Attention, Perception, & Psychophysics (2021) 83:1652–1672
https://doi.org/10.3758/s13414-020-02236-3

Statistical learning as a reference point for memory distortions: Swap and shift errors

Paul S. Scotti[1] · Yoolim Hong[1] · Julie D. Golomb[1] · Andrew B. Leber[1]

Accepted: 23 December 2020 / Published online: 18 January 2021

Abstract

Humans use regularities in the environment to facilitate learning, often without awareness or intent. How might such regularities distort long-term memory? Here, participants studied and reported the colors of objects in a long-term memory paradigm, uninformed that certain colors were sampled more frequently overall. When participants misreported an object's color, these errors were often centered around the average studied color (i.e., "Rich" color), demonstrating swap errors in long-term memory due to imposed statistical regularities. We observed such swap errors regardless of memory load, explicit knowledge, or the distance in color space between the correct color of the tested object and the Rich color. An explicit guessing strategy where participants intentionally made swap errors when uncertain could not fully account for our results. We discuss other potential sources of observed swap errors such as false memory and implicit biased guessing. Although less robust than swap errors, evidence was also observed for subtle shift errors towards or away from the Rich color dependent on the color distance between the correct color and the Rich color. Together, these findings of swap and shift errors provide converging evidence for memory distortion mechanisms induced by a reference point, bridging a gap in the literature between how attention to regularities similarly influences visual working memory and visual long-term memory.

Keywords Memory: Long-term memory · Attention in learning

Introduction

Human memory is easily distorted and prone to false memories (Bartlett, 1932; Brainerd & Reyna, 2008; Loftus, 2003; Wixted, Mickes, & Fisher, 2018). Memory distortions can occur because information from multiple memory sources was incorrectly combined. This can explain how leading questions can contaminate eyewitness testimony and induce strong false memories for recent events (Loftus & Hoffman, 1989). More recently, it has been proposed that memory errors can prove adaptive for the memory system by maximizing overall task performance at the expense of specific failures (e.g., Carpenter & Schacter, 2017; Guerin, Robbins, Gilmore, & Schacter, 2012; Newman & Lindsay, 2009; Schacter, Guerin, & St. Jacques, 2011; Yoo, Klyszejko, Curtis, & Ma, 2018). One source of information that the memory system takes advantage of comes from regularities in our surroundings. Statistical learning refers to acquired knowledge of environmental patterns, which may be expressed via changes in behavior (see Perruchet & Pacton, 2006, for review). These patterns, or regularities, are automatically and implicitly incorporated by the memory system during unsupervised learning, such as how visual attention is implicitly guided by spatial probability (Geng & Behrmann, 2002; Jiang, Swallow, Rosenbaum, & Herzig, 2013) or how infants learn word boundaries based on the statistical relationships between neighboring speech sounds (Saffran, Aslin, & Newport, 1996). Prior knowledge that an object category (e.g., apples) generally contains similar features (e.g., size, color, texture) is based on accumulated interactions with this kind of object. We exploit knowledge of these regularities when retrieving uncertain information about a single memory item, leading to systematic memory distortions (e.g., biased reports of an apple's size towards the average; Hemmer & Steyvers, 2009).

Research into the different types of memory distortions has largely focused on visual working memory rather than visual long-term memory, specifically in relation to *swap errors* and *shift errors* (e.g., Bae & Luck, 2017; Bays, Catalao, & Husain, 2009; Brady & Alvarez, 2011; Golomb, 2015; Golomb,

✉ Paul S. Scotti
scottibrain@gmail.com

[1] Department of Psychology, The Ohio State University, 225 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210, USA

L'Heureux, & Kanwisher, 2014; Huang & Sekuler, 2010). *Swaps* and *shifts* reflect distinct types of feature-binding errors found in memory and attention paradigms. Swap errors refer to the mistaken report of a non-target feature. For example, you might report a target item as being colored blue when presented with a red target and a blue distractor. *Swap errors* can reflect "misbindings" or "misassociations" between items (e.g., Bays et al., 2009; Bays, Wu, & Husain, 2011), though the term is also used more generally to reflect a type of error where the participant misreports the feature(s) of a different object (e.g., Dowd & Golomb, 2019). *Shift errors* refer to more subtle errors where the report is biased towards (attraction) or away from (repulsion) a distractor feature. For example, you might report the target as magenta when presented with a red target and a blue distractor.

Feature-binding errors have classically been described as failures of the attentional system to "glue" object features together (Treisman, 1988, 1998; Treisman & Gelade, 1980). Due to capacity limitations of the visual working memory system, attentional resources may be misallocated and result in feature-binding errors (Dowd & Golomb, 2019; Zokaei, Heider, & Husain, 2014). Feature-binding errors can also be observed in visual long-term memory, where memory items can be forgotten at different rates (Brady, Konkle, Alvarez, & Oliva, 2013) and are susceptible to recombination (i.e., swap errors; Lew, Pashler, & Vul, 2016; Utochkin & Brady, 2019). Notably, feature-binding errors can occur in long-term memory even when the attentional system is not near capacity during encoding or retrieval (slow presentation time for encoding objects one at a time, and subsequent unlimited time to respond to each memory item during testing). This suggests that while feature-binding errors observed in visual working memory may represent constraints or errors of the attentional system, feature-binding errors in visual long-term memory may represent constraints or errors of the memory system (although it is unclear when these errors arise during encoding, storage, or retrieval).

We reasoned that such long-term memory feature-binding errors could be influenced by a signal particularly important to the long-term memory system: statistical regularities. Given distinct mechanisms that may underlie feature-binding errors in visual working memory and visual long-term memory, it remains an open question how statistical regularities might induce systematic memory distortions in visual long-term memory. Previous studies showing swap errors in visual long-term memory (e.g., Lew et al., 2016; Utochkin & Brady, 2019) have focused on mistaken reports of a non-target feature, but no one has examined how statistical learning might act as the reference point for swap errors, and if so, what might be the underlying source of swap errors, including explanations such as implicit biased guessing, explicit biased guessing, and false memory (where false memory would indicate swap errors due to true mis-binding of the target feature). Regarding shift errors in visual long-term memory, previous studies have focused on attraction bias (e.g., Brady, Schacter, & Alvarez, 2018; Hemmer & Steyvers, 2009; Huttenlocher, Hedges, & Vevea, 2000), but no one has examined how the direction of shift errors (repulsion or attraction) in long-term memory might be dependent on the similarity between the memory item and the reference point in feature space.

We sought to address these questions by simultaneously probing the existence of swap and shift errors in a long-term memory paradigm where participants memorized and later reported the colors of real-world objects. Unknown to participants, we manipulated the regularity of shared visual information among learned objects. Specifically, object colors were more likely to be sampled from a certain region in color space. During test, participants were instructed to recreate the original color of every object and report how confident they were in their color selection. We hypothesized that imposed statistical regularities, namely the average studied color (hereafter referred to as the "Rich" color), might act as a reference point for the memory system to systematically distort subsequent color reports.

We used probabilistic mixture modeling to characterize memory responses according to one of four underlying distributions: a target distribution (responses around the correct color), a swap distribution (responses around the Rich color), a swap-comparison distribution (responses around the color 180° away from the Rich color), and a random guessing distribution (uniform responding across all colors). The target distribution included a flexible precision parameter as well as a flexible mean parameter to assess shift errors.

We considered a few possible outcomes. First, regularities may not bias long-term memory. If this were the case, we would expect participants to report a mix of correct responses and random guesses. The correct responses would be fit by the target distribution (with individual variations in precision), with a mean centered around 0° of error (no shift), and the incorrect responses would be uniformly distributed and therefore well-fit by the random guessing distribution. In this no-bias outcome, the proportion of trials fit by the swap and the swap-comparison distributions would both be near zero and not quantitatively different from each other.

