lead to higher memory performance than the *no-cue* condition, despite of the fact that the total duration between encoding and the probe was longer in the former.

Participants completed two blocks of trials, each involving a different testing procedure (same-different or 2AFC; the order was counterbalanced across participants). In each block there were 16 practice trials and 192 experimental trials. The experimental trials were randomly and evenly divided into two memory loads (3 or 5), two cue conditions (*retro-cue* or *no-cue*), and two responses (same/different or left/right).

Results

Figure 4 shows visual WM sensitivity results (d') as a function of testing procedure and memory load, separately for *no-cue* and *retro-cue* trials. A repeated-measures ANOVA with memory load, cue, and testing procedure as factors revealed significant main effects of all three factors. d' was higher when load was lower, $F(1, 15) = 103.17, p < .01, \eta_p^2 = .87$, and when a retro-cue was used, $F(1, 15) = 19.44, p < .01, \eta_p^2 = .56$. Importantly, we replicated Experiment 1's finding that d' was higher in the same-different than in the 2AFC task, even though the two test probes used in the 2AFC task were highly dissimilar, $F(1, 15) = 44.98, p < .01, \eta_p^2 = .75$.

We also observed a significant interaction between the cue condition and testing procedure, $F(1, 15) = 6.42, p < .05, \eta_p^2 = .30$. That is, although the same-different task led to significantly higher sensitivity than did the 2AFC in both the *retro-cue*, $F(1, 15) = 6.23, p < .05, \eta_p^2 = .29$, and the *no-cue* conditions, $F(1, 15) = 64.47, p < .01, \eta_p^2 = .81$, this effect of testing procedure

was significantly smaller in the *retro-cue* condition compared with the *no-cue* condition. No other interaction reached statistical significance, all $ps > .14$.

Discussion

Experiment 2 used large categorical changes in a color change detection task. Unlike Experiment 1 which used small within-category changes, the task did not require fine visual WM resolution (Barton et al., 2009; Jiang et al., 2008). Nonetheless, we replicated the results from Experiment 1. Visual WM sensitivity was much lower when tested in a 2AFC procedure than in a same-different procedure. These results suggest that the 2AFC disadvantage cannot be accounted for by the need to make fine perceptual comparisons between the two similar probes, or the need to make fine memory comparisons between the probes and memory. That is, the finding of a considerable 2AFC cost even when the two test probes are easy to discriminate suggests that the locus of the 2AFC cost is the requirement to make multiple comparisons with a memory representation, rather than the amount of attention required to compare the two test items with each other.

The 2AFC disadvantage was reduced when participants knew, before the delivery of the testing display, which item would be tested. Previous studies have shown that the use of a retro-cue increases the robustness of visual WM representation (Landman et al., 2003; Makovski et al., 2008; Makovski & Jiang, 2007; Matsukura et al., 2007; Sligte et al., 2008). This finding is consistent with the idea that memory performance in the 2AFC procedure is inferior to that observed in the same-different procedure because 2AFC testing causes greater interference with a relatively labile form of visual WM.

Experiment 3

Thus far we have argued that the 2AFC deficit in testing of visual WM is a general phenomenon. It is observed for visual WM tasks that require either coarse or fine memory representation, and for encoding of highly similar or dissimilar items. We have also implied that this finding is specific to testing of visual WM. This latter argument, however, rests upon past studies that may be vastly different from the visual WM task used in our study. The purpose of the next two experiments was to compare 2AFC with same-different tasks in a perceptual color-matching task (Experiment 3) and a visual long-term memory task (Experiment 4). If the 2AFC sensitivity deficit seen in Experiments 1 and 2 indeed originates from visual WM's volatile representation, then this deficit should be attenuated or absent in Experiments 3 and 4, which tap into relatively stable forms of representation in perception and long-term memory.

Method

Participants—Sixteen participants (mean age 19.4 years) completed Experiment 3.

