

# An adaptive perspective on visual working memory distortions

Chaipat Chunharas<sup>1,2,3</sup>, Rosanne L. Rademaker<sup>1,4</sup>, Timothy F. Brady<sup>1\*</sup>, & John T. Serences<sup>1,5,6\*</sup>

<sup>1</sup>Department of Psychology, University of California San Diego, La Jolla, California, USA

<sup>2</sup>Department of Medicine, King Chulalongkorn Memorial Hospital, Chulalongkorn University, Bangkok, Thailand

<sup>3</sup>Chulalongkorn Cognitive, Clinical & Computational Neuroscience research group, Chulalongkorn University, Bangkok, Thailand

<sup>4</sup>Ernst Strüngmann Institute for Neuroscience in cooperation with the Max Planck Society, Frankfurt, Germany

<sup>5</sup>Neurosciences Graduate Program, University of California San Diego, La Jolla, California, USA

<sup>6</sup>Kavli Institute for Brain and Mind, University of California, San Diego, La Jolla, California, USA

Running title: Biases in working memory

Correspondence:

Chaipat Chunharas

[chaipat.c@chula.ac.th](mailto:chaipat.c@chula.ac.th)

1873 Rama IV, Pathumwan, Bangkok, 10330, Thailand

**Conflict of interest:** the authors declare no conflict of interest

**Keywords:** visual working memory; memory biases; attraction bias; repulsion bias; color memory

## Author Note

This work was supported by NEI R01-EY025872 (JTS), by a Thai Red Cross Society grant (CC), by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No 743941 (RLR), and by an NSF CAREER award BCS-1653457 (TFB). \* JTS and TFB were co-senior authors on this project.

\* Parts of this work were also presented at the Vision Sciences Society annual meeting. The data and codes are available at [https://osf.io/qp6xk/?view\\_only=0559769c587c4c8294288451e8af239e](https://osf.io/qp6xk/?view_only=0559769c587c4c8294288451e8af239e).

© 2022, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/xge0001191

## **Abstract**

When holding multiple items in visual working memory, representations of individual items are often attracted to, or repelled from, each other. While empirically well-established, existing frameworks do not account for both types of distortions, which appear to be in opposition. Here, we demonstrate that both types of memory distortion may confer functional benefits under different circumstances. When there are many items to remember and subjects are near their capacity to accurately remember each item individually, memories for each item become more similar (attraction). However, when remembering smaller sets of highly similar but discernible items, memory for each item becomes more distinct (repulsion), possibly to support better discrimination. Importantly, this repulsion grows stronger with longer delays, suggesting that it dynamically evolves in memory and is not just a differentiation process that occurs during encoding. Furthermore, both attraction and repulsion occur even in tasks designed to mitigate response bias concerns, suggesting they are genuine changes in memory representations. Together, these results are in line with the theory that attraction biases act to stabilize memory signals by capitalizing on information about an entire group of items, whereas repulsion biases reflect a tradeoff between maintaining accurate but distinct representations. Both biases suggest that human memory systems may sacrifice veridical representations in favor of representations that better support specific behavioral goals.

Memory is a constructive rather than a passive process. For example, people will naturally fill in gaps when recalling a story in an attempt to make the story more coherent (Bartlett, 1932; Loftus, 2005; Roediger & McDermott, 1995). When people study a list of words, they often falsely recall or recognize associated words that were not on the original list (Deese, 1959; Underwood et al., 1965), and later report these words as actual memories (Schooler et al., 1988). Similarly, visual memory is not analogous to taking a photo – instead, there are many systematic biases in how visual attributes are remembered after a sensory stimulus is no longer available (Alvarez, 2011; Bar, 2004; Brady & Alvarez, 2011; Fischer & Whitney, 2014; Huang & Sekuler, 2010; Koutstaal et al., 2001; Rademaker et al., 2015; Schacter et al., 2011).

When people are tasked with remembering a visual item, such memories are often distorted toward existing, learned prototypes (Huttenlocher et al., 1991, 2000; Hemmer & Steyvers, 2009). Such distortion can also occur not toward pre-learned prototypes, but toward the central tendency of a group within a single presentation. For example, when people are asked to remember multiple visual items, these memories are 'attracted' to each other - that is, different objects are remembered as more similar than they really were (Brady & Alvarez, 2011; Dubé et al., 2014; Dubé & Sekuler, 2015; Freyd & Johnson, 1987; Huang & Sekuler, 2010; Spencer & Hund, 2002). It has been proposed that this occurs because object-level representations are imprecise, so these unstable representations are constrained by using additional information about the properties of the set of items as a whole (i.e. group-level representation). Thus, inter-item attraction biases may be the result of weighting the representation of each individual object towards the “summary” of the set to achieve a more stable memory at the expense of maintaining distinctions between individual items (Brady & Alvarez, 2011; Huttenlocher et al., 1991).

84 Interestingly, attraction biases are not ubiquitous. Under some conditions, when multiple items are  
85 shown at once, memories for individual specific items have been shown to repel each other, being  
86 remembered as more different than they really were (Bae & Luck, 2017; Golomb, 2015; O'Toole  
87 & Wenderoth, 1977; Rademaker et al., 2015; Rauber & Treue, 1998; Suzuki & Cavanagh, 1997).  
88 However, far less research has been dedicated to understanding inter-item repulsion biases.  
89 Repulsion biases have sometimes been proposed to arise from lateral inhibition, as competition  
90 between neurons representing similar feature values may lead to representations that repel away  
91 from each other (Johnson et al., 2009; Wei et al., 2012), akin to repulsion resulting from  
92 competition during early perceptual processing (Jazayeri & Movshon, 2006; Navalpakkam & Itti,  
93 2007; Purushothaman & Bradley, 2005; Regan & Beverley, 1985; Scolari & Serences, 2009;  
94 Scolari & Serences, 2010; Smith et al., 2005). However, while providing a possible mechanistic  
95 basis, such theories do not straightforwardly explain why repulsion biases sometimes arise and  
96 sometimes do not; nor why attraction biases occur for similar stimuli under other circumstances.  
97 Despite the importance and pervasiveness of these memory distortions, to date there have been  
98 few attempts to understand why memories sometimes attract, while at other times they repel.

99 Because these are rarely studied together, it is still unclear whether these inter-item memory  
100 distortions that arise for simultaneously presented items are due to changes in the representations  
101 themselves, or if they instead reflect demand characteristics that lead to systematic response biases.  
102 For example, repulsion biases can emerge in continuous report paradigms if participants want to  
103 actively communicate that they know two items are different, even if participants have access to  
104 veridical representations, and most work to date has demonstrated repulsion biases only in such  
105 continuous report situations (Bae & Luck, 2017; Golomb, 2015; Rademaker et al., 2015).

To establish when attraction biases and repulsion biases arise and whether they are properties of the memory system or a result of stimulus differences or straightforward responses biases that occur only in continuous report tasks, we present a series of experiments. First, we determine whether attraction and repulsion are simply properties of subject's communicative intent in continuous report tasks. Second, we examine whether they arise in predictable circumstances, by manipulating task difficulty and the similarity and distinctiveness of the memoranda. While these are general issues, related to nearly all kinds of memory, we tested these ideas in a well-studied domain – visual working memory for color – where memory representations can be precisely quantified. Task difficulty was increased or decreased by changing how many items must be remembered (set size), how distinctive the colors are from each other (their proximity in color space), and encoding time and memory delay.

After establishing the empirical phenomena, we adopt the perspective (in the 'Framework' section of the paper) that these inter-item biases for simultaneously presented items may be natural consequences of the memory system attempting to minimize memory error, and that systematic distortion can be adaptive in particular circumstances (Schacter et al., 2011). Specifically, when many items are present and memories for individual items are noisy, attraction biases are known to be optimal for minimizing error (e.g., Brady & Alvarez, 2011). In this case, relying on group-level statistics provides an efficient means of retaining at least some information about all items at the expense of precisely representing information about each single item. Repulsion biases can also reduce error in some situations, making them adaptive. In particular, if items would naturally be blended or confused by our memory system (Swan & Wyble, 2014; Oberauer & Lin, 2017) – that is, if similar items would interfere with each other – then repulsion can reduce this tendency and reduce error. In this case, the goal is to distinguish highly similar or noisy representations, by

reducing the confusability between memory items. In particular, if items interfere to the extent they overlap in features, then repulsion is adaptive when items overlap in representation. In discussing this framework, we examine whether attraction and/or repulsion occur in the circumstances predicted by this framework, and not in circumstances where biases would be maladaptive to memory performance (i.e. contrary to the adaptive framework).

Overall, we find that when distinctiveness between two items goes down, repulsion biases are stronger (up to the point where two items become indistinguishable, and attraction takes over as the dominant force). Repulsion biases also grow stronger with longer delays, suggesting that as memory demands increase and item representations become noisier, memories are biased to keep items individuated. In contrast, we observe attraction biases when individuating items is more difficult due to a higher memory load (in an experiment with 4 instead of 2 memory items), consistent with sacrificing single-item discriminability in order to remember at least some information about ensemble-level features. Importantly, by using a 2-alternative forced-choice paradigm we were able to test the role of demand characteristics: the results imply that repulsion biases are not the result of participants trying to communicate that they can distinguish two targets in a continuous report task. Collectively, these studies suggest that, given task-imposed constraints, attraction or repulsion biases may help to improve behavioral performance even though these biases may lead to non-veridical memories.

## Experiment 1: Memory distortion vs. response strategy

Do memory items truly 'repel' each other when people hold in mind a small number of similar items? In Experiment 1 we sought to replicate this basic repulsion effect and to determine if previously reported biases (e.g., Bae & Luck, 2017; Golomb, 2015) are more likely to reflect memory distortions, or if they are a result of changes in response strategy to communicate an understanding of the continuous reproduction task. That is, when participants remember a pair of colors, they can communicate their awareness of the colors being distinct from one another by exaggerating the difference between the two. When cued to report one of the two remembered items on a continuous color-wheel, this strategy would result in an answer repelled away from the uncued non-target item – mimicking a repulsion bias. We directly addressed this possible response strategy by having participants remember two colored items over a brief delay (Figure 1a), after which they perform a 2-alternative forced-choice (2-AFC) task comparing the correct (cued target) color to an incorrect (distorted foil) color (Figure 1a, b). By presenting participants with the correct answer on every trial, such response biases are discouraged as they are detrimental to task performance, and an understanding of the task is best communicated by picking the correct color. To distinguish between attraction and repulsion in this 2-AFC paradigm, the incorrect foil color was distorted by 6° relative to the correct target color, and the distortion was either towards the non-target (i.e. 'attracted' to the non-target) or away from the non-target (i.e. 'repelled' from the non-target). If memories for the two colors were repelled from each other, a foil color that was distorted *toward* the non-target would be less often confused with the correct answer (have a higher accuracy) than a foil color that was distorted *away* from the non-target (have a lower accuracy).

## Methods

174 The datasets from all of the current studies (plus the code used to generate the stimuli and analyze  
175 the data) are available in the OSF repository

176 [https://osf.io/qp6xk/?view\\_only=0559769c587c4c8294288451e8af239e](https://osf.io/qp6xk/?view_only=0559769c587c4c8294288451e8af239e)

177 **Participants.** 45 naïve participants were recruited from Amazon Mechanical Turk. In this and all  
178 other experiments reported, all experimental procedures were approved by the UCSD Institutional  
179 Research Board, all online participants provided written informed consent, and all reported normal  
180 or corrected-to-normal vision without color-blindness. Participants were naïve to the purpose of  
181 the study and received payment (\$6 per hour) for their time.

182 **Stimuli & Procedure.** All stimuli were drawn on a  $500 \times 500$  pixels white background with a  
183 black border around it (1 pixel wide). The fixation cross was in the middle of the canvas, and 12  
184 small circular placeholders were shown around fixation, each centered at a distance of 120 pixels.  
185 Each placeholder had a radius of 20 pixels, and the inter-placeholder distances were 62 pixels  
186 (center-to-center). Placeholders were positioned such that six of them were on the left, and the  
187 other six were on the right side of fixation. Furthermore, two placeholders were always presented  
188 directly to the left and right of fixation, centered at 35 pixels from fixation. Memory items were  
189 colors selected from a subset of CIE  $L^*a^*b$  color space ( $L = 70$ ,  $a = 20$ ,  $b = 38$ , radius = 60). Note,  
190 while one of the memory items was always selected randomly from this color space, the second  
191 item always differed from the first by  $45^\circ$ . The location probe, cueing participants which memory  
192 item to report on, was a small equilateral black triangle, 20 pixels wide and 20 pixels tall.

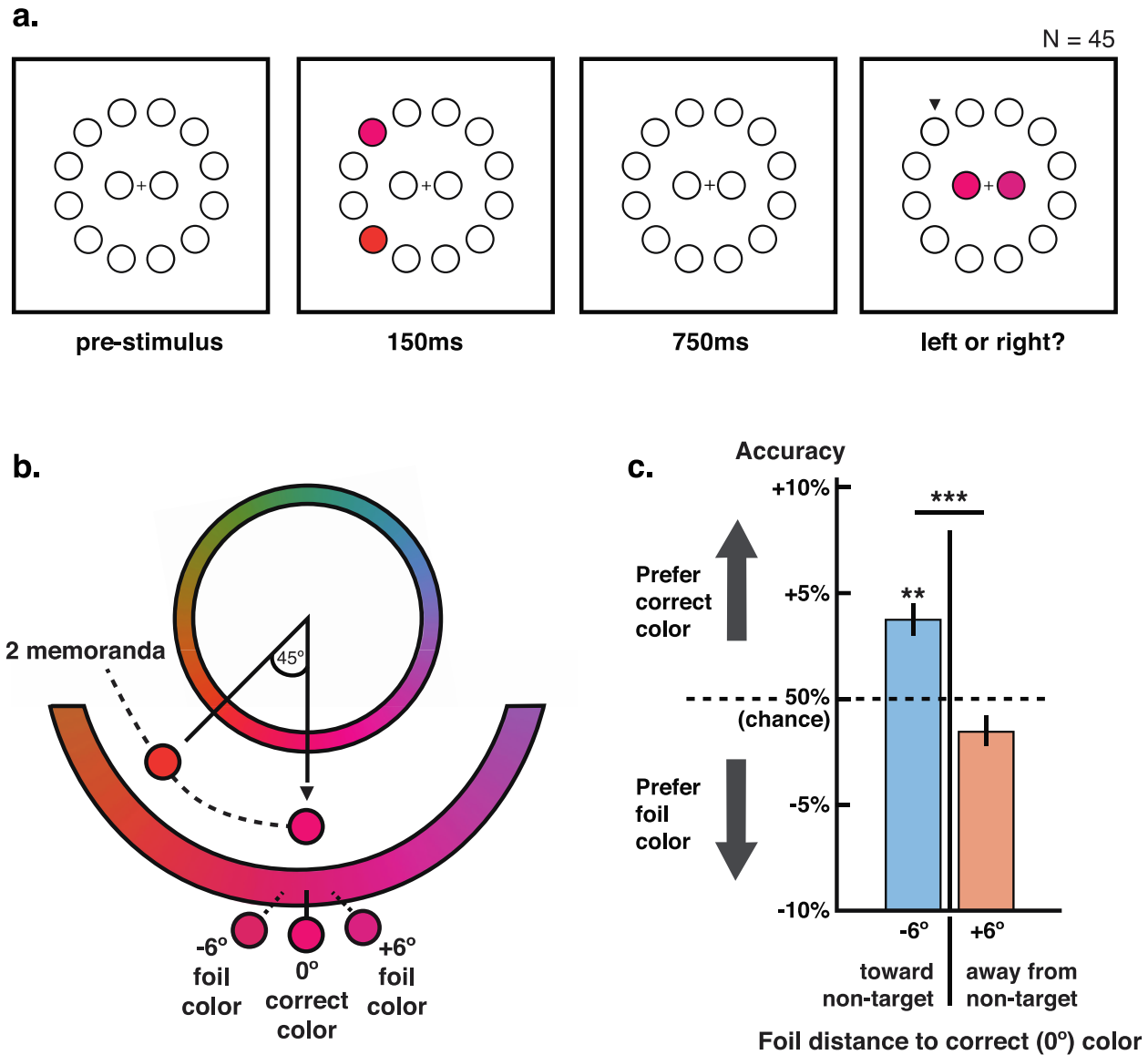
193 Participants were shown two memory items for 150ms at two randomly selected placeholders in  
194 the display (out of 12 possible placeholders), with the restriction that there were always at least 2  
195 empty placeholders between the two memory items. After a 750ms delay, a location cue (arrow)



indicated which of the items was the memory target, and two response options appeared in the placeholders directly to the left and right of fixation. One of the response options was always the correct color (i.e. identical to the color that was cued), while the other option was always a foil, and participants made a 2-AFC judgment between the two response options. The foil always differed from the correct color by 6° in color space, either in the direction towards (50% of trials) or away (50% of trials) from the non-target memory item. The positions (left or right of fixation) of the correct and foil response options were completely randomized. Participants had to press “z” or “m” to select the choice presented on the left or right of fixation, respectively, before proceeding to the next trial. There were 60 trials per condition (a total of 120 trials per participant).