A second possibility is that regularities induce swap errors in long-term memory, causing participants to report the Rich color instead of the correct color for some objects. This could come in the form of a strategic guess, where the participant does not remember the object color but has realized that most objects are a certain color, hence responding near the Rich color to maximize accuracy (we refer to this as "explicit biased guessing" from now on). Another interpretation of swap errors might be that participants guessed a color near the Rich color due to some kind of implicit, adaptive behavior (i.e., "implicit biased guessing," optimally guessing near the Rich color despite not explicitly knowing about statistical regularities). Yet another interpretation is that swap errors could reflect the participant truly misremembering the wrong color-pair association for the tested object (i.e., a "false memory" due to a feature-binding error). In all three of these cases, we would expect participants to report a mix of correct responses, swapped incorrect responses, and random guesses, though our confidence report analyses might reveal important differences in the relative proportions of these kinds of responses.

A third possibility (non-exclusive to the second) is that regularities induce shift errors in long-term memory. If this were the case, we would expect participants to report a mix of correct responses and random guesses, but the correct responses would actually be biased in color space, shifted systematically towards or away from the Rich color. This would result in the flexible mean parameter for the target distribution being shifted from zero, either away from (repulsion) or towards (attraction) the Rich color. Shift errors are expected to be small, likely biasing reports no greater than 10° in color space from the correct response (Bae & Luck, 2017; Golomb, 2015; Golomb et al., 2014), meaning that it is unlikely that shift errors would be falsely attributed to the swap or swap-comparison distributions. It is also possible that both swap and shift errors are observed, resulting in a shifted target distribution and a larger proportion of trials accounted for by the swap distribution compared to the swap-comparison distribution.

Finally, an accompaniment to the shift hypothesis is that the magnitude and direction of shift errors may depend on the distance in color space between the correct color and the Rich color. This prediction is motivated by the relational representation model described in the visual working memory literature (Bae & Luck, 2017; see also Golomb, 2015). The relational representation model explains that, given a short distance in feature space between a memory item and a reference point, the memory item is often remembered as repulsed, or farther away, from the reference point. In contrast, given a long distance in feature space between a memory item and a reference point, the memory item is often remembered as attracted to, or towards, the reference point. These relational mechanisms observed in working memory have never been reported in long-term memory but may be analogously explained by the hippocampal computational processes of pattern separation and pattern completion. Pattern separation is similar to repulsion and refers to overlapping memory representations becoming more distinct to reduce interference during retrieval (e.g., Aimone, Deng, & Gage, 2011; Yassa & Stark, 2011). Meanwhile, pattern completion is similar to attraction and refers to the generalization of new memories based on an internal model of the learned structure of our environment (e.g., S. Leutgeb & Leutgeb, 2007; Yassa & Stark, 2011), where multiple memory signals may be generalized into a single representation less prone to degradation. We thus also separately modeled memory responses depending on the distance in color space between each memory item's color and the Rich color to test this relational representation model for long-term memory.

Experiment 1: Regularities induce swaps in long-term memory

In Experiment 1, participants studied and recalled the colors of 40 unique real-world objects per each of nine blocks, where 30% of objects shared the exact same color (with the remaining objects randomly distributed across the remaining 359 colors in our circular color space). At the end of each block, participants were instructed to recreate the original color of every object in the preceding study block. Participants were not informed that 30% of all objects (12 per block) shared the exact same color (this Rich color was randomly determined per subject and constant for all blocks), while the rest of the objects were randomly sampled from the remainder of the color wheel. We used probabilistic mixture modeling to characterize memory responses according to four distributions (target, random guess, swap, or swap-comparison).

Method

Open practices

The rationale, method, and the analyses for every experiment were preregistered at the Open Science Framework (OSF) (https://osf.io/n7c5e/). Analyses not mentioned in the preregistration are declared as exploratory. All preregistered analyses are reported in either the main text or the supplemental text. For ease of exposition, Experiment 1 is reported first despite being the last experiment conducted (the chronological order was Experiment 2, Experiment 3, and then Experiment 1). Any deviations made in regard to analysis decisions are explicitly mentioned when relevant.

Participants

Experiment 1 included 44 participants (21 male, 23 female; M_{age} = 34.98 years, SD = 9.91). An a priori simulation-based power analysis estimated that we would need 44 participants to credibly detect possible shift errors with our model-based analysis described below. (This power analysis was based on simulations of the data from Experiments 2 and 3, conducted chronologically before Experiment 1; for the full details of the simulation, see preregistration: https://osf.io/n7c5e/, section 7 of Preregistration_Details_ContLTM_EXPT3.docx). Fewer subjects were estimated to be needed to detect swap errors. All participants were recruited through Amazon Mechanical Turk (MTurk) and were paid US$9 (plus bonus based on performance) for the experiment, which lasted roughly 1–1.5 h. All participants lived in the USA, held an MTurk approval rating of ≥ 98%, and successfully completed over 750 MTurk tasks prior to this experiment. All participants reported normal or corrected-to-normal vision, were naive to the purpose of the experiment, and provided informed consent in accordance with The Ohio State University institutional review board.

Ten additional participants were excluded based on preregistered exclusion criteria (https://osf.io/n7c5e/). Exclusion criteria were based on pilot data and meant to reflect the minimal threshold of performance expected from an attentive participant. Individuals were excluded if their target proportion was less than .10 or if their average standard deviation across the target and biased guessing distributions (weighted by respective proportion fits) exceeded 70°. These fits were based on a mixture model that included memory responses to objects originally sampled > 45° from the Rich color (i.e., excluding trials where the correct color was near the Rich color).

Materials and procedure

Experiment 1 was conducted online using MTurk, meaning that monitors could vary in size and viewing distance. Therefore, we report stimulus sizes in pixels (px) and not degrees of visual angle. Figure 1a illustrates an example trial sequence for study and test blocks. Each study block consisted of 40 unique real-world objects. A 250 x 250 px object was presented in the center of a gray, square background (600 x 600 px) for 1 s, followed by a blank 1-s interval. A central 15 x 15 px black fixation cross appeared on the center of the screen between image presentations. Participants were instructed to memorize the color associated with every object, knowing that they would be asked to recreate the original colors for all objects at the end of each study block. A practice study and test block of five objects (sampled evenly across the color wheel) familiarized participants with the procedure. Stimulus presentation was facilitated by a combination of HTML, CSS, and JavaScript.

Object stimuli were acquired from two image sets. We used images from Brady, Konkle, Alvarez, and Oliva (2008) and from the Bank of Standardized stimuli (BOSS; Brodeur, Dionne-dostie, Montreuil, & Lepage, 2010). Our resulting image set used across all three experiments contained 365 objects. Posterization was applied to each image such that pixel values could only be white, black, or a single color of interest (one of 360 RGB color values drawn from a one-dimensional selection of CIE Lab color space provided by MemToolbox; Suchow, Brady, Fougnie, & Alvarez, 2013). Specifically, each image was first converted to grayscale, with