Task 1: 2AFC—Participants viewed three colors: one displayed on top and two at the bottom (each color square subtended $1.6^\circ \times 1.6^\circ$). Their task was to decide which of the two alternatives at the bottom was identical to the top color (10.4° above the fixation) by pressing the left (“s”) or right (“d”) key. On half of the trials the left choice was the same as the top color, while on the other half the right choice was the same. The colors were drawn from the ones used in Experiment 1 (Figure 1 right). One of the choices at the bottom was identical to the top color, and the other belonged to the same color category but was one step away in the color continuum. We used a one-step difference instead of the four-step difference (Experiment 1) to avoid ceiling performance. One choice was shown on the left side of fixation and the other was on the right side. For half of the participants the two choices were 12° apart from each other (*far-display* condition). For the other half of the participants, the two choices were 1.86° apart from each other (*near-display* condition; Figure 5, left). We tested both displays because there was

a concern that when the choices were placed too far away from each other participants would have to hold one choice in memory while examining the other choice (Hollingworth et al., 2008).

Task 2: Same-different—The same-different procedure was similar to the 2AFC task, except that only one of the two choices was presented. Participants were asked to determine whether the choice at the bottom was the same or different from the target color at the top. The choice items were displayed 6° or 0.92° to the left or right of fixation, similar to the *far-display* and *near-display* conditions of the 2AFC task, with half of the participants tested in each condition. On half of the trials the bottom choice was the same as the top color, and on the other half they were different by one step in the color continuum.

Trial sequence, procedure, & design—Each trial began with a white fixation circle (0.4° in diameter) presented against a black background for 500 ms, followed by the test display that was presented until a response was made. Only accuracy was emphasized in the tasks. Similar to Experiments 1 and 2, participants performed the 2AFC and same-different tasks in two separate blocks of trials. The order of the two tasks was counter-balanced across participants. Each task block contained 20 practice and 300 experimental trials. The experimental trials were divided randomly and evenly into three color categories (red, green, or blue) and two correct responses (same/different or left/right).

Results and Discussion

Figure 5 (right) shows mean sensitivity as a function of testing procedure and the test display distance between the two choices. A mixed-factor ANOVA with testing procedure as a within-subject factor and display type as a between-subject factor showed no effects of display type or an interaction between display type and testing procedure, $F_s < 1$. The main effect of testing procedure was significant, revealing lower d' in the 2AFC than in the same-different task, $F(1, 14) = 9.09, p < .01, \eta_p^2 = .39$. However, the disadvantage of the 2AFC task was substantially smaller in this experiment than in the earlier experiments. In the current experiment, d' in the 2AFC task was 17% lower than the d' in the same-different task. In contrast, d' in the 2AFC task was 86% lower than that in the same-different task of Experiment 1a. The interaction between testing procedure and Experiment (Experiment 1a vs. Experiment 3) was significant, $F(1, 30) = 13.75, p < .01, \eta_p^2 = .31$. Performance in the same-different task was comparable between Experiments 1a and 3, $t < 1$, but performance in the 2AFC task was significantly lower in Experiment 1a than in Experiment 3, $t(30) = 5.00, p < .01$.

These results suggest that the interference produced by the 2AFC task was substantially greater in the visual WM task than in the perceptual comparison task. We did not fully eliminate the 2AFC deficit in Experiment 3, possibly because we did not completely remove the contribution of visual WM in the perceptual comparison task. Participants may have needed to fixate on each color to make perceptual comparisons, because color perception is best handled by cone receptors concentrated at the fovea of the retina. Given that the target color and choice colors were more than 10° apart from each other (this was the case even in the near-choice condition), it may have been necessary for participants to remember one color (e.g., the target color) when they fixated on another (e.g., a choice color), engaging visual WM (Hollingworth et al., 2008). Nonetheless, the availability of all items for perceptual analysis increased the robustness of representation and greatly attenuated the cost of 2AFC compared with a same-different task.

Experiment 4

This experiment tested visual long-term memory for objects using same-different and 2AFC tasks. If the 2AFC disadvantage is generally observed in tasks involving visual memory, then

we should replicate this disadvantage in Experiment 4. Alternatively, if the 2AFC disadvantage results from the vulnerability of visual WM to interference from the testing probes, then it should be reduced or eliminated when representation is stable, as is the case in long-term memory tasks.

Method

Participants—Twenty-four participants (mean age 21.7 years) completed this experiment. Half were tested in the same-different procedure and the other half in the 2AFC procedure. We did not use a within-subject design because our stimulus set only permitted a relatively small number of trials per condition.