## **Results**

As predicted by an account where repulsion is a genuine memory phenomenon, participants were better at rejecting a foil color that was distorted toward the non-target memory item than rejecting a foil color that was distorted away from the non-target memory item – an indicator of repulsion bias ( $t_{(44)}=3.98$ ;  $p<0.001$ ; Figure 1c). In other words, performance was higher when a foil was distorted towards the non-target memory. This shows that repulsion biases occur even in a 2-AFC format with an objectively correct answer vs. an objectively incorrect answer, implying that repulsion is not merely the result of this particular a priori plausible response strategy.



**Figure 1.** Task and results from Experiment 1. a.) Participants remembered two memory items that were always  $45^\circ$  apart in color space. Memory items were briefly presented for 150ms at two randomly chosen placeholder locations. After a 750ms delay, participants reported the color of the target item (cued with an arrow) by choosing between two options, one always being the correct color, and the other always being an incorrect foil color that was distorted from the correct color by  $6^\circ$  in a direction either toward or away from the non-target color. In the example trial shown here, the correct response is shown on the left, while the foil on the right is distorted in a direction

away from the non-target color. b.) Two memory colors were selected to lie within  $45^\circ$  of each other in color-space (at any possible position on the color wheel). The target color (cued after the delay) was always one of the response options during the 2AFC phase of the trial (i.e. the “correct color”). The other response option was a foil color. The foil color always differed  $6^\circ$  from the correct target color and could be distorted towards ( $-6^\circ$ ) or away ( $+6^\circ$ ) from non-target color. c.) Participants preferred the correct color to the foil when the foil was distorted toward the non-target color, as indicated by above-chance performance (blue bar;  $t_{(44)}=3.73$ ;  $p=0.006$ ). This differed significantly from trials on which the foil was distorted away from the non-target color (compare blue and red bars;  $t_{(44)}=3.98$ ;  $p<0.001$ ), with a trend towards participants preferring the incorrect foil over the correct answer, as indicated by numerically below-chance performance (red bar;  $t_{(44)}=-1.76$ ;  $p=0.08$ ). This is the expected result when memory for the target is distorted away from the non-target (i.e. when there is a repulsion bias). Error-bars represent  $\pm 1$  within-subject SEM. Double and triple asterisks indicates  $p\leq 0.01$  and  $p\leq 0.001$  respectively.

## **Experiment 2: Memory distortion vs. response strategy, and the role of task engagement.**

We next replicated and extended Experiment 1 with additional foil colors that were  $25^\circ$  away from the correct memory target. We added  $25^\circ$  foils in this second experiment to test the possibility that participants simply favored all colors distorted away from the non-target color by way of a response strategy, even though such a strategy would result in objectively incorrect performance in this task. After all, if participants meant to communicate their awareness of the two memory colors being distinct, they would prefer any foil away from the non-target over the correct answer. In this were the case,  $25^\circ$  foils would be favored even more than  $6^\circ$  foils, because they are more clearly away from non-target color. This hypothesis is schematically shown in Figure 2a (top

panel; ‘Prediction 1’). By contrast, if memories of the two colors were truly repelled from one another, and participants remembered the target item as further from the non-target than it actually was, performance should depend on the degree of foil distortion. Specifically, participants should be more likely to choose the foil (and give an incorrect answer) when it closely matches their distorted memory (e.g. the  $+6^\circ$  foil), but more likely to choose the correct color when the distortion of the foil becomes irreconcilable with their memory (e.g. the  $+25^\circ$  foil). This hypothesis is also schematically shown in Figure 2a (bottom panel; ‘Prediction 2’).

## Methods

**Participants.** 45 new naïve participants were recruited from Mechanical Turk for Experiment 2. For the control experiment replicating Experiment 2 (Appendix Figure 1) we recruited another independent set of 45 participants from Amazon Mechanical Turk.

**Stimuli & Procedure.** The stimuli and task were identical to Experiment 1, except that in Experiment 2 the foil could differ from the correct color by either  $6^\circ$  (45% of trials),  $25^\circ$  (45% of trials), or  $180^\circ$  (10% of trials). As in Experiment 1, on half of these trials the foil was in the direction *toward* the non-target in color space, while on the other half of trials the foil was *away* from the non-target in color space. Given how easily distinguishable the  $180^\circ$  foils were from the correct color, these trials served as catch trials. For the control experiment replicating Experiment 2 (presented in Appendix Figure 1), the foil could differ from the correct color by either  $6^\circ$  (90% of trials), or  $180^\circ$  (10% of trials). In Experiment 2, there were 30 trials per main condition (total of 4 main conditions, i.e.  $6^\circ$  vs.  $25^\circ$  foils, crossed with distortion away vs. toward non-target) plus 12 catch trials (a total of 132 trials per participant). In the replication study of Experiment 2, there were 60 trials per condition ( $6^\circ$  foils, with distortion away vs. toward non-target) plus 12 catch

trials (a total of 132 trials per participant).

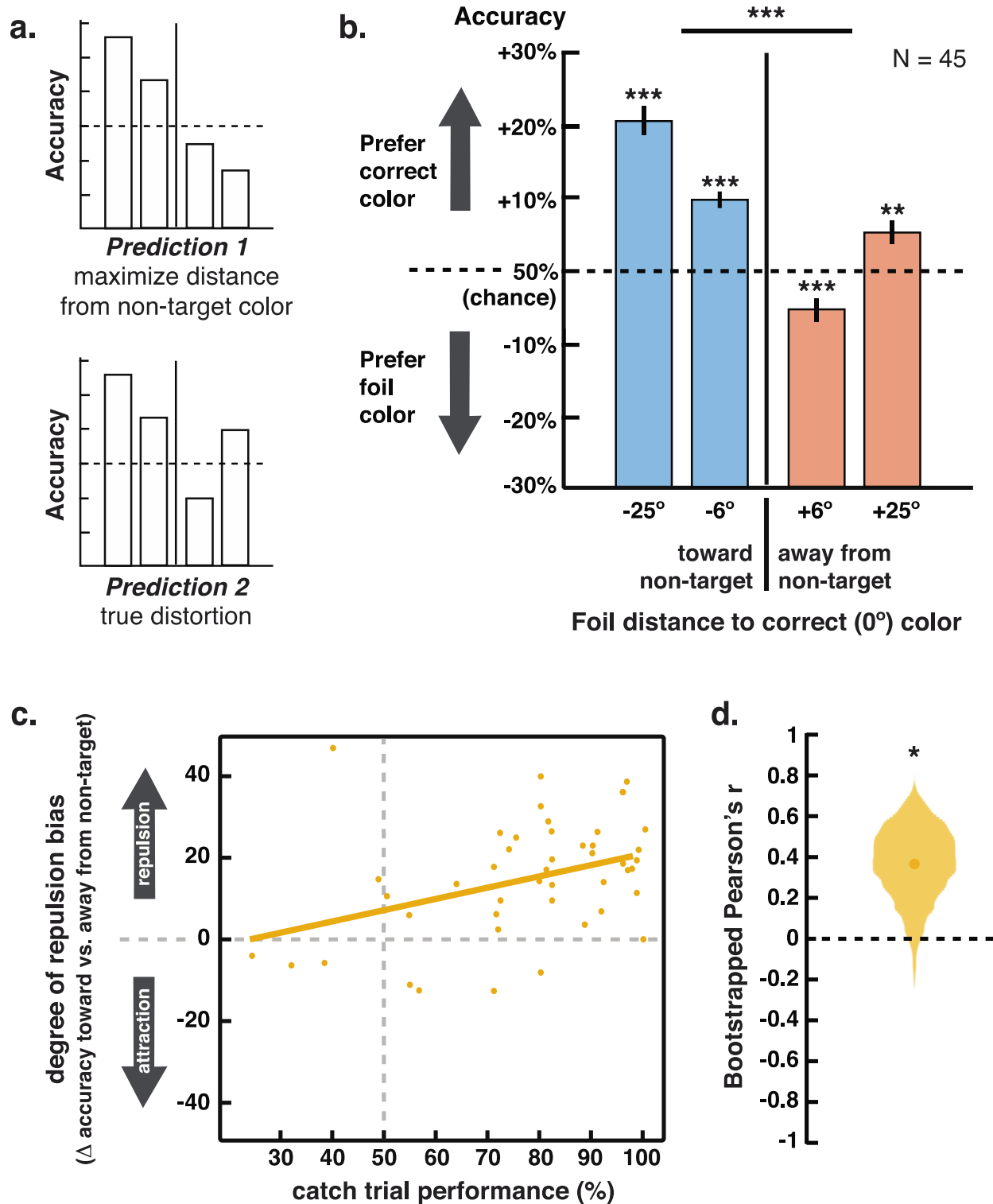
## Results

We replicated Experiment 1, as participants were again better at rejecting a foil color that was distorted *toward* compared to *away* from the non-target memory item ( $F_{(1,44)} = 49.2$ ;  $p < 0.001$ ; Figure 2b, compare blue and red bars). Interestingly, subjects more often selected foils that were  $6^\circ$  away from the non-target color compared to the correct target color, resulting in below-chance level performance in this condition ( $t_{(44)} = 3.41$ ;  $p = 0.001$ ; Figure 3b, compare  $+6^\circ$  bar against chance accuracy). This is consistent with a strong degree of memory distortion, where participants prefer a repelled foil color relative to the correct answer. In contrast, subjects successfully rejected all other foils resulting in above-chance level performance ( $t_{(44)} = 8.70, 7.70, \text{ and } 3.05$ ;  $p < 0.001, < 0.001, \text{ and } 0.004$  for foils that were  $-25^\circ, -6^\circ, \text{ and } +25^\circ$  relative to the non-target color, respectively; Figure 2b). Thus, participants showed a clear repulsion bias that cannot be easily explained by response strategy. Instead, the data are consistent with a target memory that was truly distorted away from the non-target item by several degrees.

In addition to replicating Experiment 1 and bolstering the case in favor of a true repulsion bias (and not a response strategy), we wanted to know if the degree of repulsion bias was related to the level of task engagement from our participants. To this end, Experiment 2 included foils that were  $180^\circ$  away from the cued memory target on 10% of the trials. We termed trials with a  $180^\circ$  foil “catch trials”, as subjects should rarely, if ever, confuse these foils with the correct color. Thus, performance on catch trials provides a useful measure of overall task engagement and effort. Critically, if the repulsion bias is adaptive and improves memory, one would expect the degree of

repulsion to positively correlate with overall performance. In contrast, if biases arise due to lack of effort or some other non-task related factor like response strategy, we might expect repulsion bias to be negatively correlated with performance (or uncorrelated). We quantified the degree of repulsion as performance on trials with foils distorted towards the non-target (both by 6° and 25°), minus trials with foils distorted away from the non-target (both by 6° and 25°). This metric will be larger for participants with stronger repulsion. We found a moderate positive correlation between the degree of repulsion bias and overall task engagement (Pearson's  $r=0.37$ ;  $p=0.013$ ; Figure 2c) supported by a bootstrapping analysis (bootstrapped mean Pearson's  $r=0.37$ , two-tailed  $p=0.048$ ; Figure 2d). This positive correlation between repulsion bias and overall task engagement was replicated in an independent set of 45 naïve subjects (Pearson's  $r=0.39$ ;  $p=0.009$ ; Appendix Figure 1). Thus, repulsion biases do not appear to arise solely in participants putting in low or moderate effort, instead, they are strongest in participants with the highest levels of task engagement.

Overall, Experiments 1 and 2 provide evidence for a repulsion bias that cannot be explained by these straightforward, a priori reasonable communicative strategies resulting in simple response biases, or a lower amount of effort.



309

310

**Figure 2.** Experiment 2 hypotheses and results. a.) The two panels show our predictions if participants were trying to strategically avoid non-target colors leading to a response bias (prediction 1, top) vs. if participants had memories that were truly distorted away from one another (prediction 2, bottom). b.) Data from 45 subjects showed a pattern consistent with true memory distortions as in prediction 2. Participants performed significantly below chance (i.e. preferred the foil over the correct response option) only when the foil was distorted 6° (but not 25°) away from the non-target color. This is in line with a true distortion of the remembered color and is indicative of participants finding that the foil more accurately reflected their memory representation. Presented with any other foil (foils distorted towards the non-target, or a foil distorted farther away from the non-target), participants chose the correct answer more often than chance. Error-bars represent  $\pm 1$  within-subject SEM. c.) Degree of repulsion bias (indexed as accuracy differences between all trials with foils distorted toward and away from the non-target color) plotted against general memory performance (indexed by performance on catch trials). Each dot represents a single participant. We found stronger repulsion biases in participants with better general memory performance (Pearson's  $r = 0.37$ ,  $p=0.013$ ). Note that the position of the dots are slightly independently jittered by random noise ( $\pm 5\%$ ) to aid visualization of all 45 data points. The solid yellow line represents the best fit to theunjittered data. d.) Distribution plot of bootstrapped Pearson's  $r$  between repulsion magnitude and general memory performance (5000 iterations of resampling with replacement). Single, double and triple asterisks indicate  $p \leq 0.05$ ,  $p \leq 0.01$  and  $p \leq 0.001$  respectively.

It is still possible that the repulsion bias is the result of a response strategy whereby the participant is trying to signal not only an understanding of the task (leading to repulsion), but also wants to



communicate which of the two items was being recalled (leading to repulsion only for the probed item).

Such an account would naturally predict a disappearance of the repulsion bias when not one, but both memory items were probed. To investigate this possibility, we reanalyze an existing open data set (Adam, Vogel, & Awh 2017) where participants were required to reproduce the colors of two memoranda in a random order. To quantify the repulsion bias, we took the absolute difference between the two stimulus colors presented and compared this to the absolute difference between the two responses participants made. In case of repulsion, response errors will be further apart in color space than the actual stimuli were. Indeed, we found that differences between the response errors were significantly larger than the stimulus differences ( $t_{(1,16)} = 3.11$ ,  $p < 0.01$ ). This suggests that also in a whole report task, items at set size 2 repel each other systematically.

Overall, while it is never possible to rule out all possible response strategies. Some aspects of these effects could still be happening at response stages, even if they are not explainable by the response strategies we test here and that are most plausible a priori. However, we have shown they apply not only in continuous report where a single item is probed, but also in continuous report where both items are probed, and in two kinds of forced-choice tasks, including one where there is a single objectively correct answer and a single objectively incorrect answer. While different response strategies could be at work in each task, giving rise to this pattern, this work provides significant evidence in favor of a mnemonic shift account.