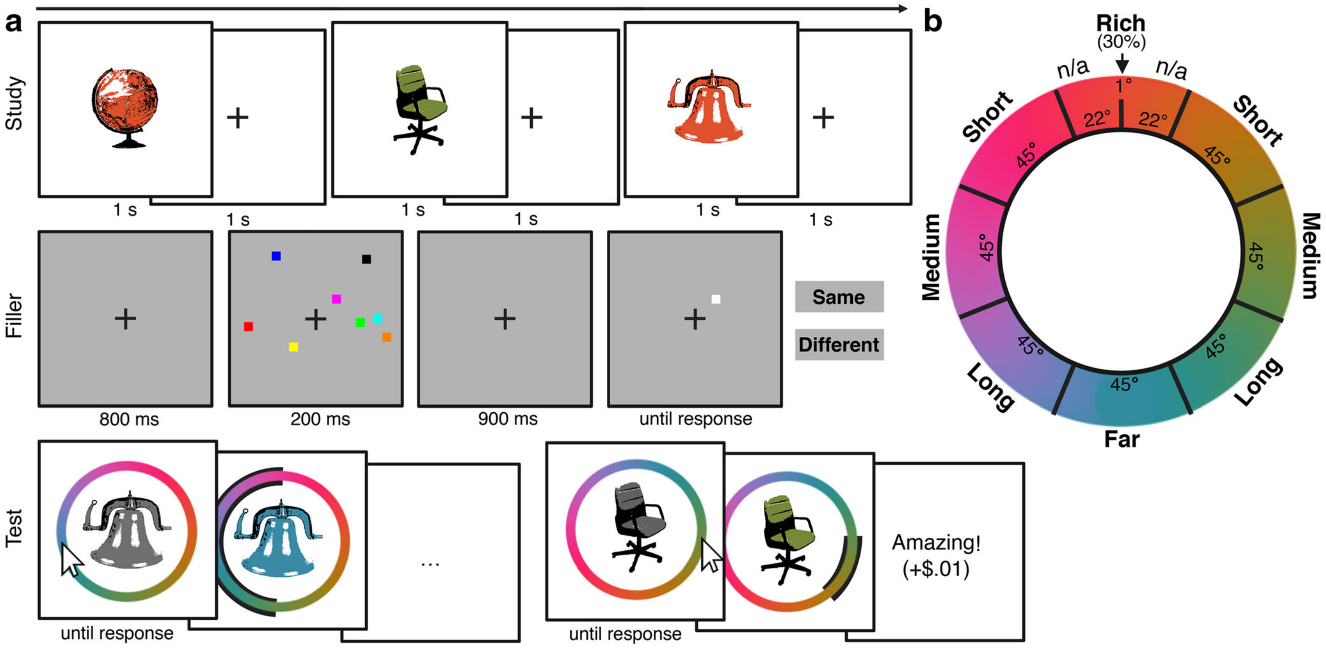

Fig. 1 (*A*) Example trial sequences for study and test blocks. Participants were instructed to memorize the color associated with all 40 objects presented during each study block. Following a color-change detection filler task, participants recreated the original color of all objects from the previous study block. Objects were presented in grayscale until mouse movement, where the object's color dynamically adjusted to match the color closest to the mouse pointer. After clicking to confirm their best guess, participants highlighted the smallest region of colors that they believed contained the original color (they were also instructed to highlight the entire color wheel if they were completely guessing). Every test trial ended with general feedback and bonus information. (*B*) Proportion of colors sampled across objects and the color-distance segments (Short, Medium, and Long) used to bin memory errors (separate models per bin). Thirty percent of all objects shared the same (Rich) color, randomly determined per subject, and all other objects were equally sampled from the rest of the color wheel and then binned accordingly

luminance values ranging from 0 to 255. Pixels with a luminance between 0 and 85 were colored white and pixels with a luminance between 170 and 255 were colored black. All other pixels were assigned to the specified object color. Objects were individually selected from the above two image sets to ensure the posterization process did not render objects unrecognizable or without a reasonable number of colored pixels (subjectively determined by PS).

Following each study block, two trials of a filler task (color-change detection task) were presented (same procedure as Brady, Konkle, Gill, Oliva, & Alvarez, 2013). Our rationale was that a change detection task involving color would occupy visual working memory (e.g., Allon & Luria, 2017; Luck & Vogel, 1997) and would help to ensure that we were primarily testing passively stored, long-term memory representations in our main experiment. A gray background was presented for 800 ms, followed by the presentation of eight colored squares for 200 ms. After a blank 900-ms retention interval, a single colored square appeared at one of the previous item locations and participants were to indicate whether the color was the same as the square that appeared at the same location during the initial display. Colored squares were 60 x 60 px each, and buttons indicating "same" or "different" were displayed immediately below the gray background after the retention interval.

Following the filler task, a memory test block was presented. Test blocks consisted of the same 40 objects as the preceding study block. On each test trial, one object – selected randomly without replacement from the study block objects – was presented in grayscale (color was replaced with RGB [128, 128, 128]), and a color wheel was presented around the object. The color wheel was randomly flipped on half of the trials and randomly rotated on each trial. Participants were tasked with selecting the original color of the object by clicking on the color wheel. As the mouse moved around the color wheel, the initially grayscale object dynamically changed to the color indicated by the mouse pointer's position on the color wheel. Following a mouse click to confirm their selected color, participants made a confidence range report highlighting the smallest portion of the color wheel that they believed contained the true color (see Chen, Leber, & Golomb, 2019). Participants were instructed to highlight the entirety of the color wheel if they were completely guessing. Highlighting involved two clicks to define the start and end points of a black, highlighted region. This confidence range report was our proxy for subjective memory strength, where a larger highlighted region indicated less confident memory retrieval. There was no time limit imposed during the test block.

Following confidence range reports, general feedback was presented for 1 s. If memory error was ≤ 15°, participants were rewarded 1 cent ("Amazing! +$.01" displayed on screen), and if memory error was > 15° and ≤ 30°, participants were rewarded half a cent ("Good +$.005" displayed on screen). Otherwise, "..." was presented on the screen and no bonus was awarded. Importantly, no color information was displayed on the screen during feedback; we used this general feedback, rather than feedback showing the correct color of the original object, to reduce the chance that memory distortions emerged from incentivized feedback more so than stimulus regularities. Feedback was based purely on the initial color report; there was neither incentive nor feedback provided in regard to reporting confidence ranges.

There were nine study and test blocks total, consisting of 360 total studied/tested objects. Importantly, the colors of studied objects were not equally sampled from the color wheel. Specifically, 30% of objects across the experiment (12 trials per block) shared the exact same (Rich) color, randomly determined for each participant. The remaining 70% of objects were randomly sampled from the remainder of the color wheel. Participants were uninformed of our color-sampling manipulation.

A post-experiment survey (following the final test block) helped to assess whether participants were aware of imposed statistical regularities. The questions included (in order):

- "What strategy (if any) did you use when selecting the colors of objects?" Open-ended, typed response.
- "Do you think object colors were sampled randomly from the color wheel or do you think objects often shared approximately the same color?" Two-alternative forced choice between "each color chosen at random" and "objects often shared the same color."
- "Thirty percent of objects shared the same color. Please select the color you think matches our preferential color sampling." A color wheel was presented and participants were instructed to click the most commonly studied color.
- "How confident are you that the color you selected is close to the actual experimental manipulation?" 6-point scale with 1 being "least confident" and 6 being "most confident."