Materials—We used color photographs (156 categories), drawings of novel abstract objects (24 categories), and drawings of common objects (24 categories), created by the same artists, for a total of 204 categories of visual objects. These objects were categorically distinct, including items like telephones, cheese, dressers, darts, etc. and, for the abstract objects, were highly distinctive between categories and highly similar within categories as shown by previous stimulus normative data (Koutstaal et al., 2003). There were 4 exemplars in each category. The images subtended $6^\circ \times 6^\circ$ and were presented at the center of the screen. Figure 6 (left) shows some examples of the images.

Learning—Stimuli from a randomly sampled half of the categories (78 photographs, 12 abstract art, and 12 drawings) were used during the encoding phase. The randomization was done separately for each participant. Participants were first shown one exemplar from each of 102 categories presented at a rate of 1.5s/item and were told to remember them for a memory test. While viewing the object on the screen, participants were asked to rate the pleasantness of the image on a scale from 1 (extremely unpleasant) to 7 (extremely pleasant). This rating procedure was established in order to encourage participants to attend to the objects. A delay interval was inserted between the learning and testing phases of the experiment. During the delay interval, individuals were tested in an unrelated task involving tracking of moving disks. Typically, a delay period of 40-45 minutes occurred. Subsequently, participants were tested using either the same-different task or the 2AFC task.

Test: Same-different—In the same-different procedure participants were presented with 306 objects, one at a time, and had to determine whether each object was the same (“s” key) or different from (“d” key) one of the memory objects. One third of the objects were old objects (*old*), one third were new objects drawn from a category not presented during encoding (*between-category*), and one third were a different exemplar of the objects the participants saw earlier (*within-category*). The task was unspeeded and each testing object was presented until a response was made. Accuracy feedback in the form of a smiley or sad face icon was presented after each response. The next trial commenced 500 ms later.

Test: 2AFC—The 2AFC task involved the presentation of two objects on each trial. The objects were presented side-by-side (the center of each picture was 6° away from fixation), and one object (left or right, randomly determined) was an old image shown earlier. The other item – the foil – could be one of three types. It could be an object from a new category not shown during encoding (*between-category*; 1/3 of trials), an object from the same category as the old object shown in that test trial (*within-category*; 1/3 of trials), or an object from the same category as an old object shown in other test trials (*other-category*; 1/3 of trials). There were 102 trials in the 2AFC task. The foils used in the *within-category* condition were a different exemplar than those used in the *other-category* condition. The *other-category* condition was included to test hypotheses unrelated to the present study. This condition produced similar results as the *within-category* condition and will not be discussed further. Participants

determined which of the two alternatives was the same as their memory by pressing the left (“s”) or right (“d”) key. Similar to the same-different task, the 2AFC procedure was unspeeded and each response was followed by immediate feedback (happy or sad face icon) regarding accuracy.

Results

In the *same-different* task, we first calculated the hit rate for trials involving the old images. We then separately calculated false alarms involving *between-category* and *within-category* foils. The hit rate for the old images and false alarms for the two types of foils were then used to calculate memory d' for *between-category* and *within-category* memory discrimination. In the 2AFC task, accuracy in each condition was converted into a d' based on Hacker and Ratcliff's (1979) table. Figure 6 (right) shows d' data for *within-* and *between-category* conditions in the 2AFC and the same-different tasks.

An ANOVA on the sensitivity scores, treating testing procedure as a between-subject factor and foil category (within or between) as a within-subject factor revealed no effect of testing procedure, $F(1, 22) < 1$, and no interaction between testing procedure and foil type, $F(1, 22) = 1.68, p > .20$. The main effect of foil category was significant, $F(1, 22) = 45.37, p < .01, \eta_p^2 = .67$, as d' was higher if the foil was drawn from a different category ($M = 2.33$) rather than the same category ($M = 1.55$) as the old object.

Because items were tested one at a time in the same-different task, and were tested pairwise in the 2AFC task, the testing phase was longer in the same-different task (there were 306 trials, as compared with 102 trials in 2AFC). To ensure that the long test did not place the same-different task at a disadvantage, we examined memory sensitivity for the first 102 trials of the same-different task. Results confirmed that d' for the first 102 trials of the same-different task ($d' = 1.56$ in the *within-category* condition and 1.86 in the *between-category* condition) was comparable to that observed in the 2AFC task (Figure 6), $F(1, 22) < 1$.

Discussion

Using a visual long-term memory task, we compared memory sensitivity between a 2AFC task and a same-different task. In contrast to results from the visual WM task, testing procedure did not significantly influence d' in the long-term memory task. If anything, the 2AFC task yielded numerically higher d' than did the same-different task when the foil was highly similar to the old object (see also Migo et al., 2009). This finding suggests that the d' deficit seen in the 2AFC compared with the same-different task is restricted to testing of visual WM.