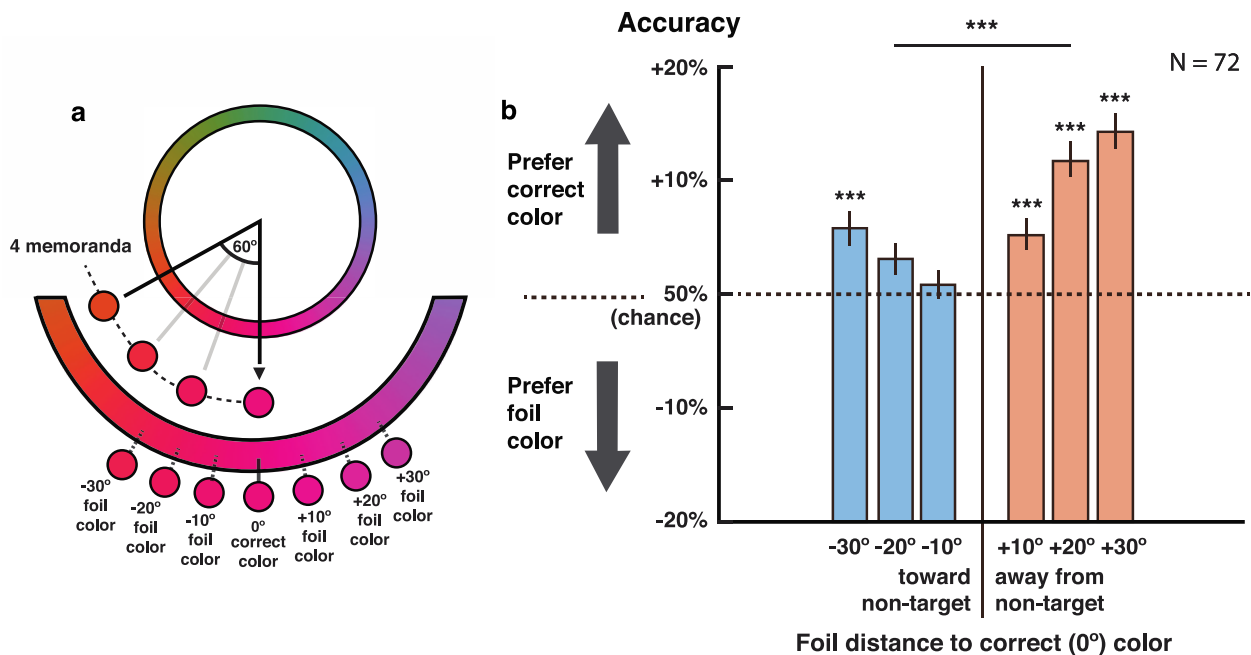
### **Experiment 3: Attraction vs. repulsion.**

We next sought to manipulate task factors to test if we could systematically flip distortions from repulsion to attraction, even for the same kind of stimuli. We used the same experimental paradigm as in Experiments 1 and 2, but increased the number of colors participants had to remember from 2 to 4 items (see also Appendix Figure 2). Given well-documented limits on the amount of information that can be retained in working memory (e.g., Bays, 2015; Bays et al., 2009; Luck & Vogel, 1997; Ma et al., 2014), remembering 4 items should be quite challenging for the majority of participants. Our adaptive framework suggests that when it is challenging to maintain individuated representations of all memory items, a partial reliance on group-level statistics (Brady & Alvarez, 2011) or partial blending between items (Swan & Wyble, 2014; Oberauer & Lin, 2017) is optimal, because it supplements the noisy information available about each of the individual items with information from the other items. In this context, when participants are presented with a foil that is distorted *towards* the colors of the other items in the set (Figure 3a), they should be more likely to confuse the foil with the correct (cued) target color (i.e. show an attraction bias) – the exact opposite of the repulsion bias observed in the previous experiments. To test this, in this experiment the four to-be-remembered colors spanned 60° of color space (in 20° steps), and we always cued one of the colors on the “edge” of this set. There were 6 possible foil conditions, of which 3 were distorted towards the other non-target items, and 3 were distorted away from the other non-target items (Figure 3a).

## Methods

**Participants.** A total of 72 naïve participants were recruited from Amazon Mechanical Turk. This is more than in Experiments 1 and 2 due to the increased difficulty of the task associated with the higher set size (thus requiring more power). Participants received \$8 per hour for their time.

**Stimuli & Procedure.** Stimulus and task presentation was identical to Experiments 1 and 2 with the following exceptions: Participants were shown four color items for 800ms, memory item locations were random (could be any 4 placeholders out of the possible 12) with the restriction that there was always at least 1 empty placeholder between each of the memory items. The four items were remembered over a 1000ms delay. The four colors were within 60° from each other in color space, and all colors were equally spaced from one another (i.e. the shortest possible color distance between two items was 20°; see also Figure 3A). The memory target probed at the end of the delay was always one of the colors at the edge of the set. Again, the correct color was always included as one of the response options, while the foil color differed by either 10°, 20°, or 30° from the correct target color option. The foil color could be either *toward* the colors of the other memory items (note how a -20° foil is identical to one of the other colors in the display, and a -30° foil is exactly the mean of all 4 colors), or it could be *away* from the other colors. There were 20 trials per main condition (total of 6 conditions, 10° vs. 20° vs. 30° foils, and distortion away vs. toward non-target) which means a total of 120 trials per participant.



**Figure 3.** a.) A set of four colors were selected to lie within 60° of each other in color-space (all separated by steps of 20°) and were presented at random spatial positions (chosen from 12 possible placeholders; see Appendix Figure 2). The cued memory target color (to be reported after the delay) was always one of the colors on the edge of the set. In this diagram, the target is the memorandum with the arrow pointing at it. After a 1000ms delay, participants performed a 2-AFC memory test. One of the options was always the correct (cued) target color, while the other choice was an incorrect foil of which the color differed by either 10°, 20° or 30° from the correct target color. The foil could be distorted towards (-10°, -20°, or -30°) or away (+10°, +20°, or +30°) from the center of the four colors in the memory set. b.) Accuracy was lower when the task was more difficult: When subjects had to choose between the correct color and a foil color that was very similar to the correct color (for example, differed by 10°) accuracy was closer to chance compared to when subjects had to choose between the correct color and a foil that differed more from the correct color (for example, differed by 30°). Importantly, performance was worse when the foil color was distorted toward the other memory colors in the set (i.e. the blue bars are lower overall than the red bars). This indicates an attraction of the cued item towards the other non-target items. Error-bars represent  $\pm 1$  within-subject SEM. Asterisks represent significance levels of differences between foils that were toward vs. away from non-target, with the triple asterisks indicating  $p < 0.001$ .

## Results

We performed a 3x2 repeated-measure ANOVA, and found a significant main effect of the distances of the foils from the target ( $F_{(2,142)} = 13.14$ ;  $p < 0.0001$ ), and a significant main effect of the direction of the foil ( $F_{(1,71)} = 15.48$ ;  $p < 0.0001$ ). There was no significant interaction ( $F_{(2,142)} = 1.93$ ;  $p = 0.15$ ). Specifically, we found that participants were more accurate when the foil

colors were more dissimilar from the correct color, making discrimination easier: Accuracy was 53%, 57.4% and 60.1% correct for foils that were 10°, 20° and 30° away from the correct color, respectively (Figure 3b, compare bars with smaller versus larger target-foil distances). Importantly, participants were also better at choosing the correct answer when the foil color was distorted *away* from the other non-target colors in the set (60.4% correct) compared to when the foil color was distorted *toward* the other non-target colors (53.2% correct; Figure 3b, compare blue and red bars). This implies an attraction bias towards the remembered non-target items, and stands in contrast to the repulsion bias found with set size 2 (in Experiments 1 and 2). Previous work has demonstrated that attraction biases in visual working memory arise from slight shifts toward the gist, and not solely from swaps or guesses based on the average color (e.g., Brady & Alvarez, 2011). Consistent with this, we found little evidence for swaps and guesses in our data as well: In particular, the -20° foil was the same color as one of the non-target items; and the -30° foil was the mean of all colors in the set. Nevertheless, neither the -20° nor the -30° foils were selected as often as -10° foil – indicative of only a slight attraction toward the other colors.

#### **Experiment 4: Biases depend on the degree of distinctiveness between items.**

In a fourth experiment (Figure 4a) we sought to determine if reducing the distinctiveness between items (by making items increasingly similar or noisy) impacts the amount of repulsion bias in a manner consistent with our framework. In particular, if the memory system naturally blends together similar items (as in the models of Swan & Wyble, 2014; Oberauer & Lin, 2017), then two items that are recognizably distinct (i.e., can still be told apart) but still similar enough to likely be blended, repulsion should arise (see ‘Framework’ section). To this end, we asked participants to remember 2 colors, and we independently manipulated both memory encoding time (50, 150 and 500ms) and distance in feature space between the two colors (0°, 20°, 45°, 90° and 135°). If less

easily distinguishable colors need to be differentiated from one another in order to improve behavioral performance, a higher degree of similarity between the 2 memory items should result in a stronger repulsion bias – but critically, there should be an exception for colors that are *so* similar that they are perceived as the same color and are thus put into a single ‘chunk’ or group. Furthermore, the color distance that creates maximal repulsion should depend on how precise the representations are: Two very precise representations at a given color distance may not require repulsion to be differentiated, while two more imprecise representations at that same color distance could be more easily differentiated with repulsion. In other words, when two memory representations are not too similar or too distinct, the magnitude of repulsion bias will depend on the precision of the memories. Repulsion bias might be necessary if the memory representations are relatively less precise. Representational precision should vary with encoding time (i.e. memory should be more precise at longer encoding times). Since Experiments 1 and 2 suggest that repulsion biases reflect changes in encoding and memory as opposed to response strategy, here we used a continuous report task where subjects had to report the remembered color by choosing from a continuous 360° color-wheel. The use of a continuous report task allowed us to generalize our findings beyond the 2-AFC paradigm, and to gain insight into how memory biases manifest in response error distributions.

## Methods

**Participants.** 24 healthy volunteers (15 female, mean age of  $19.75 \pm 1.52$ ) from the University of California San Diego (UCSD) community participated in the experiment in person. All procedures were approved by the UCSD Institutional Research Board and all participants provided written informed consent, and reported normal or corrected-to-normal vision without color-blindness. Participants were naïve to the purpose of the study and received partial course credit for their time.

**Stimuli & Procedure.** Stimuli were rendered on a CRT monitor with a 60-Hz refresh rate and a screen size of 40 x 30 cm. Stimuli were generated using MATLAB and the Psychophysics toolbox (Brainard, 1997; Pelli, 1997). Participants were instructed to maintain fixation throughout, aided by a white central fixation dot ( $0.5^\circ$  diameter) presented on a dark-gray background of 2.37 cd/m<sup>2</sup>. Memory items were colors randomly selected from a subset of CIE color space ( $L = 70$ ,  $a = 20$ ,  $b = 38$ , radius = 60), as was done in the previous three experiments. Sixteen white placeholders ( $4.3^\circ$  radius,  $0.2^\circ$  thick line) were positioned around the fixation point (centered at  $10.5^\circ$  from fixation). The locations of the two memory targets were selected at random with the exception that (1) they were always presented in the same hemifield to maximize inter-item competition (Alvarez & Cavanagh, 2005; Cohen et al., 2016; Störmer et al., 2014) and (2) there were always 2 empty placeholders between the two memory items (i.e. they were spaced  $\sim 4^\circ$  apart, center-to-center).

On each trial (Figure 4a), two colored stimuli were presented for either 50ms, 150ms or 500ms and participants had to remember the colors as precisely as possible. The colors of the two memory items could be either  $0^\circ$ ,  $20^\circ$ ,  $45^\circ$ ,  $90^\circ$  or  $135^\circ$  apart in color space (with  $\pm 3^\circ$  random jitter). After a 750ms delay, one of the two colors was probed via a spatial cue (the rim of the placeholder in one location got thicker). Along with the spatial probe, a randomly oriented color-wheel (with  $10^\circ$  radius,  $1^\circ$  wide) was presented around fixation, and a crosshair appeared at the fixation point. Participants used the mouse to move the crosshair to the hue on the color-wheel that most closely resembled the remembered color at the probed location. The next trial began  $\sim 1$ s after participants clicked the mouse and this procedure was repeated 96 times per experimental condition (i.e. a total of 1440 trials per participant). Presentation of the 5 different color distances and 3 different encoding times was fully counterbalanced.

**Analyses.** We calculated the difference between the cued target color and the reported color (reported<sup>o</sup> – target<sup>o</sup>) on each trial. To investigate the systematic relationship between the cued color and the non-target color, we flipped the sign of the error such that the non-target color was always counter-clockwise to the cued target in the error distribution. The circular standard deviation was used to quantify subjects' response precision (i.e. larger deviations indicate less precision). Biases in subjects report were quantified by computing the proportion of responses on the “clockwise” side of the error distribution (i.e. the side opposite to that of the non-target). We centered this bias onto 0 to get a percentage score for the bias as follows (see also Figure 4b):

$$bias = \frac{responses\ away * 100}{total\ responses} - 50$$

We expect this bias metric to be roughly 0% if no biases exist, >0% if there is repulsion away from the non-target, and <0% if there is attraction toward the non-target. Note that this metric reflects relative repulsion/attraction biases rather than being an absolute metric, since potential “swap” errors (where the target and non-target colors are confused, and a subject mistakenly reports the non-target) would be counted as “attraction”. Thus, this metric is conservative to the extent that potential swap errors would inflate attraction biases and underestimate repulsion biases. To benchmark our model-free metrics of memory precision and bias, we also fit a von Mises (circular analogue of a normal distribution) to our error distributions using 2 parameters: standard deviation (*vmSD*) and bias (*μ*). We used repeated-measures analysis of variance to evaluate the impact of encoding time and color similarity on both the model-free (circular standard deviation and percentage bias metric) and estimated (*vmSD* and *μ*) parameters.