Analyses

The memory response distributions were fit using Markov chain Monte Carlo (MCMC), as implemented in MemToolbox (Suchow et al., 2013). We used a modified swap model (Bays et al., 2009; Golomb et al., 2014) to account for various sources of error. Each response was first converted into an error measurement (i.e., the difference between the reported and correct color values). Errors in which participants reported a color in the direction towards the Rich color were signed positive, and errors in which participants reported a color in the direction away from the Rich color (i.e., closer to the color 180° away from the Rich color) were signed negative. In this way, we could observe a mean shift in the target distribution where responses were either towards or away from the Rich color, as well as swap errors specific to the Rich color. The model included four distributions: a target distribution, a swap distribution, a swap-comparison distribution, and a random guessing distribution. The target distribution was a circular Gaussian (von Mises) probability density function centered on the original, correct color (with flexible mean and standard deviation). The swap distribution was a circular Gaussian (von Mises) probability density function centered on the Rich color, and the swap-comparison distribution was a circular Gaussian (von Mises) probability density function centered on the color 180° away from the Rich color. The swap and swap-comparison distributions shared a flexible standard deviation parameter. Note that a recent framework by Schurgin, Wixted, and Brady (2018) criticizing mixture models is not problematic for our purposes because we are not relying on any theoretical distinctions between guess rate and precision for our analyses. The difference between our modified swap model and the original swap model (Bays et al., 2009; Golomb et al., 2014) was that we gave the target distribution a flexible mean, we included a swap-comparison distribution, and we used separate precision parameters for the target distribution and the swap/swap-comparison distributions. The probability distribution can be expressed as:

$$p(\theta) = (1{-}S{-}C{-}\gamma)\phi_{\mu,\kappa_1} + S\phi_{d,\kappa_2} + C\phi_{d+180,\kappa_2} + \gamma\left(\frac{1}{2\pi}\right),$$

where θ is the difference between the reported and correct color values, γ is the proportion of trials on which the participant responded at random, ϕ is a von Mises distribution with mean μ, d, or $180 + d$, and concentration κ_1 or κ_2 (standard deviation $= \sqrt{1/\kappa}$), S is the proportion of "swap" trials on which the participant responded around the Rich color (von Mises distribution with mean d, the distance from the original color to the Rich color, and concentration κ_2), and C is the proportion of "swap-comparison" trials on which the participant responded around the color 180° away from the Rich color.

We separately modeled memory response distributions by bin depending on the distance in color space between the target color and the Rich color (Short, Medium, and Long; see Fig. 1b). This is because the relational representation model predicts repulsion for closer distances and attraction for farther distances. We also wanted to observe whether swap errors might be more or less likely depending on color distance.

In line with our preregistration, for this particular experiment we created an aggregated "super-subject" that contained trials across all participants, and the model was fit to this dataset for each color-distance segment. We preregistered a "super-subject" model, as opposed to fitting a model for every participant, because we anticipated such an approach would be necessary due to insufficient power to individually model memory errors for each color-distance segment. (Note that individual subject modeling procedures are employed in later experiments.) For each color-distance segment, we sampled three parallel chains across as many iterations as needed to reach convergence, according to the method of Gelman and Rubin (1992). We collected 15,000 post-convergence samples and used the posterior distributions to compute the maximum a posteriori estimates of the parameters μ, κ_1, κ_2, γ, S, and C. We also computed 95% highest posterior density intervals (HDIs), which indicate that the true parameter value has a 95% probability of lying within this interval (Kruschke, 2011). Swap errors were considered credible if the 95% HDIs for the swap and swap-comparison distributions did not overlap, and shift errors were considered credible if the 95% HDI for μ did not contain zero. Shift errors were characterized as reporting a color close to the correct color but slightly shifted either towards (attraction) or away from (repulsion) the Rich color ($\mu \neq 0$), and swap errors were characterized as misreporting the original color as the Rich color more frequently than misreporting it as the color 180° away from the Rich color ($S > C$). Additional exploratory analyses designed to explore the relationship between memory distortions and confidence range reports, as well as analyses on the post-experiment survey, are reported in the *Collapsed experiments* section.

Results and discussion

Swap errors

We examined whether statistical regularities, namely predisposing a certain color to occur most often among studied objects, might distort subsequent long-term memory reports. Figure 2a depicts histograms of memory errors, binned into trials where the original object color was sampled from a Short (23–67°), Medium (68–112°), or Long (113–157°) distance in color space from the Rich color. The bump in memory errors centered around the Rich color (swap errors) is visually obvious across all color-distance segments. Figure 2b depicts the model parameter estimates for γ, S, and C for each color-distance segment. The proportion of trials attributed to the swap distribution (S) was credibly larger than the proportion of trials attributed to the swap-comparison distribution (C) for all color-distance segments, as indicated by non-overlapping HDIs between the swap and swap-comparison parameters (0% of posterior probability densities overlapped for each comparison (S_{short} = .163, 95% HDI: [.136 .201], C_{short} = .007 [.000 .021]; S_{medium} = .148 [.122 .173], C_{medium} = .024 [.005 .040]; S_{long} = .106 [.088 .131], C_{long} = .005 [.000 .023])). This indicates that participants made swap errors where they

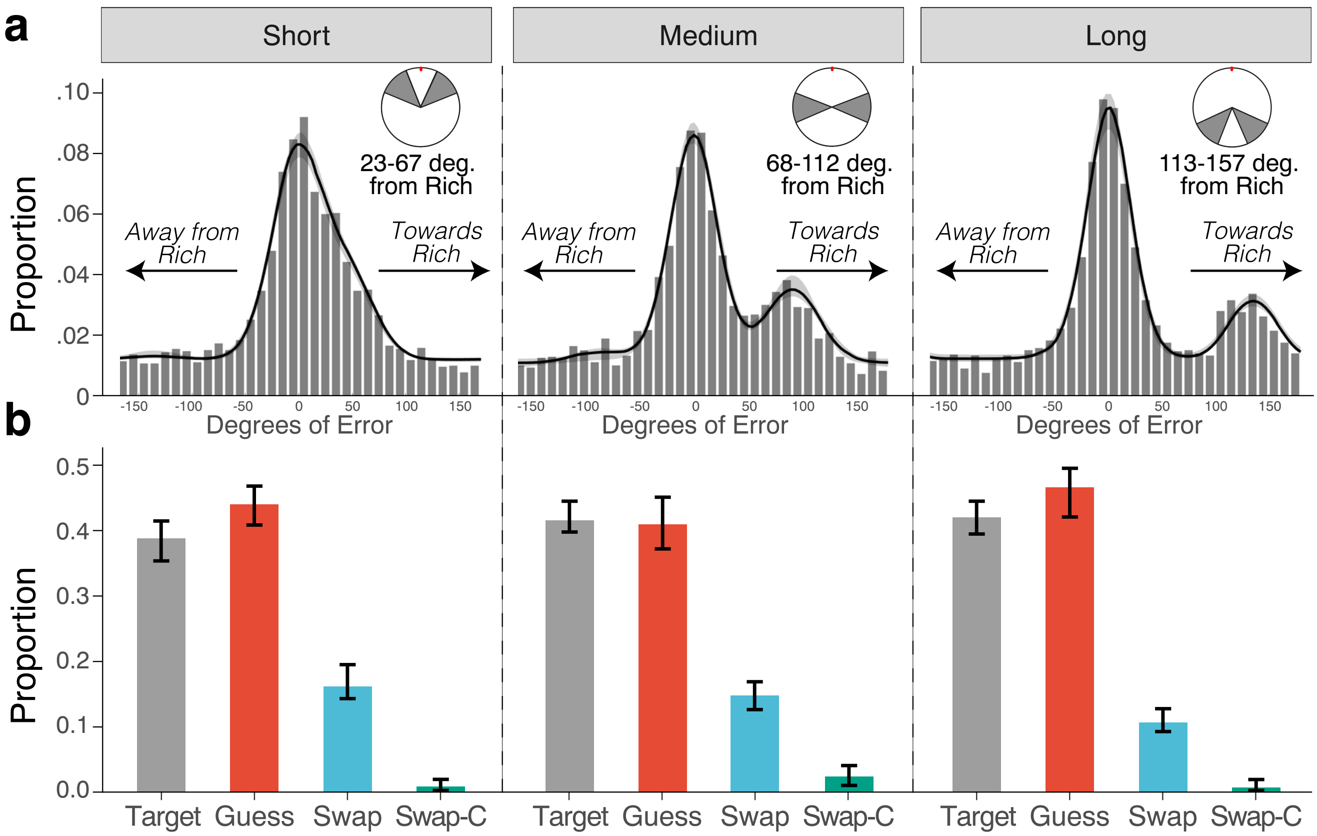

Fig. 2 Response histograms and model fits according to color-distance segments. "Short" trials contained objects originally sampled 23–67° away in color space from the Rich color, "Medium" trials were 68 to 112° away, and "Long" trials were 113 to 157° away. Circle plots represent the analyzed color-distance segment, with red dots depicting the location of the aligned, Rich color. (**A**) Response histograms where error corresponds to the difference between the reported color and the correct, originally presented color. Errors were signed positive if the memory response was in the direction towards the Rich color and signed negative otherwise. Gray shading around lines of best fit reflects 95% credible intervals. A bump in the histogram is easily visible where participants falsely reported a color centered around the Rich color. (**B**) Bar plots depict maximum a posteriori mixture proportions. Error bars represent 95% highest posterior density intervals. Models were fit separately for each color-distance segment, collapsed across subjects

reported values around the Rich color instead of the correct color.