Whereas increasing similarity among the to-be-encoded items improves change detection performance (Johnson et al., 2009; Lin & Luck, 2009; Mate & Baques, 2009), increasing memory-to-test similarity reduces change detection performance (e.g., Awh, Barton & Vogel, 2007). Similarly, the results of Experiment 4 showed that long-term memory sensitivity was enhanced when memory-to-test similarity was low (*between-category* condition) rather than high (*within-category* condition). The categorical effect (as measured in d') in LTM was more salient than that observed in accuracy in a recent memory study (Brady, Konkle, Alvarez, & Oliva, 2008), suggesting that categorical information plays a significant role in the long-term representation of visual objects.

General Discussion

This study investigated how memory sensitivity in a visual WM task is influenced by the way it is measured. Answers to this question have implications for the way visual WM is conceptualized. Previous research in psychophysics and long-term memory has suggested that

the underlying perceptual or memory sensitivity should be largely independent of the testing procedure (Bayley et al., 2008; Khoe et al., 2000; MacMillan & Creelman, 1991, but see Yeshurun et al., 2008). In contrast, we have demonstrated that memory sensitivity in visual WM tasks is dramatically influenced by small changes in the testing procedure. d' was nearly doubled when estimated with a same-different task than with a 2AFC task. This difference was found for the encoding of highly similar or dissimilar items (Experiment 1), for visual WM tasks that required fine memory resolution (Experiment 1) or coarse resolution (Experiment 2), and was reduced but not eliminated by the use of a retro-cue before testing (Experiment 2). On the other hand, the cost of the 2AFC procedure was relatively small in perceptual comparison (Experiment 3) and was absent in a visual long-term memory task (Experiment 4).

Together these results are inconsistent with the view that visual WM is a robust storage system. Instead, they suggest that WM may be highly susceptible to interference. The need to attend to and process two testing probes in the 2AFC procedure led to greater interference on visual WM representation than that produced by the same-different task. This interference was minimal in a perceptual matching task and was absent on long-term memory representations. Thus, we suggest that these results are consistent with the idea that visual WM representations are unstable and can be influenced by the need to perceive and attend to new visual input (Johnson et al., 2009; Soto et al., 2010). Indeed, the only condition where the 2AFC deficit in the visual WM task was reduced was when a retro-cue was used to stabilize visual WM representation. The 2AFC deficit was observed even when the testing displays for the 2AFC and the same-different tasks were largely equated (Experiment 1b) and fully equated (follow-up Experiment to 1b), therefore the deficit must originate from the particular task demands imposed by the 2AFC format rather than from bottom-up, stimulus differences. That is, the elevated demand of attending and matching two, rather than one, test probes to a memory item can lead to a larger decrease in memory sensitivity (Huang, Treisman, & Pashler, 2007). Thus, the additional probe information provided in the 2AFC procedure that otherwise would have helped participants (Migo et al., 2009) led to more harm than good in testing of visual WM.

Although this study is not the first to emphasize the fragile nature of visual WM (e.g., Landman et al., 2003; Makovski et al., 2008; Sligte et al., 2008), it is the first to demonstrate this point by examining the 2AFC-cost, a finding that sets visual WM apart from visual long-term memory and visual perception. Furthermore, this study extends previous findings by showing that not only is visual WM susceptible to interference from new visual input but is also highly sensitive to “top-down” task demands. This observation opens up the possibility that some of the memory effects found in the past (e.g., between small versus large memory-to-probe similarities; Awh et al., 2007; Barton et al., 2009; Jiang et al., 2008) may in part reflect an effect of testing interference. Moreover, the sensitivity of visual WM to testing procedures and task demands suggests that visual WM is more than a “storage locker”. Central, executive processes are a critical component of visual WM (see also Baddeley, 2000; Makovski et al., 2006; Vergauwe, Barrouillet & Camos, 2010; Vogel, McCollough & Machizawa, 2005).