## **Results**



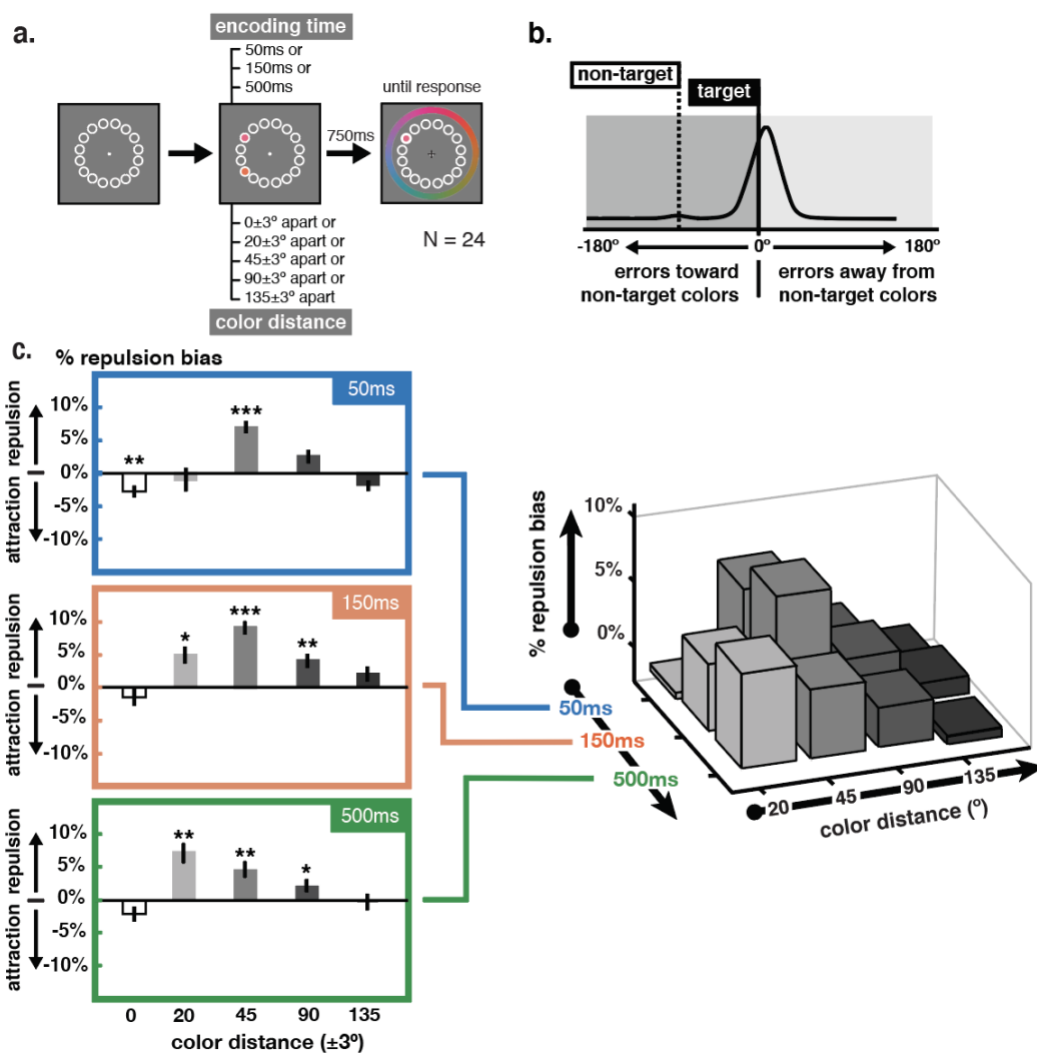
We confirmed that memory precision was higher at longer encoding times, with circular standard deviations of 26.5°, 25.0°, and 22.0° for encoding times of 50, 150 and 500ms, respectively ( $F_{(2,46)}=65.17$ ,  $p<0.001$ ). Memory precision also differed as a function of color distance, with circular standard deviations of 17.4°, 21.4°, 25.3°, 29.1° and 29.4° for color distances of 0°, 20°, 45°, 90° and 135°, respectively ( $F_{(4,92)}=69.49$ ,  $p<0.001$ ), showing increasingly noisy responses as two colors differed more.

To quantify the repulsion bias, we used our model-free bias metric (as discussed above), where values  $>0$  indicate repulsion, and values  $<0$  indicate attraction. We found differences in repulsion at longer encoding times, with biases of 0.8%, 3.4% and 2.4% for encoding times of 50, 150 and 500ms, respectively ( $F_{(2,46)}=9.19$ ,  $p<0.001$ ; compare the 3 panels on the left in Figure 4c). The amount of repulsion also differed as a function of distance in color space between the two memory items, with biases of -2.2%, 3.8%, 7%, 3% and -0.6% for color distances of 0°, 20°, 45°, 90° and 135°, respectively ( $F_{(4,92)}=13.14$ ,  $p<0.001$ ; compare values along the x-axis in the left panels in Figure 4c).

Importantly, there was an interaction between encoding time and color distance ( $F_{(8,184)}=3.78$ ,  $p<0.001$ ; Figure 4c). For example, the strongest repulsion bias shifted from 45° at the shortest encoding time (50ms) to 20° at the longest encoding time (500ms). This is in line with the idea that the maximum amount of repulsion depends on both color distance and representational precision. Note how two very similar colors presented at very short encoding times show a decreasing amount of repulsion (with repulsion disappearing when two items were 20° apart and shown for only 50ms). This pattern likely emerges because people are no longer able to individuate the two items, as shown in a control experiment (Appendix Figure 3). Interestingly, the repulsion of two memory representations away from one another is not a simple lateral shift, but instead

leads to significantly skewed response distributions (Appendix Figure 4).

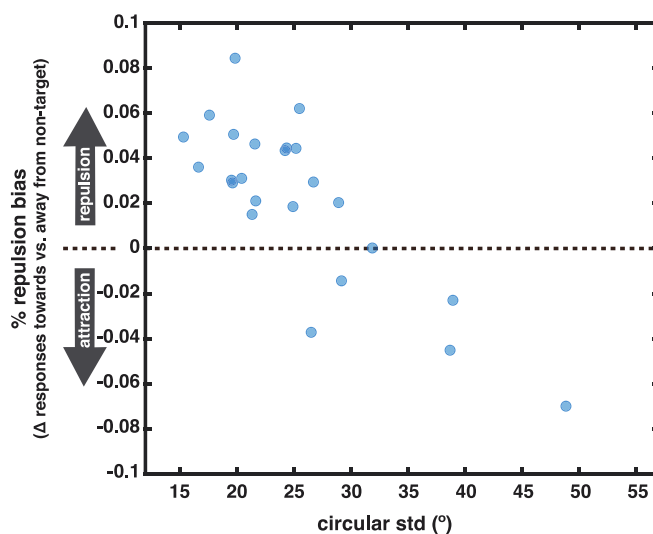
Together, these results are consistent with our framework and suggest that representations are biased to become more distinctive in order to maintain individuated representations (although in the limit people need to be able to dissociate item colors during encoding before any repulsion can occur). This means that with shorter encoding times we see maximal repulsion when two items are sufficiently distant in feature space (i.e. at  $45^\circ$  but not  $20^\circ$ ). It also means that when longer encoding time leads to representations that are more precise, items must be very similar (i.e. differ by  $20^\circ$  in color space) to achieve maximum repulsion.



**Figure 4.** a) In Experiment 4, participants remembered two memory items that were either 0°, 20°, 45°, 90° or 135° apart in color space (each with  $\pm 3^\circ$  of jitter), and that were briefly presented for either 50ms, 150ms or 500ms. Participants reported the color of the cued item (indicated by a thicker outline at one of the placeholder locations) by choosing the remembered color on a color-wheel. b) While non-targets could have a color that was either counterclockwise or clockwise in feature-space relative to the cued color, error distributions were constructed (for each subject and condition) by always plotting the non-target color as counter-clockwise from the cued color. This cartoon depicts one such error distribution. Attraction and repulsion biases were operationalized as the difference in the percentage of responses that were toward (dark gray shading) vs. away from (light gray shading) the non-target color. c.) The 3D bar plot (right) shows repulsion as a function of both encoding time (z-axis) and inter-item distance in color space (x-axis). Repulsion at each encoding time is replotted in the three sub-panels (left) to show the within-subject standard error ( $\pm 1$  SEM) for each condition, and to show the data from trials with a 0° inter-item difference (not shown in the 3D bar plot) where no repulsion or attraction should exist. Overall, repulsion biases were more prevalent when the two memory colors were more similar. Especially when encoding time increased, and responses become more precise, did the remembered colors need to be very similar to observe maximal repulsion. Single, double and triple asterisks indicate  $p \leq 0.05$ ,  $p \leq 0.01$  and  $p \leq 0.001$  respectively (tested against no-bias; uncorrected for multiple comparisons).

Note that the above analyses, based on non-parametric quantifications of precision and bias, were confirmed with an additional analysis based on the standard deviation and bias parameters of a von Mises distribution fit to the error distributions (Appendix Figure 5).

In Experiment 2 we had found that the degree of repulsion bias was related to the level of task engagement (Figure 2c). This indicated that a lack of effort was not the source of the repulsion biases found in that experiment. To make sure this finding was not due to the specific 2AFC or online nature of Experiment 2, we also analyzed the data from the current experiment, which was collected in the lab using a continuous report paradigm. Here, baseline performance was quantified by the circular standard deviation of each subject (with lower circular standard deviation indicating better performance), while bias was quantified by the percentage of responses away from non-target color (values  $>0$  indicating repulsion). We found strong negative correlation between circular standard deviation and bias (Pearson's  $\rho = -0.81$ ,  $p < 0.001$ , Bayes factor = 3872; Figure 5) supporting and extending our findings from Experiment 2. In the current analysis, the correlation is very prominent, possibly owing to the high number of trials (1440) per subject. Participants with better performance (smaller circular deviation) tended to have stronger repulsion bias (responses away from non-target colors were higher than 0), showing that repulsion biases are strongest in participants with the highest levels of task engagement.



**Figure 5.** A strong negative correlation between bias (y-axis) and memory performance (as indexed by the circular standard deviation on the x-axis) demonstrates that repulsion is stronger in participants whose performance is better. This replicates the correlation between task performance and magnitude of repulsion biases in Experiment 2 (Figure 2c; and see also a replication experiment in Appendix Figure 1), and clearly demonstrates that a lack of effort cannot explain repulsion biases.

### **Experiment 5: Repulsion biases grow with longer delays**

Finally, we tested whether repulsion biases become stronger with increasing memory noise. In Experiments 1–4, biases emerging during encoding cannot be dissociated from those emerging during the delay. Therefore, here we focus on memory noise that arises during the delay. To manipulate memory noise, we compared performance across different memory delay durations. Note that while some have argued that memory noise does not change as a function of delay interval (e.g., Huang & Sekuler, 2010; Magnussen & Greenlee, 1992; Regan & Beverley, 1985; Zhang & Luck, 2009, 2011), subsequent studies have since demonstrated that, with adequate power, representations do become noisier over time (Rademaker et al., 2018; Shin et al., 2017). We reasoned that if repulsion bias functions to keep two memory representations distinct, then this repulsion bias should grow stronger as the memory delay (and thus memory noise) increases. Alternatively, it is possible that when the two representations become increasingly noisy over time, responses may instead become biased toward the average of the two colors, and thus repel less, or even attract. We tested these predictions in an experiment where we manipulated delay duration (250ms, 750ms, or 5000ms; see Appendix Figure 6 for stimulus presentation details) as participants remembered two items. Encoding time was fixed at 250ms, and color distance between

the two items was fixed at 45° (i.e., values that yielded the largest repulsion bias in Experiment 4). Subjects recalled the target color using a continuous report paradigm. We quantified bias in a model-free manner as in Experiment 4.

## Methods

**Participants.** A total of 60 naïve participants were recruited using Amazon Mechanical Turk. For the control experiment (presented in Appendix Figure 7), an additional 50 naïve participants were recruited from Amazon Mechanical Turk. All participants provided their informed consent, and were paid approximately \$8 per hour for their time. Five participants out of 60 were excluded because of poor baseline performance (mean circular standard deviation more than 70° which was > 2 SD of the group). For the control experiment, 3 participants were excluded for the same reason.

**Stimuli & Procedure.** Stimuli and task procedures were identical to Experiments 1–2 (i.e., two stimuli at a 45° color distance were briefly shown at two of 12 placeholders on the screen and remembered over a delay before responding) with the following exceptions: There were no placeholders next to fixation, instead, there was always a light gray circle visible (237 pixel radius, 2 pixels wide, #d3d3d3 hex color) outside of the placeholders (see Appendix Figure 6). This grey circle turned into a randomly rotated color wheel during the response period (color wheel of the same dimensions as the grey circle). The two memory stimuli were presented for 250ms and participants remembered the color of each stimulus for a 250ms, 750ms, or 5000ms delay period. After the delay, one of the two colors was probed, and participants reported the cued color by moving a white circle along color wheel (i.e. via a continuous recall procedure as in Experiment 4). This procedure was repeated 60 times for each of the 3 delay period conditions (i.e., 180 trials

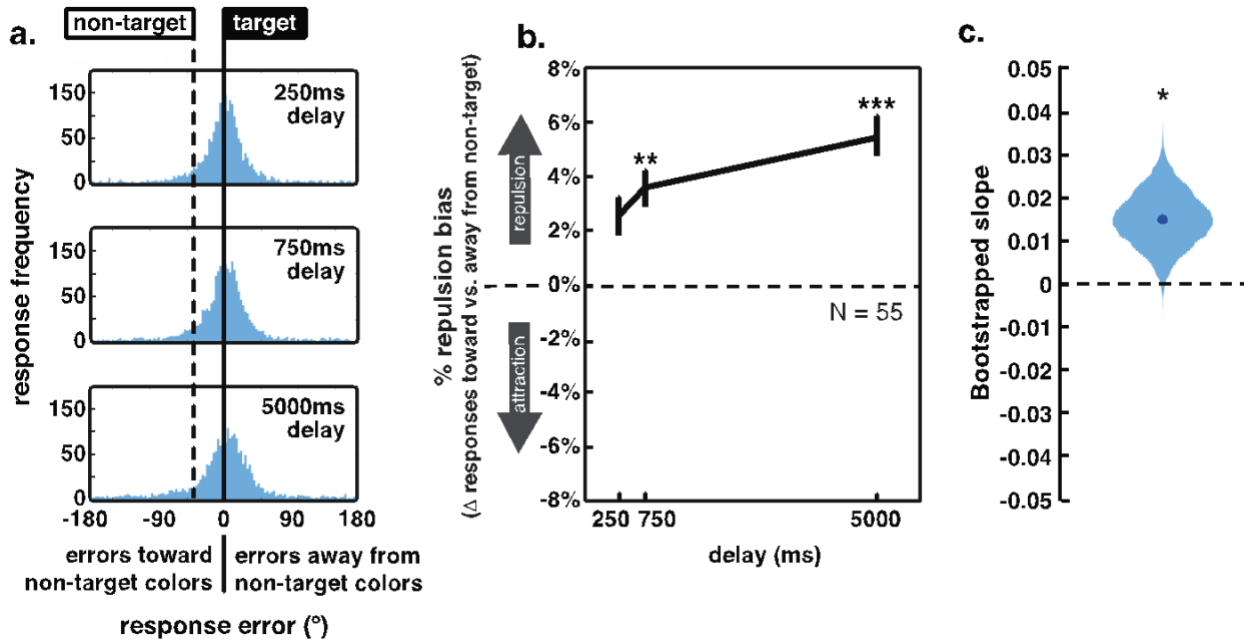
per participant in total). For the replication experiment (Appendix Figure 7), the procedure was identical, with the exception that stimuli were only presented for 150ms (instead of 250ms).

## Results

First, we found that the width of recall error distributions significantly differed across the three memory delays (Figure 6a), with circular standard deviations of 33.8°, 34.5° and 38.6° for delays of 250ms, 750ms and 5000ms, respectively ( $F_{(1,54)}=38.33$ ,  $p<0.001$ ). This is consistent with the notion that there is an increase in memory noise as items have to be remembered over longer delays. We also found that the repulsion bias grew monotonically with delay duration, from 2.5%, to 3.6%, and 5.6% for delays of 250ms, 750ms and 5000ms, respectively (Figure 6b and 6c;  $F_{(1,54)}=5.36$ ,  $p=0.025$ ), suggesting larger repulsion biases with increasing delay duration. This effect was replicated in a control experiment using an independent set of subjects (Appendix Figure 7) and cannot be explained by changes in swap rate with delay (i.e., swaps happen when subjects mistakenly report the non-target color instead of the target color; see Appendix Figure 8).

Thus, when two similar (but dissociable) items have to be remembered, we observe repulsion. As the items are held in memory for increasingly longer durations, they repel further apart as they become noisier (we do not observe a switch to attraction biases). The increase in repulsion with longer delays suggests that the repulsion bias is at least partly related to the storage of information in memory, and is not purely due to perceptual factors or response strategies.