Shift errors

We explored the existence of shift errors as indicated by credibly positive or negative μ parameter estimates for the target distribution (see Fig. 3). No credible shift errors were observed for the Short color-distance segment (μ_{short} = -2.32 [-4.28 0.60]), although it was close to passing the credible cutoff because 92.8% of the posterior probability density was below the threshold of 0° of error, towards repulsion (95% was needed for us to consider the shift as credible). There was a credible repulsion bias observed for the Medium color-distance segment (μ_{medium} = -1.38 [-3.46 -0.03], 97.2% of the posterior probability density below 0° of error, towards repulsion). No significant shift errors were observed for the Long color-distance segment (μ_{long} = 0.42 [-1.06 2.19], 81.7% of the posterior probability density above the critical threshold of 0° of error, towards attraction). Although not all of these effects were credible, the direction of shift errors was in line with the relational representation model (Bae & Luck, 2017; see also Golomb, 2015), with repulsion away from the Rich color for the Short and Medium color-distance segments and attraction towards the Rich color for the Long color-distance segment. These findings converge to form a weak but intriguing pattern, tentatively suggesting that reference-based shift errors may arise from statistical regularities in long-term memory, but more conclusive evidence is needed to solidify this claim. The more robust finding from Experiment 1 was that swap errors were observed due to imposed statistical regularities.

Other parameters

Exploratory analyses of the target proportion, random guessing proportion, and the standard deviation of the target and swap/swap-comparison distributions suggested that none of these parameters differed as a factor of color distance (overlapping 95% HDIs). A full list of model parameter estimates

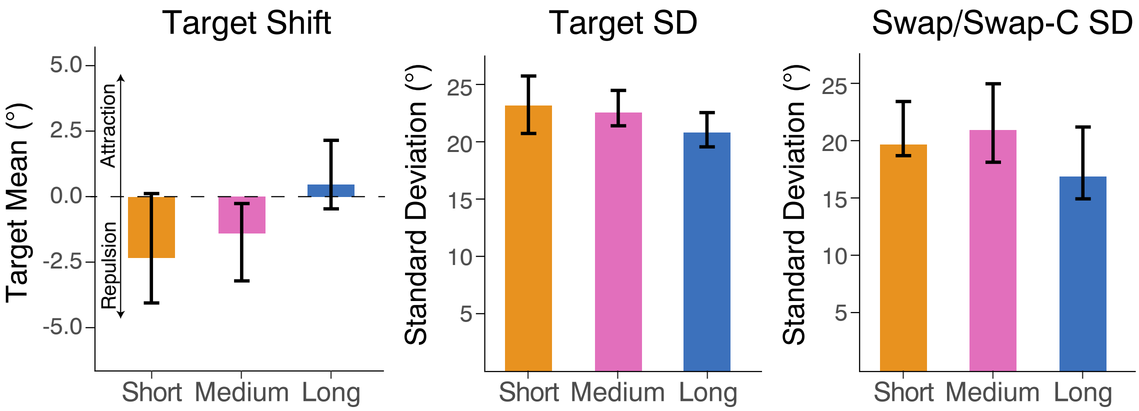

Fig. 3 Bar plots depict maximum a posteriori estimates for the mean and standard deviation parameters for the target distribution and the shared standard deviation parameter for the swap and swap-comparison distributions. Error bars represent 95% highest posterior density intervals. Models were fit separately for each color-distance segment, collapsed across subjects

for Experiment 1 can be found in Table S1 (Online Supplemental Material, OSM).

Experiment 2: Swaps with imprecise regularities

Experiment 1 demonstrated long-term memory distortions due to statistical regularities present during object encoding. Regardless of the original color of the object, participants sometimes misreported objects as being colored as the Rich color (swap errors). Patterns in our environment are often imprecise, however, in contrast to the regularity in Experiment 1: A well-defined, single Rich color that may have been easily exploited by the memory system. Experiment 2 specifically focuses on whether swap errors would persist given a less precise Rich color region as the basis for statistical regularities. Moreover, we collected two simultaneous datasets with this manipulation, one online using Amazon Mechanical Turk (MTurk) and one offline in the lab, to confirm the reliability of MTurk data collection. Our preregistration stated that if results were similar, data would be collapsed to increase statistical power.

Method

Participants

Two samples of 26 participants were simultaneously collected. Twenty-six participants (six male, 20 female; $M_{age} = 23.88$ years, SD = 5.48) were recruited from The Ohio State University (US$15 plus bonus) and 26 participants (age/sex not available due to experimenter error) were recruited through Amazon Mechanical Turk (MTurk) (US$9 plus bonus). All MTurk workers were from the USA, held an approval rating of ≥ 98%, and had previously completed at least 750 MTurk tasks prior to this experiment. To determine sample size for this preregistered experiment, we analyzed separate pilot data using the statistical tool G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). We estimated that a sample size of 26 participants would be necessary to detect swap errors with 90% power, given a .05 criterion of significance and a Cohen's d of 0.666; more details are provided in the OSF preregistration (https://osf.io/n7c5e/). We used a different power analysis from Experiment 1 because Experiment 1 was conducted chronologically after Experiments 2 and 3, and we used data from Experiments 2 and 3 in the power analysis for Experiment 1.

Additional participants were excluded in accordance with preregistered exclusion criteria (see Experiment 1, *Methods*): one participant was excluded from the in-lab sample and eight participants were excluded from the MTurk sample. All participants reported normal or corrected-to-normal vision, were naive to the purpose of the experiment, and provided informed consent in accordance with The Ohio State University institutional review board.

Materials and procedure

In Experiment 2, the biased color sampling was increased from 30% to 40% of objects, and these objects' colors were randomly sampled from a constrained 90° quadrant of color space (i.e., Rich quadrant) rather than a single color (see Fig. 4).