So far we have interpreted our results in terms of an interference account. But could these data be explained by an alternative account, according to which the differences between 2AFC and same-different tasks were attributed to obligatory encoding of the test probes into visual WM? On this account, given that visual WM has limited storage capacity, if the test probes must be encoded into visual WM to perform the task, then the effective WM load would be the actual load (e.g., 3 colors) *plus* either two in the 2AFC task, or one in the same-different task. The former (a total of 5) would typically exceed memory capacity whereas the latter (a total of 4) would still be within capacity [footnote²]. This hypothesis, however, hinges on the assumption that test items must be encoded into visual WM for the comparison to be made. This assumption does not receive strong empirical support. There is no doubt that the test items must be

attended, but merely attending to an item is not the same as actively remembering it in visual WM (see for example, Makovski & Jiang, 2008a, 2008b; Soto, Heinke, Humphreys & Blanco, 2005; Soto et al., 2010). Furthermore, when an object is perceptually available, people often consult the object perceptually rather than remembering its properties in visual WM (e.g., Ballard, Hayhoe, & Pelz, 1995; Kunar, Flusberg, & Wolfe, 2008), and such a perceptually-based approach could be readily applied to the simple “color patch” stimuli used here. Therefore, it seems unlikely that visual WM’s limited capacity is the source of the difference between two testing procedures as there is no reason to believe that observers had unnecessarily encoded the test stimuli into visual WM. Nevertheless, future research should examine the intriguing possibility that attended items are obligatorily encoded into visual WM.

If, as we have argued, the need to attend to more test probes in the 2AFC condition produced greater interference on visual WM, why did previous studies fail to find a difference between whole-display and single-probe procedures (Jiang et al., 2000; Luck & Vogel, 1997)? The probable reason is that although there are more probes to attend to in the whole-display procedure, the cost is compensated for, or is offset, by the presence of contextual information. It is known that items in visual WM are not represented in isolation, but are represented in relation to one another (Jiang et al., 2000). The cost of making multiple comparisons may be offset by the presence of contextual information for the target item. A second possibility is that unlike 2AFC tasks where the same memory representation is compared with two probes, in whole-display procedures each memory representation is compared with a single test stimulus. The memory may be less disrupted when it is compared with one, rather than two probes.

By rejecting the assumption that all visual WM procedures are similar in indexing memory representations, the present findings have important methodological and substantive implications. On the one hand, the results justify the common use of the same-different procedure (as opposed to 2AFC) in measuring visual WM. This task appears to produce less interference on visual WM representation. On the other hand, it is possible that unlike the same-different procedure, the 2AFC procedure tests a different aspect of visual WM, specifically the *durable* form of visual WM (Slight et al., 2008). Our results also raise the possibility that findings from studies that use different testing procedures may not be directly comparable. Variations in testing procedures of visual WM representation have been common, including the use of 2AFC (Jiang et al., 2008; Makovski & Jiang, 2008b), color wheel (Zhang & Luck, 2008; Wilken & Ma, 2004), and memory-to-test displacement (Bays & Husain, 2008). These procedures likely produce greater interference on visual WM representation than does the same-different task. What is measured in these procedures may be a more durable form of visual WM than what is revealed in same-different tasks. Because the durable format of visual WM may have different characteristics than the more fragile form of visual WM (Makovski & Jiang, 2007; Slight et al., 2008), and because different procedures can tap into different proportions of the durable and fragile WM, caution should be exercised when comparing results from different testing procedures.

Finally, in our experiments we have used d' and the M-alternative-forced-choice version table (Hacker & Ratcliff, 1979) as an index of memory strength. It is important to note that our main finding – a 2AFC disadvantage in visual WM – cannot be attributed to intrinsic differences in the measurement of d' for same-different and 2AFC tasks. These measures – d' and the M-AFC tables for 2AFC tasks – have been validated by previous psychophysics and long-term memory studies (e.g., Bayley et al., 2008; Khoe et al., 2000; Macmillan & Creelman, 1991). The

²We thank Weiwei Zhang for raising this possibility and for suggesting a potential way to rule out the alternative hypothesis. Following Dr. Zhang’s suggestion, we tested participants in an experiment with a memory load of 1 and 2 using procedures similar to those used in Experiment 1a (N=8). The idea is that with such low loads, the addition of 2 probes should still typically be within visual WM capacity, and therefore the alternative account would predict no cost of 2AFC compared with same-different. Our results revealed a significant 2AFC cost even at such low load ($p < .01$ for both load 1 and 2), seemingly inconsistent with the alternative account.

measures did not reveal a 2AFC disadvantage in tasks that did not depend on visual WM (see Experiments 3 & 4), so the disadvantage found in visual WM tasks could not be attributed to intrinsic differences in the measurement of d' for same-different and 2AFC. In addition, an alternative measure of d' (see footnote ¹) produced similar results as the d' derived from Hacker and Ratcliff's (1979) table.