647



648

649 **Figure 6.** a.) At increasing delays, error distributions become wider (larger circular standard  
 650 deviation), indicating increasing memory noise. The distributions also reveal a high number of  
 651 responses biased away from the non-target. b.) The proportion of responses biased away from the  
 652 non-target, when quantified for the three delay-duration conditions, revealed a repulsion bias that  
 653 grew monotonically stronger as the delay time increased. Error-bars represent  $\pm 1$  within-subject  
 654 SEM. Double and triple asterisks indicate  $p \leq 0.01$  and  $p \leq 0.001$  respectively (tested against no-  
 655 bias; uncorrected for multiple comparisons). c.) To assess the increase of repulsion bias with delay,  
 656 one can fit a line through the three points in (b) and calculate the slope – a positive slope indicating  
 657 an increasing repulsion. Shown here is a distribution plot of bootstrapped slopes (5000 iterations  
 658 of resampling with replacement). The single asterisk indicates  $p < 0.05$ . This confirms a statistically  
 659 robust effect, with repulsion bias growing as a function of delay duration.

660

### An adaptive framework



In five main experiments (and three control experiments), we found that memory representations were repelled from each other when the memoranda were highly similar (Experiments 1–2), when memory representations were noisier (Experiment 4), and more when representations were remembered over longer delay intervals (Experiment 5). We confirmed that these effects do not simply reflect straightforward demand effect or straightforward response biases, and they hold across different experimental paradigms. Moreover, we showed that participants with excellent performance and task engagement showed large repulsion biases, suggesting that these biases do not simply reflect a lack of effort to precisely remember the colors. Finally, when memory load increased and it was harder for participants to maintain individuated representations, memory biases reversed from repulsion to attraction (Experiment 3).

In this framework, we focus specifically on memory biases between amongst two or more simultaneously presented memory stimuli – which is different from categorical biases and the serial dependence effect. Overall, the experiments we presented here argue against the idea that some studies find attraction biases and some find repulsion biases purely as an artifact of using different stimuli. They also argue against the idea that such biases arise primarily from some form of motor-response strategies.

We instead suggest that these inter-item biases can be thought of as adaptive distortions by our memory system, designed to reduce error. The broad framework we adopt is that visual working memory faces at least two distinct problems. First, the capacity of working memory is limited, and when more items must be stored, they are stored with more noise (Bays & Husain, 2008; Ma et al., 2014; Zhang & Luck, 2008). In such cases, summary statistics or other ways of blending across items can be used to somewhat improve memory of individual items (Brady et al., 2011; Brady & Alvarez, 2015; Lew & Vul, 2015). The second problem is that access to memories is not automatic

and not independent of cues and context. Instead, there can sometimes be confusion between items that arises when items are similar in context and features. Indeed, prominent process models of working memory that focus on feature-location binding predict that items are automatically blended if they are similar (Swan & Wyble, 2014; Lin & Oberauer, 2017). Avoiding such confusion is important to reducing error when such blending is not optimal (e.g., when item representations are not noisy, but are similar and so likely to be blended).

We do not attempt to make a precise quantitative model that could be fit to performance on our tasks. However, it is useful to formalize these ideas to see if it is plausible that reducing error is the overall goal of attraction and repulsion, and to ask whether the factors that affect the magnitude of each problem determine when we should expect attraction and repulsion to be strong or weak. We do that here.

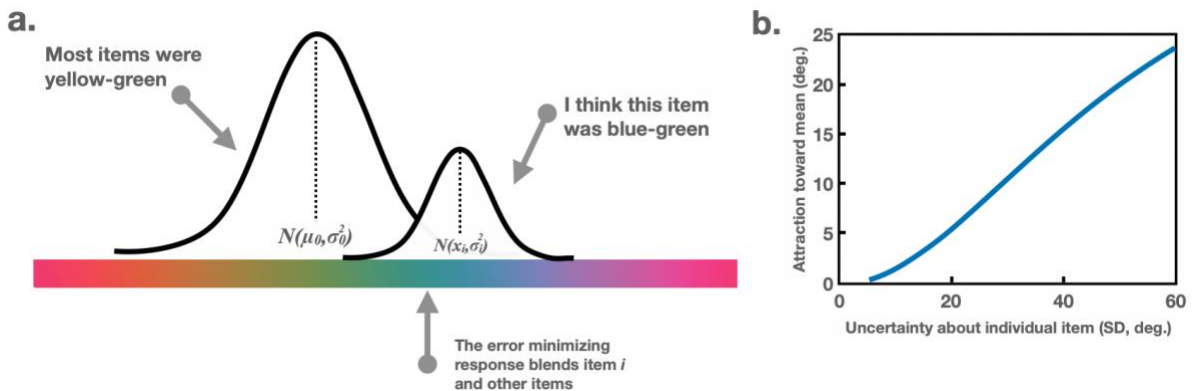
**Attraction.** For the purposes of considering attraction, we assume that the information subjects have about the display is (1) information about the entire set of colors (i.e., participants know if the items were all red), and (2) information about each specific item, with, for now, the simplifying assumption that there is no confusion as to which color goes with which item (i.e., when a subject remembers the color of the  $i^{th}$  item, they never mistakenly retrieve the color of the  $j^{th}$  item). Given these assumptions, we can predict if memory distortions would be optimal to minimize error if subsequently asked to report the feature associated with an individual item.

In general, the observer has an estimate of the mean ( $\mu_0$ ) and the uncertainty ( $\sigma_0$ ) about the color of the entire set of colors – i.e., the ensemble – and a noisy estimate of the color of a given item (with mean  $x_i$ , and uncertainty  $\sigma_i$ ). This gives rise to a hierarchical situation because the color of each item is part of the overall set of colors. Given this hierarchy, the optimal error-minimizing

color to assign to an item follows from hierarchical Bayesian models, which for the simplest case of two nested normal distributions is:

$$optimal = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_0^2} \mu_0 + \frac{\sigma_0^2}{\sigma_i^2 + \sigma_0^2} x_i$$

That is, remembering and reporting colors according to this rule results in less error on average than reporting based only on your memory of an individual item (i.e., reporting only  $x_i$ ). However, the cost for this increased accuracy is distortion: following this rule results in attraction toward the mean color of the set. Intuitively, this distortion actually increases performance because if there is a noisy sample of a given color that is green-ish blue, but the mean of the entire set of colors is yellow-ish-green, it is more likely the sample was inaccurate by being too blue as opposed to being too green (Figure 7a). Thus, when taking into account information from both levels, the optimal color to report is slightly greener than the actual sample associated with that one color alone. That is, reporting colors in this way is actually more accurate – resulting in less error on average – than reporting the color you believe an item to be without pulling it toward the average of the set (Brady & Alvarez, 2011; Huttenlocher et al. 2000).



**Figure 7.** a) Attraction is the error-minimizing thing to do when you have uncertainty about an individual item, but know how that item related to the entire set. Blending the information about the individual item with the information about the other similar items improves performance in this circumstance. Intuitively, this distortion actually increases performance because if there is a noisy sample of a color that is green-ish blue, but the mean of the entire set of colors is yellow-ish-green, it is more likely the sample was inaccurate by being too blue as opposed to inaccurate by being too green. Thus, when taking into account information from both individual item and group levels, the optimal color to report is slightly greener than the actual sample. b) The amount of attraction that is optimal depends on several factors, but it most clearly depends on the uncertainty about the individual item you are probed on: The more uncertain you are about its color (the wider the normal distribution associated with it), the more attraction is optimal.

Three aspects of the optimizing equation above are relevant for attraction in a typical working memory tasks. For a given set size, more uncertainty about each item will lead to a greater reliance on information about the entire set as opposed to information about the specific item (as  $\sigma_i$  goes up, you weigh  $x_i$  less and  $\mu_0$  more, Figure 7b). Thus, in general, manipulations that increase uncertainty about individual items, such as decreasing encoding time or increasing delay time (Rademaker et al., 2018; Shin et al., 2017; Schurgin et al., 2020), should result in more attraction if all else is held equal.

The second relevant factor is related to the clustering of individual item values in feature space. Consider a display with a single well-formed cluster of colors that are all some shade of yellow-

ish green, as illustrated in Figure 7<sup>1</sup>. If all the items are part of this single cluster, then as the colors get more similar to each other, the uncertainty ( $\sigma_0$ ) associated with the group mean will go down and the group color will have a bigger influence on the optimal decision. When  $\sigma_0$  gets very small, as would happen if the colors were all very similar, this factor assigns nearly all the weight to the group color and none to individual items, regardless of the uncertainty associated with the individual items.

A final relevant factor for attraction is that increases in memory set size don't *just* increase the uncertainty associated with each item (i.e., drive up  $\sigma_i$ , which would increase attraction). Instead, larger set sizes also lead to more precise estimates of the mean and less uncertainty about the entire set of colors ( $\mu_0$  and  $\sigma_0$ ), since there are more samples to constrain these values. Thus, if the items are relatively tightly clustered on the color wheel at all set sizes, then, as set size goes up, your certainty about the color of the whole set (the ensemble color) goes up (in the same way that having more trials would decrease the standard error of your estimate of the mean in a typical experimental setting). This decreases  $\sigma_0$ , exacerbating the attraction effect even more than just increasing  $\sigma_i$  alone.

As a result, at larger set sizes, and particularly when the items are tightly clustered in feature space, this framework predicts a stronger attraction effect than at smaller set sizes, even with similar clustering. This follows because there are two factors driving attraction – as set size goes up, certainty about the average color of the set goes up, and the item representations themselves get noisier. In contrast, for small set sizes, only in very noisy individual-item conditions or in

---

<sup>1</sup> Of course, more complex scenarios exist: i.e., if 3 items are red-ish and 3 are blue-ish on a display of 6 items, participants may form 2 clusters and items may be selectively attracted toward the cluster they are part of (Chunharas & Brady, 2019), but we set that aside here for simplicity.

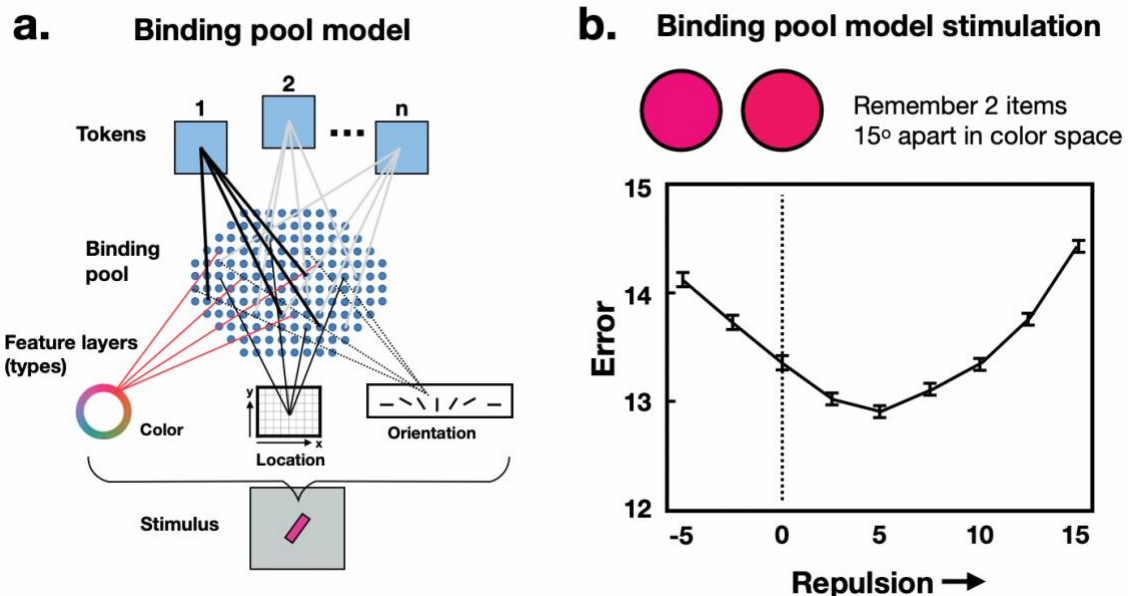
conditions where the set of items are so similar that  $\sigma_0$  is much smaller than  $\sigma_i$  – would the framework predict any appreciable attraction effects, even though such attraction effects should be robust in displays at higher set sizes when there is clustering of the features.

**Repulsion.** In contrast to attraction effects, which should be amplified at large set sizes, our framework suggests that repulsion biases should be error-reducing primarily at small set sizes when items are highly confusable.

When considering attraction biases, our model assumed that when subjects seek to retrieve information about color  $i$ , they can successfully retrieve only information about color  $i$  (i.e.,  $x_i$  reflects only color  $i$ ). However, human memory in general is based on cued-retrieval: content-based access rather than direct access (Gallistel & King, 2011). That is, unlike a computer, which stores an item in a given spot in RAM and then accesses that exact address again later, human memories are retrieved by matching operations based on content. As a result, more similar memories are more likely to be confused at retrieval or to interfere with each other. While widely recognized in long-term memory (e.g., Criss, Malmberg & Shiffrin, 2011), this aspect of memory retrieval is typically also present in models of visual working memory when they focus on cued-retrieval (Swan & Wyble, 2015; Oberauer & Lin, 2017).

Importantly, such models of memory blend together the representation of different items *all the time* because of interference between memory representations, as a natural consequence of cued retrieval. For example, when storing just two item similar items, the “binding pool” model of Swan and Wyble (2015) predicts that the two items will be attracted to each other significantly (see Figure 8). As we have seen, however, this is not in any way optimal: with strong memories, and few items to give rise to a tight ensemble distribution, attraction will not reduce error.

Thus, in this scenario, an adaptive system must balance the need to avoid overlap between item representations and the need to maintain an accurate memory. If the representations are encoded veridically, they will have significant interference and be blended inappropriately. If they are represented as more distinct from each other than they really were, this will come with its own reduction in accuracy although it will also reduce inappropriate blending. The memory system must strike a balance, with systematic repulsion to offset the blending that would otherwise occur, but not so much repulsion that it impairs accuracy overall. We can simulate this in the Binding Pool model (with all of the default parameters) simply by adding an attraction or repulsion step to the encoding process, and seeing what happens to (1) the resulting bias, and (2) error. In the binding pool model, the error minimizing amount of repulsion for storing 2 items that are 15 degrees apart in color space is  $\sim 5$  deg (Figure 8). More repulsion is required to minimize error when items are more similar and/or when items are represented with more uncertainty.



**Figure 8.** a) Schematic of the Binding Pool model, reproduced from Swan and Wyble (2014). In this model, each stimulus evokes an activation in a set of feature layers (here: color, location, and

orientation). These features are then encoded into a shared binding pool layer and tied to a particular ‘token’. When provided a cue at test (like the item location) the associated token can be activated and the color or orientation retrieved. Notably, the binding pool layer, which is shared between all items and the source of the capacity limits of the system, also results in the items features being necessarily blended (e.g., Swan & Wyble, 2014, Figure 11), even when only 2 items are represented. Thus, by default, this model, like many others, always predicts attraction between memory representations. b) We simulated what would happen if repulsion was added to the encoded information in the binding pool model, to provide a concrete case study for how repulsion could be used to overcome the blending inherent in a model such as the Binding Pool model, and reduce overall error. In particular, we asked the model to store and recall 2 items that were 15° apart in color space. As part of the encoding stage, we added an additional step that introduced repulsion of the colors of the two items before they were put in the binding pool. In 100 simulations of the model at each of 9 levels of additional attraction or repulsion added at encoding, we calculated the model’s error. We found that error was minimized when the items were repelled away from each other by ~5° before being entered into the binding pool layer.