In-lab participants for Experiment 2 were tested in a controlled environment. Participants were seated in a small room without access to electronic devices and positioned approximately 56 cm from the monitor (head position was not fixed). Stimuli were presented using MATLAB (Mathworks, Natick, MA, USA) and Psychophysics Toolbox (Brainard, 1997) on a 24-in. widescreen LCD monitor with a screen resolution of 1,920 x 1,080 pixels. The monitor was color-calibrated with a Minolta CS-100 colorimeter. All stimuli were presented on a

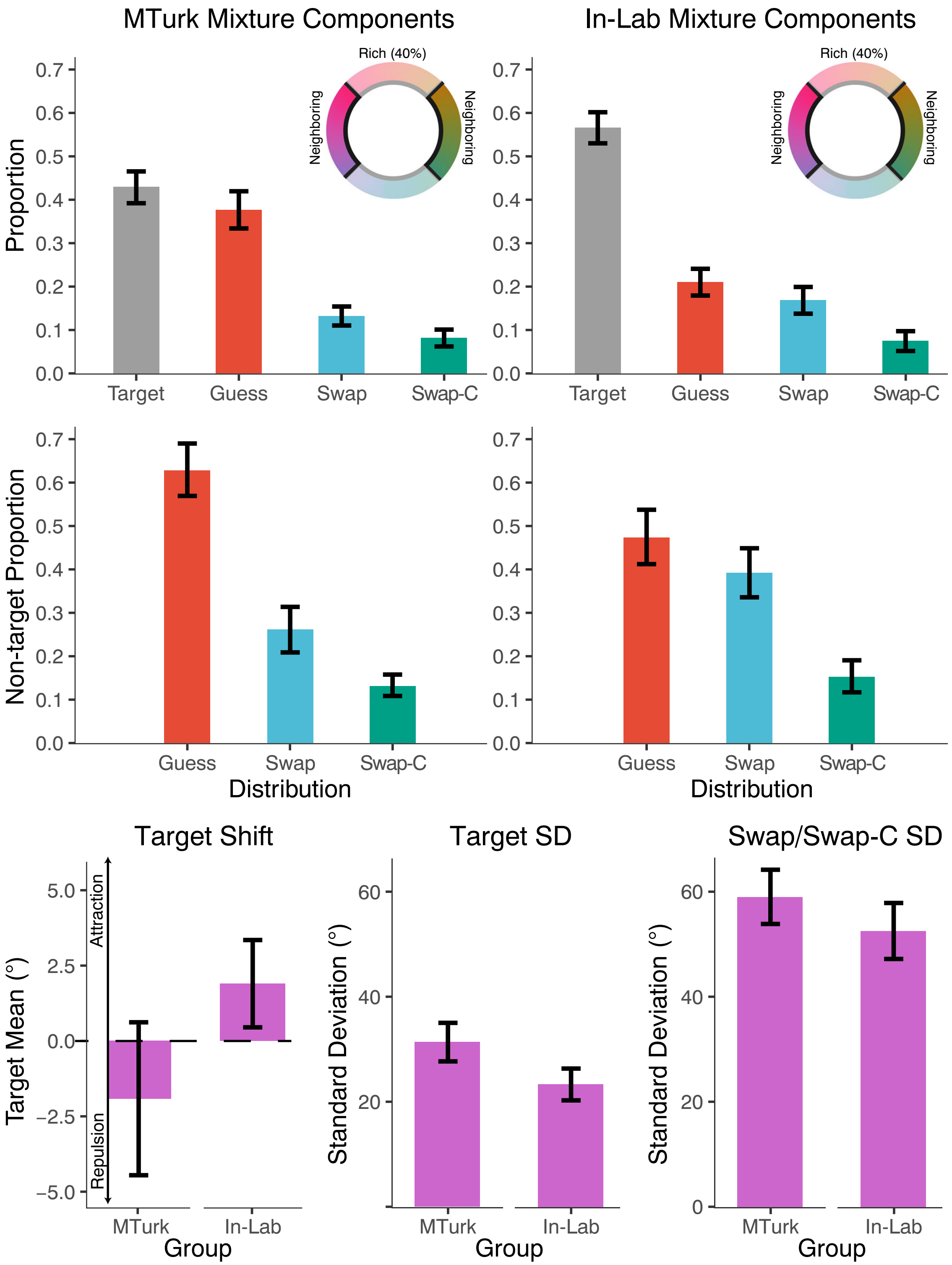

Fig. 4 Model fits for Experiment 2 for each group (MTurk and in-lab). Models were fit to each participant individually, including only memory responses for objects originally sampled from the Neighboring quadrants (objects sampled 46–135° from the center of Rich quadrant). Bar plots depict average parameter estimates and error bars depict standard error of the mean. The top panel depicts the absolute proportion of each mixture distribution. The middle panel depicts the relative, non-target proportion of each mixture distribution (relative proportion = absolute proportion / (1-absolute target proportion)). The bottom panel depicts the remaining parameter estimates. Overall, both groups similarly demonstrated significant swap errors (swap > swap-comparison). The only significant differences between groups were in the proportion of target responses and the (absolute) proportion of random guessing: MTurk participants demonstrated worse memory performance as indicated by these parameters

white background. At the typical viewing distance, objects subtended a visual angle of 10.24°, the fixation cross presented between image presentations subtended 0.20°, and change detection squares each subtended 1.23°. Aside from the above changes, the materials and procedure for Experiment 2 were the same as Experiment 1.

Analyses

Each participant's memory response distribution was fit using the same probabilistic mixture modeling approach described for Experiment 1, with a few modifications. In contrast to Experiment 1, where the swap distribution was centered on each participant's Rich color (which was a single value), in Experiment 2, the swap distribution was centered on the center of each subject's Rich quadrant (i.e., the average studied color, which we will refer to as the Rich color to keep terminology consistent across experiments). Additionally, we could not use the same "Short" (23–67°), "Medium" (68–112°), and "Long" (113–157°) color-distance segments as Experiment 1 because of the use of a Rich quadrant (the "Short" color-distance segment would have overlapped with the Rich quadrant). Thus, our main analyses focused on memory responses from objects originally sampled from the Neighboring quadrants (objects sampled 46–135° from the Rich color, see Fig. 4), and later in the paper we report data collapsed across experiments that examine shift errors as a function of more finely binned color distances. This is a deviation from our preregistration where we stated that we would fit separate models for objects sampled from within the Rich quadrant and objects sampled from outside the Rich quadrant (we realized such modeling would be uninformative; for further details see amendment.docx on OSF: https://osf.io/n7c5e/; note this also applies to Experiment 3). The use of a Neighboring quadrant afforded us sufficient statistical power to fit models to individual participant data (see "Assessing parameter reliability" section of the OSM), in contrast to Experiment 1, so within-subjects statistics were used to assess significance. Swapping was assessed by comparing the proportion of swap (S) to swap-comparison (C) errors; shifts were considered significant if μ significantly deviated from zero. We additionally conducted between-subjects statistics to compare model parameter estimates between MTurk and in-lab participants.

Results and discussion

We first tested whether overall memory performance was different between groups. We observed that MTurk participants showed worse overall performance as indicated by a significant difference in the proportion of target responses (M_{MTurk} = .421, $M_{\text{In-Lab}}$ = .554; $t(50) = 2.57$, $p = .013$, $d = .71$).

Because of this difference in target proportions, for the subsequent analyses comparing the different types of errors across groups, we examined both the absolute proportions (directly from the mixture model) and the non-target, or *relative*, proportions of each non-target distribution (guessing, swap, swap-comparison). For example, the relative proportion of random guessing was calculated by dividing the (absolute) proportion of random guessing by the total proportion of all non-target errors (1 – target proportion). (Note that our preregistrations only specified that we would examine "mixture proportions" with no reference to absolute vs. relative proportions.) The relative proportion of random guessing may be a more informative measure than the absolute proportion of random guessing here because it accounts for how changes in the target proportion alter the proportions of the other mixture distributions (i.e., mixtures are dependent on each other). For example, the MTurk group demonstrated a decreased target proportion compared to the in-lab group. This means that the sum of the non-target mixture proportions in the MTurk group is larger than the sum of the non-target mixture proportions for the in-lab group, and we need to correct for this baseline difference to detect non-target proportional differences between groups. In other words, the difference in the target proportion between groups could automatically lead to a significant difference in the proportions for all the other mixture distributions; we want to test whether these non-target distributions are different between groups after accounting for the difference in the target proportions.