To conclude, this study reports large differences in WM memory sensitivity between two variations of the change detection task. Unlike perceptual or long-term memory representation, visual WM sensitivity is dramatically reduced when estimated in a 2AFC task than in a same-different task. This finding illustrates the fragile nature of visual WM representation and highlights both the theoretical and methodological importance of taking testing format into account when comparing different studies of visual WM.

Acknowledgments

This research was supported in part by NIH MH071788 to YVJ. We thank Ben Kline and Jennifer Decker for help with data collection, and Khena Swallow, Weiwei Zhang, Paul Verhaeghen, and two anonymous reviewers for comments.

Appendix

Average percent correct across all conditions in all experiments. Standard errors of the mean are presented in parentheses.

	Low encoding similarity			High encoding similarity		
	2AFC	Same-Different		2AFC	Same-Different	
		Hits	False alarms		Hits	False alarms
Exp. 1a	72.4 (1.6)	71.4 (1.7)	21.8 (2.6)	72.6 (2.4)	70.7 (3.3)	15.6 (2.0)
Exp. 1b	70.4 (2.4)	75.1 (2.3)	26.9 (2.9)	80.7 (2.5)	74.2 (3.1)	16.4 (2.6)

	Set-size 3			Set-size 5		
	2AFC	Same-Different		2AFC	Same-Different	
		Hits	False alarms		Hits	False alarms
Exp. 2 no-cue	82.3 (2.0)	92.2 (1.9)	15.6 (3.3)	70.3 (2.4)	86.7 (2.2)	34.4 (4.0)
Exp. 2 retro-cue	91.8 (2.4)	92.2 (1.5)	11.5 (2.7)	84.8 (2.0)	88.0 (2.2)	23.4 (3.1)

	Far display			Near display		
	2AFC	Same-Different		2AFC	Same-Different	
		Hits	False alarms		Hits	False alarms
Exp. 3	82.2 (1.1)	70.0 (3.0)	17.4 (2.3)	82.1 (1.8)	73.0 (2.6)	18.3 (2.9)

	Old	Between-category		Within-category	
	Same-Diff (Hits)	2AFC	Same-Diff (Correct rej.)	2AFC	Same-Diff (Correct rej.)
Exp. 4	70.3(4.2)	92.6 (2.2)	93.7(1.8)	85.5 (3.9)	70.3(4.2)

References

- Awh E, Barton B, Vogel EK. Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science* 2007;18:622–628. [PubMed: 17614871]
 Baddeley, AD. Working memory. Oxford: Oxford University Press; 1986.