**Summary: attraction and repulsion.** Our adaptive framework holds that attraction biases (when memory is very noisy) can be understood as optimal using a straightforward hierarchical Bayesian integration model. Effectively, attraction biases arise because integrating summary statistic information results in reduced error even if it results in systematic distortion (Brady & Alvarez, 2011; Huttenlocher et al., 2000). This framework makes a clear set of predictions about when attraction should occur: when items are clustered in color space and individual items are associated with a higher degree of uncertainty than the ensemble color. In practice, this ends up happening primarily when set size is high, or when set size is low but items are very similar relative to the



item-level uncertainty. In contrast, repulsion bias can be understood as balancing the avoidance of overlapping representations with the need for accurate representations. Insofar as overlap is present and attraction is not adaptive, this model predicts that items should repel from each other. At low set sizes, this means that repulsion is expected whenever items are similar enough, and uncertainty high enough, that the memory representations overlap substantially. At high set sizes, the extent to which repulsion will be useful in lowering error is severely reduced by the crowding of the feature space with other items, and the fact that attraction and repulsion pull in different directions, with attraction likely being dominant. Overall, we believe this adaptive framework can providing a guiding theory for conceiving of when attraction and repulsion arise in memory.

## **General Discussion**

Our memory is susceptible to systematic distortions. Even across short periods of time, specific memories become affected by the overarching categories that memory items belong to (categorical biases) or by information viewed in the immediate past (serial dependence). The research presented here focused on a different kind of distortion: inter-item distortions that occur in memory when we try to hold multiple items in mind. When encoding and remembering multiple items at once, mnemonic representations can be subject to systematic distortions that can make items either more separable (repulsion biases) or more similar (attraction biases). While both types of inter-item distortion are well documented, it is not clear when repulsion or attraction will occur as a function of the type of information being remembered and current task demands. Here, we examined when each type of bias arises. We found that memory representations were repelled away from each other when the memoranda were highly similar (Experiments 1–2), when memory representations were noisier (Experiment 4), and when representations were remembered over longer delay intervals (Experiment 5). We confirmed that these effects do not simply reflect

straightforward response strategies, and occur in many distinct tasks, and we showed that high-performing participants showed larger repulsion biases which suggests that these biases do not simply reflect a lack of effort to precisely remember the colors. Finally, when memory load increased and it was harder for participants to maintain individuated representations, memory biases reversed from repulsion to attraction (Experiment 3).

Past work has found evidence for attraction biases (Brady et al., 2011; Brady & Alvarez, 2011, 2015; Dubé et al., 2014; Dubé & Sekuler, 2015; Huang & Sekuler, 2010; Lew & Vul, 2015; Lorenc et al., 2018; Utochkin & Brady, 2020), repulsion biases (O’Toole & Wenderoth, 1977; Rauber & Treue, 1998; Suzuki & Cavanagh, 1997; Scotti et al., 2021), or both (Bae & Luck 2017; Rademaker et al 2015; Golomb 2015). Our model and empirical work identifies several key factors that drive these effects and provides evidence that both can arise even in similar paradigms. This is important, as using highly comparable paradigms and memory for a single feature (color) argues against the more mundane explanation that differences in stimulus features (such as orientation in Bae & Luck, 2017; Dubé & Sekuler, 2015; Huang & Sekuler, 2010; Lorenc et al., 2018; Utochkin & Brady, 2020, spatial location in Lew & Vul, 2015; Suzuki & Cavanagh, 1997, motion direction in Kang & Choi, 2015, or color in Brady & Alvarez, 2015; Golomb, 2015) lead to attraction in some studies and repulsion in others.

Here, we tested the general account that when subjects were trying to encode items in a memory display, repulsion and attraction were driven largely by the inter-item relationship between memoranda. We proposed a way to conceive of these biases and when they arise based on adaptive framework. In particular, we suggested that these biases may be natural consequences of the memory system attempting to minimize memory error, if systematic distortion is adaptive in particular circumstances (Schacter et al., 2011). When many similar items are present and so

868 memories for individual items are noisy, attraction biases are known to be optimal for minimizing  
869 error (e.g., Brady & Alvarez, 2011). Repulsion biases can also reduce error in some situations,  
870 making them adaptive. In particular, if the items would naturally be blended or confused by our  
871 memory system (Swan & Wyble, 2014; Oberauer & Lin, 2017), then repulsion can reduce this  
872 tendency and reduce error when we have strong and distinct item memories. Importantly, these  
873 biases are not simply inherited from perceptual processing: as noise accumulates in memory over  
874 time (reducing the signal-to-noise if memory items), and the need to keep memoranda distinct  
875 grows, a corresponding increase in the repulsion bias is observed. Importantly, very recent work  
876 (performed since the first presentation of the experiments in the current paper) has confirmed  
877 various key aspects of our framework: As memories get weaker, biases switch from repulsion to  
878 attraction (Lively, Robinson, & Benjamin, 2021), and repulsion biases increase with longer  
879 memory delays (Scotti et al., 2021).

880 Based on these results, the degree and type of bias likely depends on the overall discriminability  
881 of a stimulus feature under investigation (such as color, space, orientation, etc.): If features are  
882 very readily discriminable, then repulsion will only occur when two items are very similar. Poorly  
883 discriminable features will need to differ more before they are susceptible to repulsion. In other  
884 words, the data suggest that the extent and type of bias will directly map onto the just-noticeable-  
885 differences (JND) of a given stimulus feature (and of individual subjects). Using JND as a standard  
886 unit might be an interesting approach that allows us to compare the various effects previously  
887 reported. Even though we tried to use the same stimulus feature and investigate various task  
888 manipulation in this paper, it is still not easy to compare the results with previously reported  
889 findings where many interesting inconsistencies are waiting to be explored.

Even though our experiments were designed to rule out specific forms of response strategy, it is still possible that our findings could be explained by other response strategies that closely resemble the framework proposed here. For one example, it is possible that foils in Experiments 1 and 2 were too similar to the true answer, and that subjects might choose between the two response options by the process of elimination (i.e., “I did not know which one was the target color so I am going to choose the one that is less similar to the non-target”). In this hypothetical case, it is still unclear how the subject would know which response option is less similar to non-target without knowing which one is more similar to the target – making it a possible but implausible strategy. We would like to note how recent neuroscience studies have demonstrated that memory representations drift over time (Compte, Brunel, Goldman-Rakic, & Wang, 2000) – a process which is not likely to be susceptible to response strategies.

**Mechanisms of memory biases.** Attraction biases can occur both in absolute stimulus space (e.g., towards particularly salient colors (Bae et al., 2015)) or arise from the similarity between items in an individual display (as in the current work). These attraction biases are straightforwardly explained as arising from gist-based or ensemble-based representations, and a combination of these global representations with item specific representations. Many models claim that attraction biases are the result of weighting the representation of each object towards the “summary” of the set to achieve a more stable memory at the expense of maintaining distinctions between individual items (Brady & Alvarez, 2011), or via blending items together if they are similar (e.g., Swan & Wyble, 2014; Oberauer & Lin, 2017). The category learning literature has carefully demonstrated that this is in general an adaptive strategy that serves to minimize error (Huttenlocher et al., 2000).

Repulsion biases have traditionally been more difficult to understand. Previous studies have shown that repulsion biases occur when two items are task-relevant and proximal in feature space

(Rademaker et al., 2015; Bae & Luck, 2017, Golomb 2015). However, the benefits of repulsion biases are still unclear. Here, we suggest that repulsion biases serve to maximize distinctiveness between items, when individual item representations are strong but items are similar enough to be more difficult to distinguish. This helps reduce blending between items that naturally occurs in the memory system (Swan & Wyble, 2013; Oberauer & Lin, 2017). Any factor that affects distinctiveness in memory should thus impact the degree of repulsion biases (e.g., encoding time, feature similarity, memory delay). Interestingly, previous work has frequently found repulsion not only between items, as in the current work, but in absolute terms as well. For example, when asked to remember an orientation that is near, but not quite at, vertical, people will systematically report the orientation as further from vertical than it really was (Jastrow, 1892; Smith, 1962). One framework that has been useful to understand these absolute biases is to dissociate the physical space of the stimuli (e.g., absolute orientation) from the psychological representation of the stimuli (e.g., people may over-weight certain values in a systematic manner). A clear example of a warped psychological space is the massive overrepresentation of vertical and horizontal orientations, presumably to efficiently code environmental regularities (Girshick et al., 2011; Wei & Stocker, 2015). Accounting for this selective over-representation of certain stimulus values in psychological space can explain biases like repulsion from cardinal axes, and the reason why these biases tend to arise in parts of stimulus space where discrimination thresholds are lowest (e.g., the most overrepresented stimulus values) (Wei & Stocker, 2015, 2017).

This conception of psychological space is designed to address long-term biases that are likely crystalized in the neural architecture of the visual system, whereas the biases we examined in the current work are more dynamic. Despite the apparent disconnect, a common mechanism such as the warping of psychological space may be at play in both stable long-term phenomena and in

more dynamic short-term regimes. In the current work, this would mean that a strong representation of an item “stretches” the psychological representation of stimulus space near that item, resulting in repulsion of other items in a manner similar to how cardinal orientations repel nearby items. This is consistent with other short-term effects: For instance, spatial judgments are distorted by top-down factors such that there is repulsion bias away from currently attended locations (Suzuki & Cavanagh, 1997). Attention, which leads to well-documented changes in visual sensitivity (i.e., lower discrimination thresholds, see Carrasco, 2011), may also adaptively bias perception and memory on demand, as biases typically manifest when discrimination thresholds are low across a variety of visual features such as orientation, motion direction, spatial frequency, and visual speed (see Zhang & Luck 2011 for a summary). Thus, attention amplifying discrimination at a single color may strengthen the representational space there, resulting in repulsion. In sum, conceptions of psychological space, and how it is distorted when particular sets of stimuli are over-represented, may be a useful framework for considering biases at all possible time scales (see also Schurgin et al., 2020, for details on the widely applicable utility of this concept).

What might be the neural substrates of biased representations? When a task requires focal attention to a small set of items to remember – as is the case in paradigms that create repulsion bias – the discriminability of the relevant items can be improved by biasing responses in early visual cortex to maximize the separability of their corresponding neural representations. For example, attention to highly similar features, akin to remembering two highly similar colors in Experiments 1, 2, and 5, has been shown to modulate neurons tuned just away from the attended features. This ‘off-target’ gain can improve performance because neurons tuned away from the attended features undergo the largest change in firing rates because the two features fall along the steepest part of

their bell-shaped tuning curves. In turn, this off-target gain gives rise to systematic biases in behavioral reports such that people see stimuli as repelled from the actual feature values (Jazayeri & Movshon, 2007; Navalpakkam & Itti, 2007; Scolari & Serences, 2009). Such repulsion would be expected if the off-target gain happening in early visual cortex was interpreted as a veridical representation of the world at higher stages of processing. While previous work in this domain has focused on selective attention to continuously present stimuli, a similar type of modulation in the domain of working memory might give rise to repulsive biases in mnemonic representations. Indeed, repulsion biases grow with delay only when a memory is actively held in mind (but disappears when an attention-demanding task is performed during the delay), suggesting that the repulsion bias is not a product of some passive process, but instead requires active maintenance (Scotti et. al., 2021). While speculative, this type of adaptive neural modulation may map onto the psychological space framework, such that changes in the discriminability of stimuli in early visual cortex – either due to a lifetime of experience or to dynamic changes in the focus of attention – lead to a warping of perception and memory.

## References

The data and codes are available at

[https://osf.io/qp6xk/?view\\_only=0559769c587c4c8294288451e8af239e](https://osf.io/qp6xk/?view_only=0559769c587c4c8294288451e8af239e).

Adam, K. C., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive psychology*, 97, 79-97.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3), 122-131.

Alvarez, G. A., & Cavanagh, P. (2005). Independent resources for attentional tracking in the left and right visual hemifields. *Psychological Science*, 16(8), 637-643.

Bae, G.-Y., & Luck, S. J. (2017). Interactions between visual working memory representations. *Attention, Perception & Psychophysics*, 79(8), 2376-2395.

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

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 617-629.

Bartlett, F. C. (1932). Some Experiments on the Reproduction of Folk-Stories. *Folklore*, 31(1), 30-47.

Bays, P. M. (2015). Spikes not slots: noise in neural populations limits working memory. *Trends in Cognitive Sciences*, 19(8), 431-438.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10), 7.1-11.

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



997 Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory:  
 998 ensemble statistics bias memory for individual items. *Psychological Science*, 22(3), 384–  
 999 392.

1000 Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal  
 1001 hierarchically structured memory representations. *Journal of Vision*, 15(15), 6.

1002 Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond  
 1003 individual items and toward structured representations. *Journal of Vision*, 11(5), 4.

1004 Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436.

1005 Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Research*, 51(13), 1484–1525.

1006 Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition  
 1007 memory. *Journal of Memory and Language*, 64(4), 316–326.

1008 Cohen, M. A., Rhee, J. Y., & Alvarez, G. A. (2016). Limits on perceptual encoding can be  
 1009 predicted from known receptive field properties of human visual cortex. *Journal of*  
 1010 *Experimental Psychology: Human Perception and Performance*, 42(1), 67–77.

1011 Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and  
 1012 network dynamics underlying spatial working memory in a cortical network model.  
 1013 *Cerebral cortex*, 10(9), 910–923.

1014 Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate  
 1015 recall. *Journal of Experimental Psychology*, 58(1), 17–22.

1016 Dubé, C., & Sekuler, R. (2015). Obligatory and adaptive averaging in visual short-term memory.  
 1017 *Journal of Vision*, 15(4), 13.

1018 Dubé, C., Zhou, F., Kahana, M. J., & Sekuler, R. (2014). Similarity-based distortion of visual  
 1019 short-term memory is due to perceptual averaging. *Vision Research*, 96, 8–16.

1020 Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*,

1021 17, 738.

1022 Freyd, J. J., & Johnson, J. Q. (1987). Probing the time course of representational momentum.

1023 *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(2), 259–268.

1024 Gallistel, C. R., & King, A. P. (2011). *Memory and the computational brain: Why cognitive*

1025 *science will transform neuroscience* (Vol. 6). John Wiley & Sons.

1026 Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation

1027 perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7), 926–

1028 932.

1029 Golomb, J. D. (2015). Divided spatial attention and feature-mixing errors. *Attention, Perception*

1030 *& Psychophysics*, 77(8), 2562–2569.

1031 Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces.

1032 *Journal of Experimental Psychology: Human Perception and Performance*, 35(3), 718–734.

1033 Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in*

1034 *Cognitive Science*, 1(1), 189–202.

1035 Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: two classes of

1036 attractors at work. *Journal of Vision*, 10(2), 24.1–27.

1037 Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: prototype

1038 effects in estimating spatial location. *Psychological Review*, 98(3), 352–376.

1039 Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus

1040 judgment? *Journal of Experimental Psychology: General*, 129(2), 220–241.

1041 Jastrow, J. (1892). Studies from the University of Wisconsin: on the judgment of angles and

1042 positions of lines. *The American Journal of Psychology*, 5(2), 214–248.

1043 Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural

1044 populations. *Nature Neuroscience*, 9(5), 690–696.

1045 Jazayeri, M., & Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory  
 1046 decoding. *Nature*, 446(7138), 912–915.

1047 Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A dynamic neural field model of  
 1048 visual working memory and change detection. *Psychological Science*, 20(5), 568–577.

1049 Kang, M.-S., & Choi, J. (2015). Retrieval-Induced Inhibition in Short-Term Memory.  
 1050 *Psychological Science*, 26(7), 1014–1025.

1051 Koutstaal, W., Verfaellie, M., & Schacter, D. L. (2001). Recognizing identical versus similar  
 1052 categorically related common objects: further evidence for degraded gist representations in  
 1053 amnesia. *Neuropsychology*, 15(2), 268–289.

1054 Lew, T. F., & Vul, E. (2015). Ensemble clustering in visual working memory biases location  
 1055 memories and reduces the Weber noise of relative positions. *Journal of Vision*, 15(4), 10.

1056 Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory.  
 1057 *Psychological review*, 124(1), 21.

1058 Loftus, E. F. (2005). Planting misinformation in the human mind: a 30-year investigation of the  
 1059 malleability of memory. *Learning & Memory*, 12(4), 361–366.