Swap errors

We performed a repeated-measures ANOVA with (absolute) proportion (swap, swap-comparison) as the within-subjects factor and group (MTurk, in-lab) as the between-subjects factor. Robust swap errors were observed as indicated by a significant main effect of proportion ($F(1,50) = 9.59$, $p = .003$, $\eta^2_p = .161$). The presence of swap errors did not significantly vary across the two groups, as indicated by a non-significant main effect of group and a non-significant distribution x group interaction (all $Fs < 1$).

Similar results were observed when relative proportion was used as the dependent variable instead of absolute proportion. Robust relative swap errors were observed as indicated by a significant main effect of proportion ($F(1,50) = 16.68$, $p < .001$, $\eta^2_p = .250$), but the presence of relative swap errors did not significantly vary across the two groups, as indicated by a non-significant main effect of group ($F < 1$) and a non-significant proportion x group interaction ($F(1,50) = 1.47$, $p = .231$, $\eta^2_p = .029$).

Overall, the use of a Rich quadrant as a more imprecise statistical regularity induced swap errors similar to those induced by a single Rich color in Experiment 1. Moreover, these swap errors were reliably observed using individual models per subject (in contrast to the "super-subject" approach from Experiment 1). Swap errors were similarly observed across MTurk and in-lab groups, indicating that our findings can generalize across these populations and justifying the use of MTurk participants in our other experiments.

Shift errors

We explored the existence of shift errors as indicated by significant positive or negative μ parameter estimates for the target distribution. There was no evidence for attraction or repulsion for the MTurk group (one-sample t-test against 0; M = -1.93, SEM = 2.54; $t(25) = 0.76$, $p = .455$, $d = .21$) or the in-lab group (M = 0.96, SEM = 2.90; $t(25) = 1.32$, $p = .194$, $d = .37$), and no significant difference between groups ($t(50) = 1.31$, $p = .195$, $d = .36$). Collapsing across groups, there was evidence for neither attraction nor repulsion (M = -0.01, SEM = 1.47; $t(51) = 0.01$, $p = .995$, $d = .00$). This null result is not particularly surprising, given that we only analyzed trials where the object color was sampled from the Neighboring quadrant (46–135° from the Rich color), which spanned color distances that might be expected to produce repulsion (shorter distances) and attraction (longer distances) in accordance with the relational representation model (Bae & Luck, 2017; see also Golomb, 2015). In the later *Collapsed experiments* section, we bin the memory responses from this experiment into smaller color-distance segments to potentially observe systematic repulsion and attraction.

Other parameters

We tested for a difference in the relative proportion of random guessing between groups and observed that there was no significant difference in relative proportions ($M_{\text{MTurk}} = .615$, $M_{\text{In-Lab}} = .463$; $t(50) = 1.77$, $p = .082$, $d = .49$), despite there being a significant difference in absolute proportion of random guessing (greater for the MTurk group compared to the in-lab group: $M_{\text{MTurk}} = .369$, $M_{\text{In-Lab}} = .207$; $t(50) = 3.03$, $p = .004$, $d = .84$). This suggests that, while the MTurk participants demonstrated worse overall performance as measured by a decreased proportion of target responses, there was not unequal allocation of the remaining proportion of non-target responses to the random guessing distribution. Exploratory analyses of the standard deviation of the target and swap/swap-comparison distributions revealed no significant difference between groups (MTurk, in-lab) in the target standard deviation parameter ($t(50) = 1.68$, $p = .100$, $d = .47$) or the swap/swap-comparison standard deviation parameter ($t(50) = 0.87$, $p = .390$, $d = .24$).

Experiment 3: Increased memory load

In Experiment 3, we increased the number of studied/tested objects per block from 40 to 90 objects to explore the role of memory load on systematic memory distortions. We predicted that increasing the working memory load would further tax the memory system, which could influence the likelihood of swap and/or shift errors. Moreover, this manipulation might help differentiate between the sources of swap errors. For example, when memory load is increased, participants may be less certain about the correct color of an object, which could result in more guessing. If swap errors arise from an implicit or explicit biased guessing strategy, then we might expect to also see a greater proportion of swap errors with higher memory load. We will further explore the source(s) of swap errors in the Combined Experiments section, where we analyze confidence range reports and the post-experiment survey.

Here we collected two datasets, one where participants studied 90 objects per block (four blocks total), and a control dataset replicating Experiment 2 where participants studied 40 objects per block (nine blocks total). We simultaneously collected these datasets to try to control for random variables (e.g., time of day, day of the week).

Method

Participants

The 90-objects and 40-objects groups were simultaneously collected on MTurk. The 90-objects group included 26 participants (eight male, 18 female; M_{age} = 44.19 years, SD = 10.28), and the 40-objects group included 26 participants (13 male, 13 female; M_{age} = 39.77 years, SD = 12.56). We used the same preregistered sample size as Experiment 2 (more details are provided in the OSF preregistration: https://osf.io/n7c5e/).

Additional participants were excluded in accordance with preregistered exclusion criteria (see Experiment 1, *Methods*): 21 participants were excluded from the 90-objects group and six participants were excluded from the 40-objects group. As discussed below, the 90-objects task was intended to increase memory load and be a harder task, but we were not expecting such a large number of excluded participants, and it is unclear whether these exclusions are participants who were non-compliant or who were performing the task properly, just with much lower accuracy. We thus report parameter estimates for the subjects who passed our preregistered exclusion criteria as well as parameter estimates when no participants were excluded. All MTurk workers were from the USA, held an approval rating of ≥ 98%, had previously completed at least 750 MTurk tasks prior to this experiment, and were paid US$9 (plus bonus) for participation. All participants reported normal or corrected-to-normal vision, were naive to the purpose of the experiment, and provided informed consent in accordance with The Ohio State University institutional review board.

Materials and procedure

The materials and procedure used for the 40-objects group were identical to those for the MTurk group from Experiment 2. The only difference for the 90-objects group was that the number of objects per block increased from 40 to 90. The total number of studied and tested objects remained the same at 360 objects (nine blocks for the 40-objects group and four blocks for the 90-objects group).

Analyses

Each participant's memory response distribution was fit using the probabilistic mixture modeling approach described for Experiment 2. Like in Experiment 2, the swap distribution was centered on the center of each subject's Rich quadrant, and only memory responses from objects originally sampled from the Neighboring quadrants were analyzed. Swapping was assessed by comparing the proportion of swap (*S*) to swap-comparison (*C*) errors; shifts were considered significant if μ significantly deviated from zero. We conducted between-subject statistics to compare model parameter estimates between the 40-objects and 90-objects groups.

Results and discussion

We first tested whether overall performance was different between groups. Somewhat surprisingly, our initial analysis showed no significant difference in the proportion of target responses between groups (M_{40} = .468, M_{90} = .383; $t(50)$ = 1.66, p = .103, d = .461), suggesting that there was similar overall performance despite a difference in memory load between groups. However, as noted above, our preregistered exclusion criteria prompted the removal of 21 participants in the 90-object group and six participants in the 40-object group. While we implemented the exclusion criteria to eliminate inattentive and/or non-compliant individuals, it appears likely that the increased difficulty for the 90-object group led to more exclusions in that group. We thus looked at overall performance differences across groups when not excluding any participants (N = 79). Results here showed the 40-objects group indeed demonstrated better overall performance compared to the 90-objects group as indicated by a larger target proportion (M_{40} = .379, M_{90} = .241; $t(77)$ = 2.95, p = .004, d = .67). Although our preregistered exclusion criteria may have been inappropriate because it excluded too many participants, to provide an unbiased analysis of the data, we will proceed to report all analyses for the subjects who passed our preregistered exclusion criteria in addition to reporting all analyses when no participants were excluded (see Fig. 5 for model parameter estimates when no participants were excluded and see Tables S3 and S4 (OSM) for complete parameter estimates when participants were excluded and not excluded).