- Baddeley AD. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences* 2000;4:417–423. [PubMed: 11058819]
- Ballard DH, Hayhoe MM, Pelz JB. Memory representations in natural tasks. *Journal of Cognitive Neuroscience* 1995;7:66–80.
- Barton B, Ester E, Awh E. Discrete resource allocation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance* 2009;35:1359–1367. [PubMed: 19803642]
- Bastin C, Van der Linden M. The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology* 2003;17:14–24. [PubMed: 12597069]
- Bayley PJ, Wixted JT, Hopkins RO, Squire LR. Yes/no recognition, forced-choice recognition, and the human hippocampus. *Journal of Cognitive Neuroscience* 2008;20:505–512. [PubMed: 18004951]
- Bays PM, Husain M. Dynamic shifts of limited working memory resources in human vision. *Science* 2008;321:851–854. [PubMed: 18687968]
- Brady TF, Konkle T, Alvarez GA, Oliva A. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences* 2008;105:14325–14329.
- Brainard DH. The Psychophysics Toolbox. *Spatial Vision* 1997;10:433–436. [PubMed: 9176952]
- Cowan N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral & Brain Sciences* 2001;24:87–185. [PubMed: 11515286]
- Gold J, Fuller R, Robinson B, McMahon R, Braun E, Luck S. Intact attentional control of working memory encoding in schizophrenia. *Journal of Abnormal Psychology* 2006;115:658–673. [PubMed: 17100524]
- Green, DM.; Swets, JA. *Signal detection theory and psychophysics*. New York: Wiley; 1966.
- Griffin IC, Nobre AC. Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience* 2003;15(8):1176–1194. [PubMed: 14709235]
- Hacker MJ, Ratcliff R. A revised table of d' for M-alternative forced choice. *Perception & Psychophysics* 1979;26:168–170.
- Hollingworth A, Richard AM, Luck SJ. Understanding the function of visual short-term memory: Transsaccadic memory, object correspondence, and gaze correction. *Journal of Experimental Psychology: General* 2008;137:163–181. [PubMed: 18248135]
- Huang L, Treisman A, Pashler H. Characterizing the limits of human visual awareness. *Science* 2007;317:823–825. [PubMed: 17690299]
- Jiang, YV.; Makovski, T.; Shim, WM. Visual memory for features, conjunctions, objects, and locations. In: Brockmole, JR., editor. *The visual world in memory*. Hove, UK: Psychology Press; 2009. p. 33–65.
- Jiang YH, Olson IR, Chun MM. Organization of visual short-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition* 2000;26:683–702.
- Jiang YV, Shim WM, Makovski T. Visual working memory for line orientations and face identities. *Perception & Psychophysics* 2008;70:1581–1591. [PubMed: 19064500]
- Johnson JS, Spencer JP, Luck SJ, Schöner G. A dynamic neural field model of visual working memory and change detection. *Psychological Science* 2009;20:568–577. [PubMed: 19368698]
- Khoe W, Kroll NEA, Yonelinas AP, Dobbins IG, Knight RT. The contribution of recollection and familiarity to yes-no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologia* 2000;38:1333–1341. [PubMed: 10869576]
- Koutstaal W, Reddy C, Jackson EM, Prince S, Cendan DL, Schacter DL. False recognition of abstract versus common objects in older and younger adults: Testing the semantic categorization account. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 2003;29:499–510.
- Kunar MA, Flusberg SJ, Wolfe JM. The role of memory and restricted context in repeated visual search. *Perception & Psychophysics* 2008;70:314–328. [PubMed: 18372752]
- Kyllingsbæk S, Bundesen C. Changing change detection: Improving the reliability of measures of visual short-term memory capacity. *Psychonomic Bulletin & Review* 2009;16:1000–1010. [PubMed: 19966249]
- Landman R, Spekreijse H, Lamme VAF. Large capacity storage of integrated objects before change blindness. *Vision Research* 2003;43(2):149–164. [PubMed: 12536137]