1060 Lorenc, E. S., Sreenivasan, K. K., Nee, D. E., Vandenbroucke, A. R. E., & D’Esposito, M.  
 1061 (2018). Flexible Coding of Visual Working Memory Representations during Distraction.  
 1062 *Journal of Neuroscience*, 38(23), 5267–5276. [https://doi.org/10.1523/jneurosci.3061-](https://doi.org/10.1523/jneurosci.3061-17.2018)  
 1063 17.2018

1064 Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and  
 1065 conjunctions. *Nature*, 390(6657), 279–281.

1066 Magnussen, S., & Greenlee, M. W. (1992). Retention and disruption of motion information in  
 1067 visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and*  
 1068 *Cognition*, 18(1), 151–156.

1069 Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature*  
1070 *Neuroscience*, 17(3), 347–356.

1071 Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, 53(4),  
1072 605–617.

1073 Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory.  
1074 *Psychological Review*, 124(1), 21–59.

1075 O’Toole, B., & Wenderoth, P. (1977). The tilt illusion: repulsion and attraction effects in the  
1076 oblique meridian. *Vision Research*, 17(3), 367–374.

1077 Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers  
1078 into movies. *Spatial Vision*, 10(4), 437–442.

1079 Purushothaman, G., & Bradley, D. C. (2005). Neural population code for fine perceptual  
1080 decisions in area MT. *Nature Neuroscience*, 8(1), 99–106.

1081 Rademaker, R. L., Bloem, I. M., De Weerd, P., & Sack, A. T. (2015). The impact of interference  
1082 on short-term memory for visual orientation. *Journal of Experimental Psychology: Human*  
1083 *Perception and Performance*, 41(6), 1650–1665.

1084 Rademaker, R. L., Park, Y. E., Sack, A. S., & Tong, F. (2018). Evidence of gradual loss of  
1085 precision for simple features and complex objects in visual working memory. *Journal of*  
1086 *Experimental Psychology: Human Perception and Performance*, 44(6), 925–940.

1087 Rauber, H. J., & Treue, S. (1998). Reference repulsion when judging the direction of visual  
1088 motion. *Perception*, 27(4), 393–402.

1089 Regan, D., & Beverley, K. I. (1985). Postadaptation orientation discrimination. *Journal of the*  
1090 *Optical Society of America. A, Optics and Image Science*, 2(2), 147–155.

1091 Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not  
1092 presented in Lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,

1093 21(4), 803.

1094 Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: an adaptive  
 1095 perspective. *Trends in Cognitive Sciences*, 15(10), 467–474.

1096 Schooler, J. W., Foster, R. A., & Loftus, E. F. (1988). Some deleterious consequences of the act  
 1097 of recollection. *Memory & Cognition*, 16(3), 243–251.

1098 Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified  
 1099 theory of visual memory strength. *Nature human behaviour*, 4(11), 1156–1172.

1100 Scolari, M., & Serences, J. (2010). Estimating the shape of the feature-based attentional gain  
 1101 function. *Journal of Vision*, 8(6), 996–996.

1102 Scolari, M., & Serences, J. T. (2009). Adaptive allocation of attentional gain. *The Journal of*  
 1103 *Neuroscience*, 29(38), 11933–11942.

1104 Scotti, P. S., Hong, Y., Leber, A. B., & Golomb, J. D. (2021). Visual working memory items  
 1105 drift apart due to active, not passive, maintenance. *Journal of Experimental Psychology:*  
 1106 *General*. Advance online publication. <https://doi.org/10.1037/xge0000890>

1107 Serences, J. T. (2016). Neural mechanisms of information storage in visual short-term memory.  
 1108 *Vision Research*, 128, 53–67.

1109 Shin, H., Zou, Q., & Ma, W. J. (2017). The effects of delay duration on visual working memory  
 1110 for orientation. *Journal of Vision*, 17(14), 10.

1111 Smith, M. A., Majaj, N. J., & Anthony Movshon, J. (2005). Dynamics of motion signaling by  
 1112 neurons in macaque area MT. *Nature Neuroscience* 8(2), pp. 220–228.  
 1113 <https://doi.org/10.1038/nn1382>

1114 Smith, S. L. (1962). Angular estimation. *The Journal of Applied Psychology*, 46(4), 240–246.

1115 Spencer, J. P., & Hund, A. M. (2002). Prototypes and particulars: geometric and experience-  
 1116 dependent spatial categories. *Journal of Experimental Psychology: General*, 131(1), 16–37.

1117 Störmer, V. S., Alvarez, G. A., & Cavanagh, P. (2014). Within-hemifield competition in early  
1118 visual areas limits the ability to track multiple objects with attention. *The Journal of*  
1119 *Neuroscience*, 34(35), 11526–11533.

1120 Suzuki, S., & Cavanagh, P. (1997). Focused attention distorts visual space: an attentional  
1121 repulsion effect. *Journal of Experimental Psychology: Human Perception and Performance*,  
1122 23(2), 443–463.

1123 Swan, G., & Wyble, B. (2014). The binding pool: A model of shared neural resources for distinct  
1124 items in visual working memory. *Attention, Perception, & Psychophysics*, 76(7), 2136–2157.

1125 Tong, K., Dubé, C., & Sekuler, R. (2019). What makes a prototype a prototype? Averaging  
1126 visual features in a sequence. *Attention, Perception & Psychophysics*, 81(6), 1962–1978.

1127 Underwood, B. J., Ekstrand, B. R., & Keppel, G. (1965). An analysis of intralist similarity in  
1128 verbal learning with experiments on conceptual similarity. *Journal of Verbal Learning and*  
1129 *Verbal Behavior*, 4(6), 447–462.

1130 Utochkin, I. S., & Brady, T. F. (2020). Individual representations in visual working memory  
1131 inherit ensemble properties. *Journal of Experimental Psychology: Human Perception and*  
1132 *Performance*, 46(5), 458–473.

1133 Wei, X. X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding  
1134 can explain “anti-Bayesian” percepts. *Nature Neuroscience*, 18(10), 1509–1517.

1135 Wei, X. X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and  
1136 discriminability. *Proceedings of the National Academy of Sciences of the United States of*  
1137 *America*, 114(38), 10244–10249.

1138 Wei, Z., Wang, X.-J., & Wang, D.-H. (2012). From distributed resources to limited slots in  
1139 multiple-item working memory: a spiking network model with normalization. *The Journal*  
1140 *of Neuroscience*, 32(33), 11228–11240.

1141 Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working  
1142 memory. *Nature*, [453\(7192\), 233–235. https://doi.org/10.1038/nature06860](https://doi.org/10.1038/nature06860)  
1143 Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory.  
1144 *Psychological Science*, 20(4), 423–428.  
1145 Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory.  
1146 *Psychological Science*, 22(11), 1434–1441.

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159 Appendix: Supplementary Information for

1160

1161 *An adaptive perspective on visual working memory distortions*

1162

1163 Chaipat Chunharas, Rosanne L. Rademaker, Timothy F. Brady, & John T. Serences

1164

1165 Correspondence: Chaipat Chunharas

1166 [chaipat.c@chula.ac.th](mailto:chaipat.c@chula.ac.th)

1167

1168

1169 **This PDF file includes:**

1170

1171 Appendix Figures 1–8

1172

1173 **Other supplementary materials for this manuscript include the following:**

1174

1175 Data, code

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

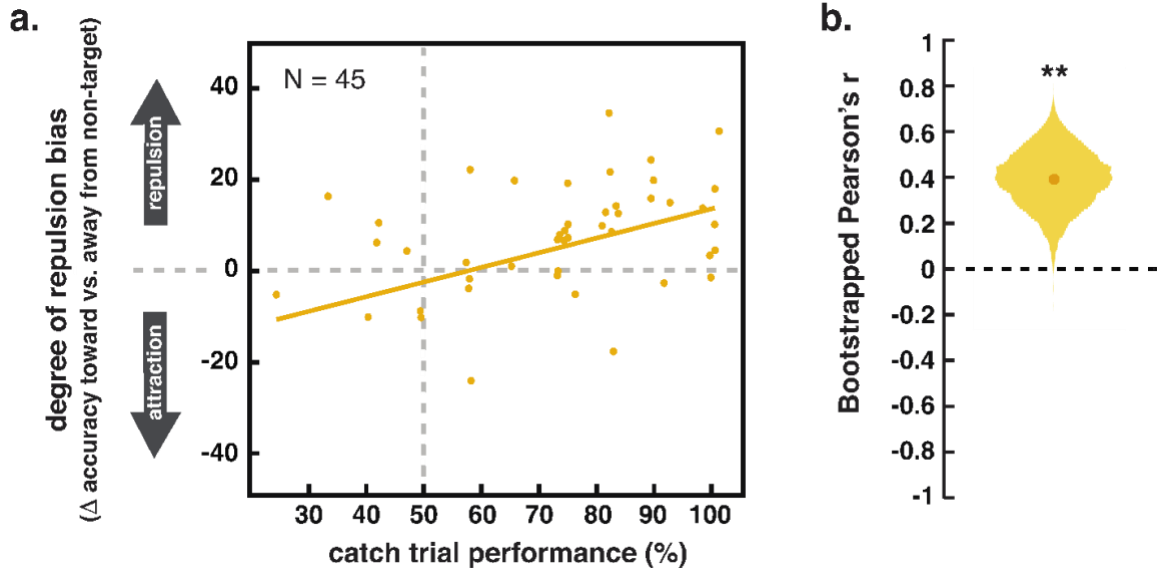
1187

1188

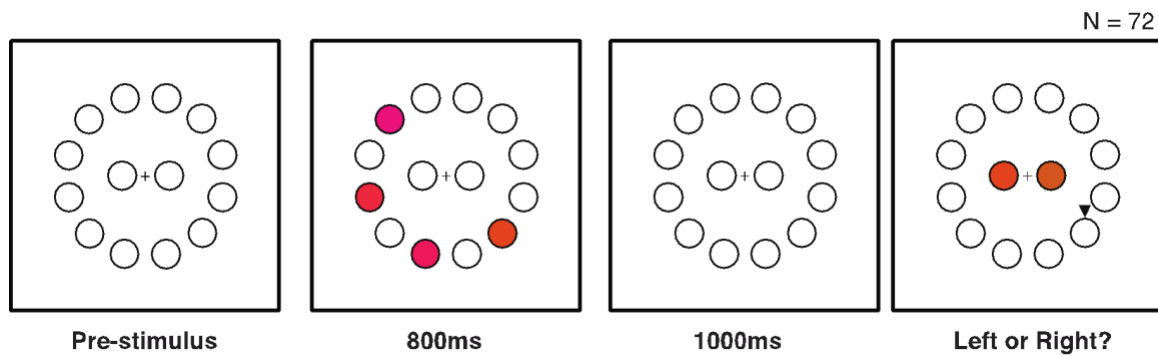
1189

1190



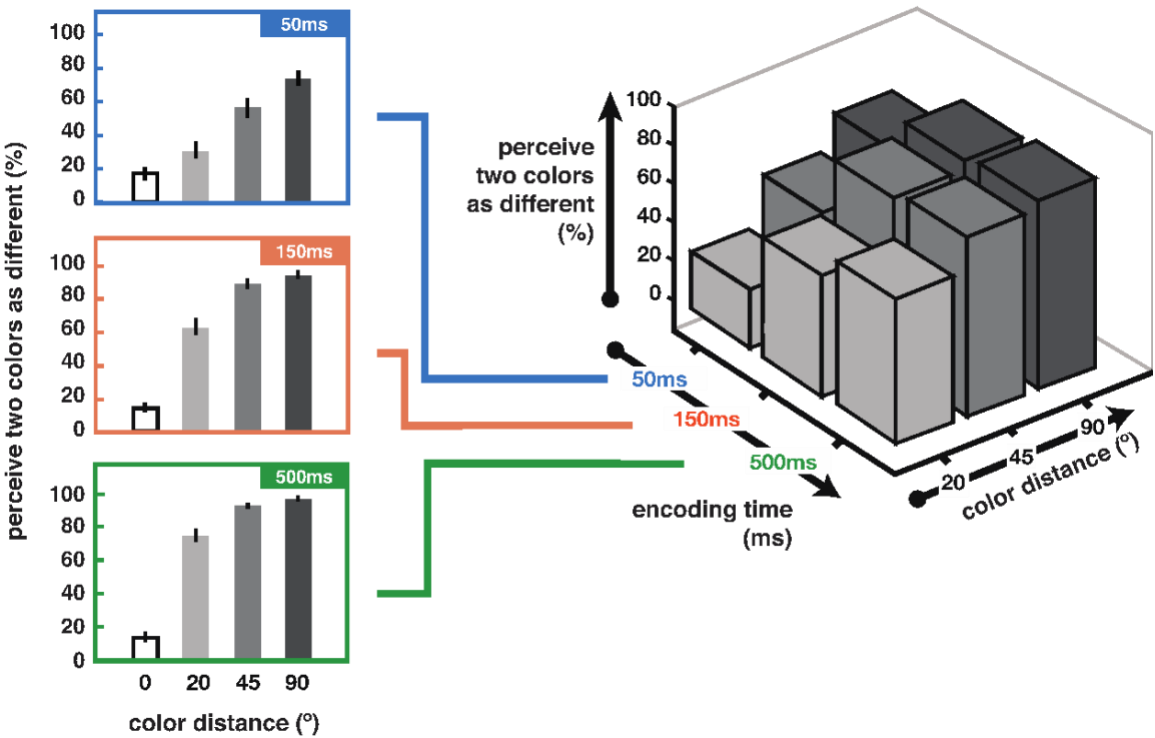


**Appendix Figure 1.** Results from a control experiment (N=45) replicating Experiment 2: In this experiment, only foils distorted by 6° relative to the correct color were used (towards and away from the non-target – similar to Exp. 1), while we also included 10% of catch trials (similar to Exp. 2). Participants were an entirely new and independent set of 45 naïve Amazon mechanical Turk workers. a.) The degree of repulsion bias (indexed as the difference in accuracy between trials with foils distorted toward, and trials with foils distorted away from the non-target color), plotted against people's general level of engagement with the memory task (indexed by performance on catch trials). Each dot represents a single subject. These data demonstrate stronger biases away from the non-target color in participants with higher levels of task engagement. b.) We bootstrapped the data in (a.) 5000 times: on each bootstrap we sampled 45 subjects with replacement, and re-calculated the correlation between repulsion bias and general task engagement. This gives a distribution of bootstrapped Pearson's r, which is depicted in the violin plot. The dot in the middle indicated the mean bootstrapped correlation ( $r = 0.39$ ). The double asterisks indicate a p-value of  $p < 0.01$ .



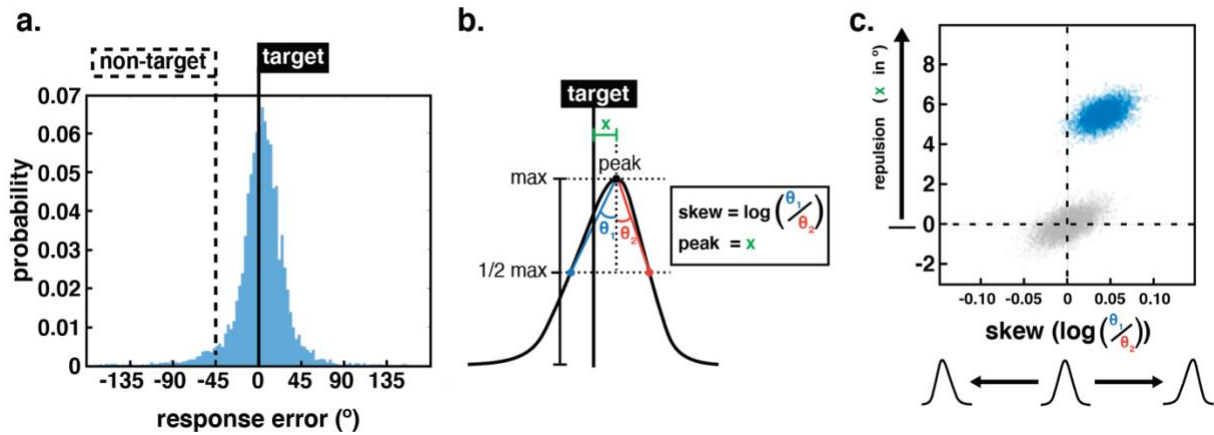
**Appendix Figure 2.** Task progression in Experiment 3: Participants had to remember a set of four colors (shown at randomly selected locations from a set of 12 possible locations, with at least one empty placeholder between items). The four colors were presented for 800ms, after which participants remembered them during a 1-second memory delay. Subsequently, participants saw a location cue (triangle) indicating which memory item to respond to, as well as two response options presented directly left and right of fixation. Participants chose between the correct (cued) color and a foil color.