- Lepsien J, Griffin IC, Devlin JT, Nobre AC. Directing spatial attention in mental representations: Interactions between attentional orienting and working-memory load. *Neuroimage* 2005;26(3):733–743. [PubMed: 15955482]
- Lepsien J, Nobre AC. Attentional modulation of object representations in working memory. *Cerebral Cortex* 2007;17(9):2072–2083. [PubMed: 17099066]
- Lin PH, Luck SJ. The influence of similarity on visual working memory representations. *Visual Cognition* 2009;17:356–372. [PubMed: 19430536]
- Luck SJ, Vogel EK. The capacity of visual working memory for features and conjunctions. *Nature* 1997;390(6657):279–281. [PubMed: 9384378]
- Macmillan, NA.; Creelman, CD. *Detection Theory: A User's Guide*. New York: Cambridge University Press; 1991.
- Makovski T, Jiang YV. Distributing versus focusing attention in visual short-term memory. *Psychonomic Bulletin & Review* 2007;14:1072–1078. [PubMed: 18229477]
- Makovski T, Jiang YV. Indirect assessment of visual working memory for simple and complex objects. *Memory & Cognition* 2008a;36:1132–1143.
- Makovski T, Jiang YV. Proactive interference from items previously stored in visual working memory. *Memory & Cognition* 2008b;36:43–52.
- Makovski T, Shim WM, Jiang YHV. Interference from filled delays on visual change detection. *Journal of Vision* 2006;6:1459–1470. [PubMed: 17209748]
- Makovski T, Sussman R, Jiang YV. Orienting attention in visual working memory reduces interference from memory probes. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 2008;34:369–380.
- Mate J, Baques J. Visual similarity at encoding and retrieval in an item recognition task. *Quarterly Journal of Experimental Psychology* 2009;62:1277–1284.
- Matsukura M, Luck SJ, Vecera SP. Attention effects during visual short-term memory maintenance: Protection or prioritization? *Perception & Psychophysics* 2007;69:1422–1434. [PubMed: 18078232]
- Migo E, Montaldi D, Norman KA, Quamme J, Mayes A. The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology* 2009;62:1198–1215.
- Miller EK, Erickson CA, Desimone R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience* 1996;16:5154–5167. [PubMed: 8756444]
- Parkin AJ, Yeomans J, Bindschaedler C. Further characterization of the executive memory impairment following frontal lobe lesions. *Brain and Cognition* 1994;26:23–42. [PubMed: 7986491]
- Pashler H. Familiarity and visual change detection. *Perception & Psychophysics* 1988;44:369–378. [PubMed: 3226885]
- Pelli DG. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision* 1997;10:437–442. [PubMed: 9176953]
- Phillips WA. Distinction between Sensory Storage and Short-Term Visual Memory. *Perception & Psychophysics* 1974;16(2):283–290.
- Potter, MC.; Jiang, YV. Visual short-term memory. In: Wilken, P., editor. *Companion to Consciousness*. Oxford: Oxford University Press; 2009.
- Rensink RA. Change detection. *Annual Review of Psychology* 2002;53:245–277.
- Sligte IG, Scholte HS, Lamme VA. Are there multiple visual short-term memory stores? *PLoS ONE* 2008;3:e1699. [PubMed: 18301775]
- Soto D, Heinke D, Humphreys GW, Blanco MJ. Early, involuntary top-down guidance of attention from working memory. *Journal of Experimental Psychology: Human Perception & Performance* 2005;31:248–261. [PubMed: 15826228]
- Soto D, Wriglesworth A, Bahrami-Balani A, Humphreys GW. Working memory enhances visual perception: Evidence from signal detection analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 2010;36:441–456.
- Vergauwe E, Barrouillet P, Camos V. Do mental processes share a domain-general resource? *Psychological Science* 2010;21:384–390. [PubMed: 20424075]

- Viswanathan S, Perl DR, Visscher KM, Kahana MJ, Sekuler R. Homogeneity computation: How interitem similarity in visual short-term memory alters recognition. *Psychonomic Bulletin & Review* 2010;17(1):59–65. [PubMed: 20081162]
- Vogel EK, McCollough AW, Machizawa MG. Neural measures reveal individual differences in controlling access to working memory. *Nature* 2005;438:500–503. [PubMed: 16306992]
- Westerberg CE, Paller KA, Weintraub S, Mesulam MM, Holdstock JS, Mayes AR, Reber PJ. When memory does not fail: Familiarity-based recognition in mild cognitive impairment and Alzheimer's disease. *Neuropsychology* 2006;20:193–205. [PubMed: 16594780]
- Wilken P, Ma WJ. A detection theory account of change detection. *Journal of Vision* 2004;4(12):1120–1135. [PubMed: 15669916]
- Yeshurun Y, Carrasco M, Maloney LT. Bias and sensitivity in two-interval forced choice procedures. *Vision Research* 2008;48:1837–1851. [PubMed: 18585750]
- Zhang W, Luck SJ. Discrete fixed-resolution representations in visual working memory. *Nature* 2008;453:233–235. [PubMed: 18385672]

Figure 1.

Left: The trial-sequence in Experiment 1 for the Same-Different procedure (top) and the 2AFC procedure (bottom). Right: Color stimuli used in Experiment 1. Items are not drawn to scale.

Figure 2.

Memory sensitivity as a function of testing procedure and encoding similarity for Experiment 1a (left) and 1b (right). Error bars show ± 1 S.E.

Figure 3.
The trial-sequence for Experiment 2 in the retro-cue (left) and the no-cue (right) conditions.
Items are not drawn to scale.

Figure 4.

Results from Experiment 2: visual WM sensitivity as a function of memory load, testing procedure, and cue conditions. Error bars show ± 1 S.E.

Figure 5.

Left: Displays used in Experiment 3. Right: Mean d' results in Experiment 3 as a function of display type and testing procedure. Error bars show ± 1 S.E.

Figure 6.

Left: sample stimuli used in Experiment 4. Four categories are shown here (each with 4 exemplars), from top to bottom: abstract objects, artist's drawings of objects, photographs of bagels, and photographs of binoculars. Right: Mean d' results in Experiment 4 as a function of foil category and testing procedure. Error bars show ± 1 S.E.