1218

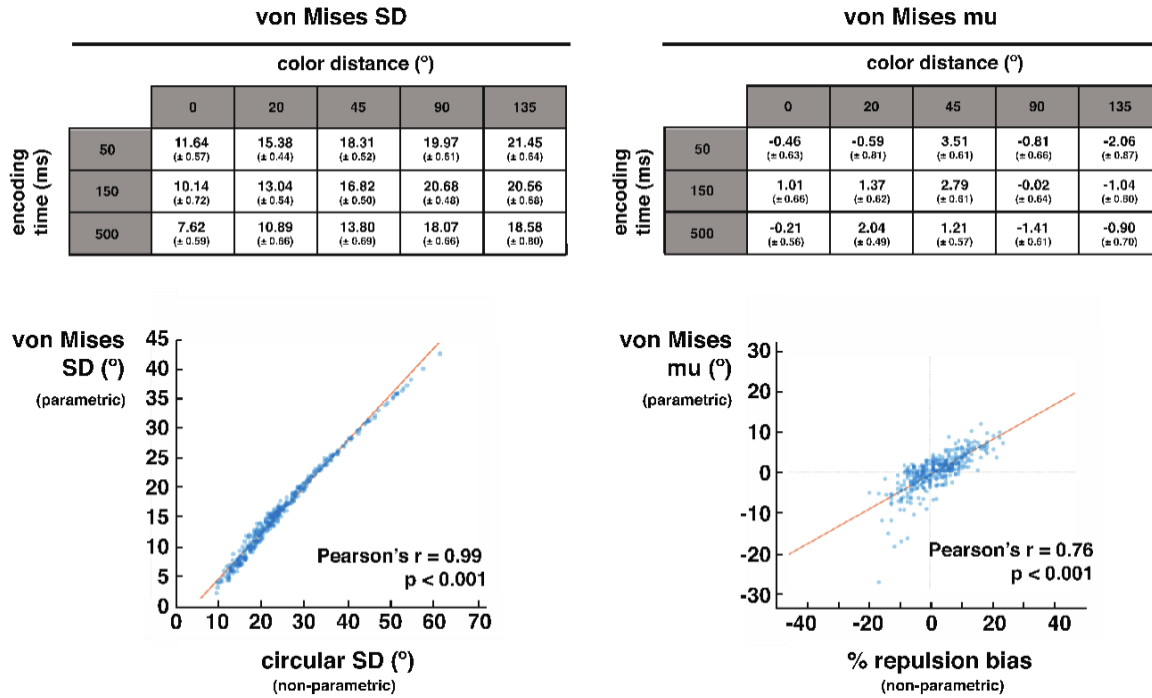


1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233

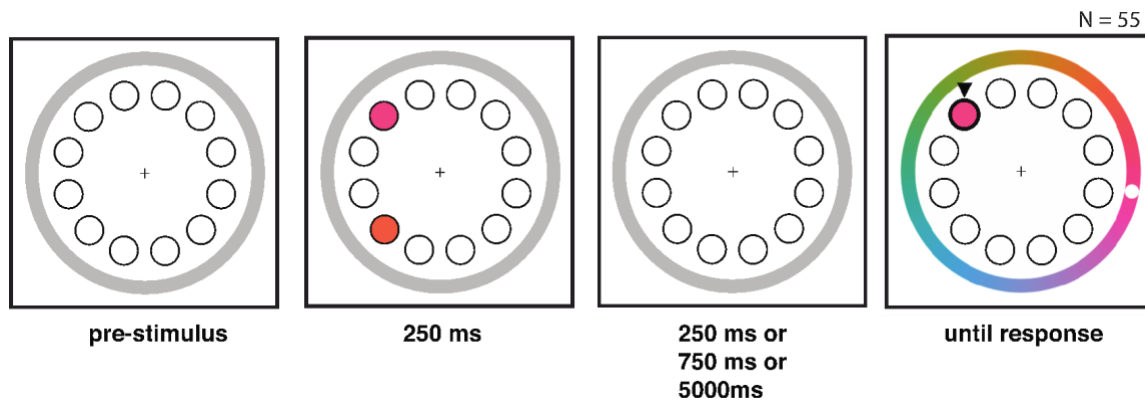
**Appendix Figure 3.** Results from a same-different color discrimination task as a control for Experiment 4. This control experiment probed whether two colors can or cannot be perceptually discriminated at various encoding times and color distances: Two colors that were either exactly the same (50% of trials) or differed by 20°, 45°, or 90° in CIE  $l^*a^*b^*$  color space (50% of trials) were simultaneously presented for either 50ms (blue), 150ms (orange), or 500ms (green). Participants on Amazon mechanical Turk (18 in total) reported whether the two colors were the same or different. Each participant completed 90 trials in total. The 3D bar plot (right) shows accuracy as a function of encoding time and color distance. Repeated-measures ANOVA's demonstrate both main effects of encoding time ( $F_{(2,34)}=36.7, p<0.001$ ), color distance ( $F_{(3,51)}=212.5, p<0.001$ ) and an interaction ( $F_{(6,102)}=9.32, p<0.001$ ). This means that participants could not tell two colors apart when they were presented very briefly and were very similar to one another (i.e., encoding time of 50ms and color distance of 20°). The inability of subjects to tell two very similar colors apart at very short encoding times explains why repulsion biases were not found in these extreme cases.



**Appendix Figure 4.** Asymmetry of Experiment 4 error distributions: a) An example error distribution from all 24 participants combined, in the condition showing the strongest repulsion bias (i.e. encoding time of 150ms and color distance of 45°). First, note how the peak of the error distribution is not aligned with the cued color, but instead is shifted away from the non-target color. Second, note how the shape of the distribution is asymmetrical, with the side away from the non-target being steeper. b) Due to the possible presence of non-target responses (i.e. where a subject mistakenly reports the color of the non-target instead of the target), we did not wish to measure skewness using circular skewness measures on the raw response distribution. Instead, we first derived a kernel density estimator (KDE). The peak of the distribution ( $x$ ) was defined as the degree of error with maximum probability. The skewness was defined by the log ratio between the angle toward ( $\theta_1$ ) vs. away ( $\theta_2$ ) from the non-target color at half maximum height of the KDE ( $\log(\theta_1/\theta_2)$ ). c) A scatter plot showing the relationship between skew and peak. Each dot represents skew and peak on one bootstrapping iteration (of 5000 total iterations) calculated by randomly resampling the data from 24 participants with replacement (data from the condition shown in a.). The horizontal zero line represents scenarios with no shift in the distribution peak, while the vertical zero line represents scenarios without any skew (thus, the 0,0 point represents a perfectly symmetrical distribution). We found both a systematic shift of the peak ( $p < 0.001$  from bootstrapping) as well as skew ( $p < 0.01$  from bootstrapping). Furthermore, the shape of the dot cloud shows that stronger repulsion is associated with a stronger skew ( $r = 0.45$ ;  $p < 0.001$ ). To test the validity of the metrics, we reanalyzed the same data with randomized signed errors and plotted in grey color. The randomized signed errors distribution centers at zero in both skew (x-axis) and bias (y-axis) suggesting that the significant bias and skew were not spurious.

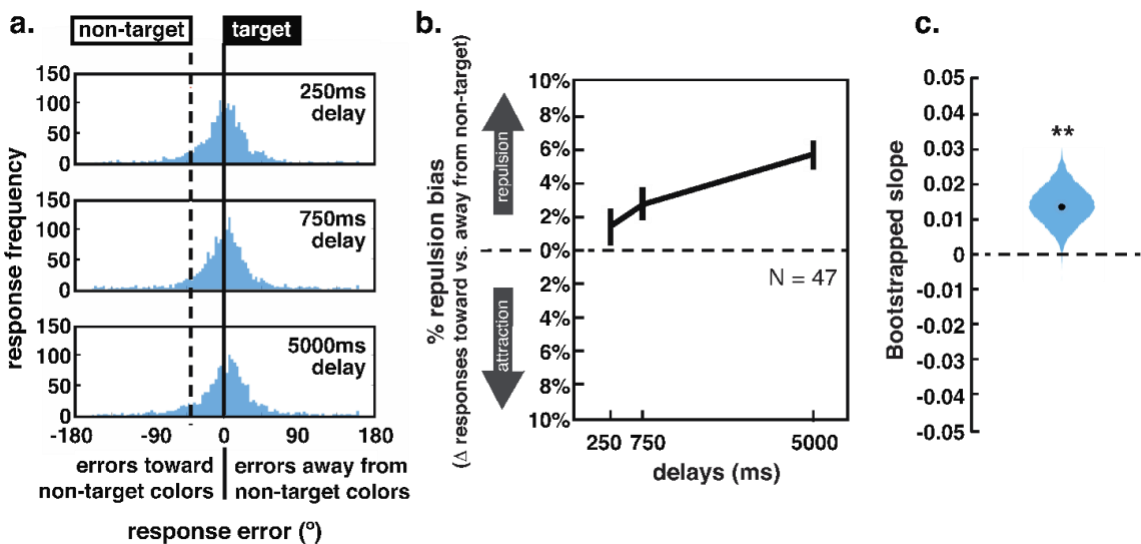


**Appendix Figure 5.** Parametric vs. non-parametric quantifications of memory precision and bias in Experiment 4. For this experiment we non-parametrically quantified memory precision as the circular standard deviation (with smaller standard deviations indicating higher precision) and we quantified biases as the difference in the percentage of responses that were toward vs. away from the non-target color (with a negative bias indicating attraction, and positive bias indicating repulsion). To validate these measures, we also parametrically fit the data using a von Mises distribution with two independent parameters to reflect memory precision (vmSD) and bias ( $\mu$ ). We found a high agreement between parametric vs. non-parametric measurements (Pearson's  $r = 0.99$  and  $0.76$ , for precision and bias, respectively; both  $p < 0.001$ ). The correspondence between these measures is shown in the scatter plots at the bottom of this figure. Furthermore, we repeated our statistical analyses with the parametric von Mises parameter estimates (tables in the top of this figure), showing significant differences in memory precision as a function of encoding time ( $F_{(2,46)}=13.7$ ,  $p < 0.001$ ), color distance ( $F_{(4,92)}=21.09$ ,  $p < 0.001$ ), and an interaction ( $F_{(8,184)}=3.76$ ,  $p < 0.001$ ). The repulsion bias is marginally impacted by encoding time ( $F_{(2,46)}=3.08$ ,  $p=0.056$ ), significantly impacted by color distance ( $F_{(4,92)}=9.54$ ,  $p < 0.001$ ), and there is a significant interaction ( $F_{(8,184)}=2.66$ ,  $p < 0.01$ ). Note that the mixture modelling assumes that the error distribution follows a symmetric circular distribution. However, the true error distributions were skewed which makes it less accurate in estimating the true biases and the memory strengths.



**Appendix Figure 6.** Task sequence in Experiment 5: Two color stimuli were presented for 250ms, and the color distance between the two items was fixed at 45°. The memory delay period was either 250ms, 750ms or 5000ms. After the delay, participants were cued to report one of the two memory items with an arrow cue, and they moved a white dot along a continuous color wheel to choose the color that matched their memory as closely as possible. For clarity, the grey circle and color wheel are shown wider here than they were presented during the actual experiment.

1289



1290

1291

1292

1293

1294

1295

1296

1297

1298

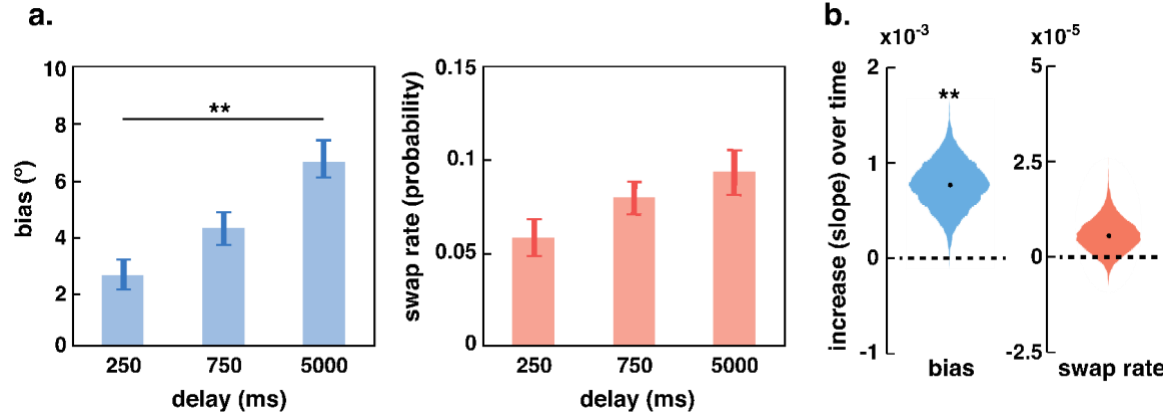
1299

1300

1301

1302

**Appendix Figure 7.** Results from a control experiment (N=47) replicating the finding from Experiment 5 that memory biases increase with longer delays. Here, we collected 36 trials per condition per subject (a total of 108 trials per subject) a.) Error distributions at each delay, revealing a high number of responses biased away from the non-target. b.) The quantified repulsion bias (i.e. percentage of responses away from the non-target color) shows that repulsion grew monotonically stronger as the delay duration increased (1.4%, 2.7%, and 5.6% for delays of 250ms, 750ms and 5000ms respectively;  $F_{(1,46)}=6.62$ ,  $p=0.013$ ). Error-bars represent  $\pm 1$  within-subject SEM. c) To assess the increase of repulsion bias with delay, one can fit a line through the three points in (b) and calculate the slope – a positive slope indicates repulsion bias growing as a function of delay duration. Shown here is a distribution plot of bootstrapped slopes (5000 iterations of resampling with replacement). The double asterisk indicates  $p<0.01$  confirming a statistically robust effect.



**Appendix Figure 8.** a.) Fitting a mixture model with swap errors to the data in Experiment 5 confirms that repulsion bias grows stronger with longer delay intervals (blue;  $F_{(1,54)}=10.2$ ;  $p=0.002$ ), confirming what we found with our non-parametric repulsion bias measure. The frequency of swap errors did not significantly change across time (red;  $F_{(1,54)}=1.87$ ;  $p=0.178$ ). b.) We computed slopes of bias and swap errors as a function of time – positive slopes indicating an increased repulsion or swap rate over time. We evaluated significance by resampling with replacement 10,000 times. Repulsion bias grew significantly stronger as the delay interval increased (blue), replicating our findings using a non-parametric bias measure. Swap errors did not increase significantly as the delay interval increased. These results suggest that the increase in repulsion bias that we found when using either parametric or non-parametric methods cannot be explained by a reduction in swap errors (if anything, swap errors increase with delay, numerically).