Swap errors

To identify whether swap errors ($S > C$) were observed in either group, we performed a repeated-measures ANOVA with (absolute) proportion (swap, swap-comparison) as a within-subjects factor and group (40-objects, 90-objects) as a between-subjects factor. Robust swap errors were observed as indicated by a significant main effect of proportion (subjects passed exclusion criteria: $F(1,50)$ = 22.42, p < .001, η^2_p = .310; all subjects: $F(1,77)$ = 9.30, p = .003, η^2_p = .108). The presence of swap errors did not significantly vary across the two groups, as indicated by a non-significant main effect of group (subjects passed exclusion criteria: $F(1,50)$ = 1.17, p = .285, η^2_p = .023; all subjects: $F(1,77)$ = 0.72, p = .400, η^2_p = .009) and a non-significant interaction between proportion and group (subjects passed exclusion criteria: $F(1,50)$ = 0.17, p = .682, η^2_p = .003; all subjects: $F(1,77)$ = 0.04, p = .839, η^2_p = .001).

Results were similar between groups when relative mixture proportions were used as the dependent variable instead of absolute proportions. Figure 5 shows that the 90-objects group seemed to increase the non-target mixtures proportionally. A repeated-measures ANOVA with relative proportion (relative swap, relative swap-comparison) as the within-subjects factor and group (40-objects, 90-objects) as the between-subjects factor revealed that robust relative swap errors were observed as indicated by a significant main effect of proportion (subjects passed exclusion criteria: $F(1,50)$ = 25.54, p < .001, η^2_p = .338; all subjects: $F(1,77)$ = 13.71, p < .001, η^2_p = .151). Again, the presence of relative swap errors did not significantly vary across the two groups, as indicated by a non-significant main effect of group and a non-significant proportion x group interaction (all *F*s < 1).

Together, these results suggest that while increased memory load seemed to induce poorer overall memory performance (as demonstrated by a smaller target proportion if we include all participants), the remaining mixtures seemed to increase proportionally, suggesting that memory load did not interact with the relative proportion of swap errors. The lack of an interaction between overall memory performance and relative swap errors is in line with the results of Experiment 2, where the MTurk group demonstrated poorer memory performance compared to the in-lab group. If an explicit or

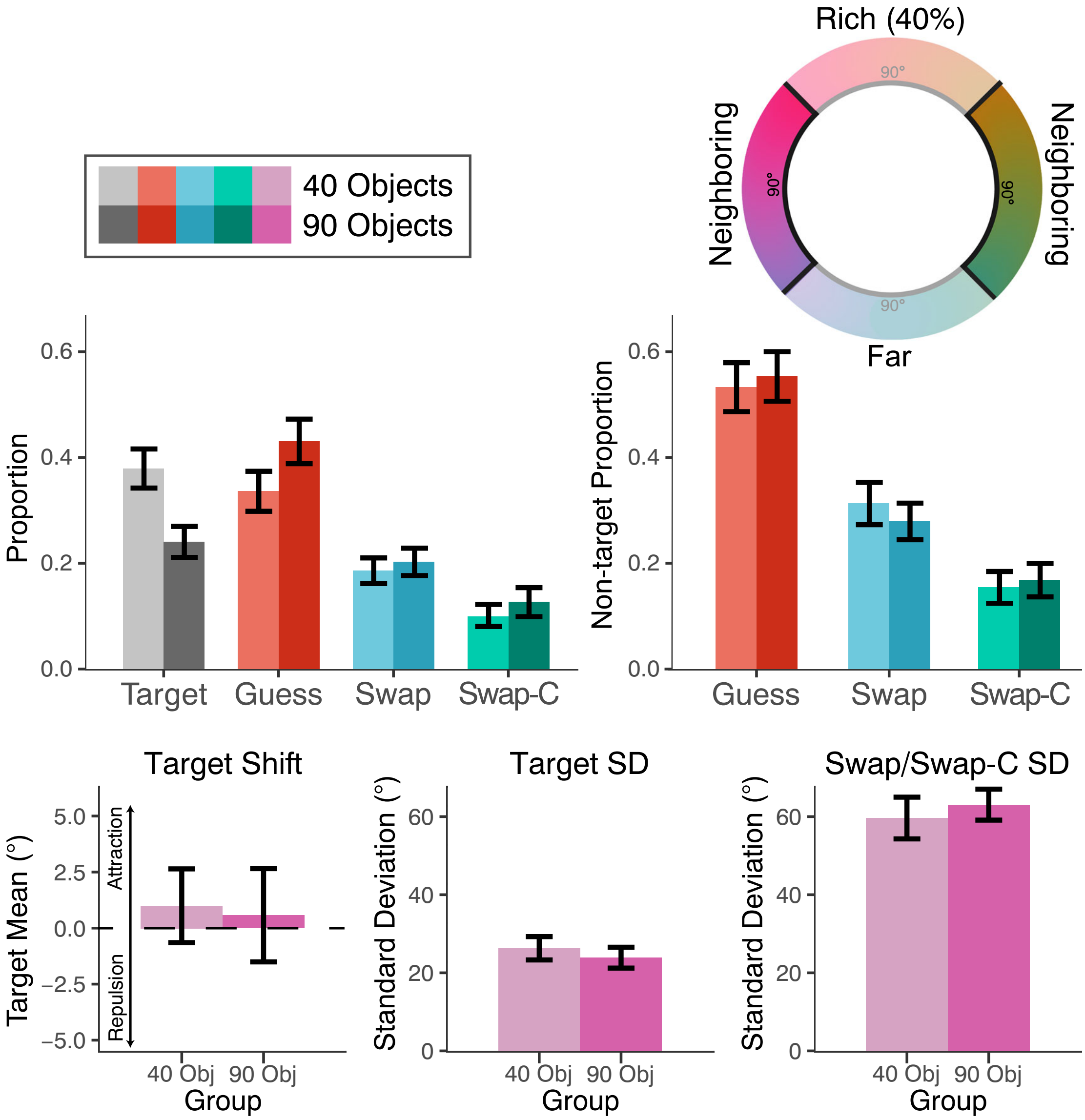

Fig. 5 Model fits for Experiment 3 for each group (40-objects and 90-objects), not excluding any participants. Models were fit to each participant individually, including only memory responses for objects originally sampled from the Neighboring quadrants. Bar plots depict average parameter estimates and error bars depict standard error of the mean. The top-left plot depicts the absolute proportion of each mixture and the top-right panel depicts the relative, non-target proportion of each mixture (relative proportion = absolute proportion / (1-absolute target proportion)). The bottom panel depicts the remaining parameter estimates. Overall, both groups similarly demonstrated significant swap errors (swap > swap-comparison). The only significant difference found between groups was in the proportion of target responses: the 90-objects group showed overall worse memory performance as indicated by a smaller target proportion

implicit biased guessing strategy was the predominant source of swap errors, then increasing memory load should have led to a larger proportion of swap errors. The lack of an increase in swapping suggests that swap errors may reflect more than an implicit or explicit biased guessing strategy, perhaps reflecting some portion of responses where participants reported around the Rich color because they truly believed that it was the correct color (i.e., false memory). We further explore the idea of different sources of swap errors in the *Collapsed experiments* section, where we examine confidence reports.

Shift errors

Like Experiment 2, systematic repulsion or attraction was not observed for items sampled from the Neighboring quadrants. There was no evidence for either attraction or repulsion overall for either the 40-objects group (one-sample t-test against 0; subjects passed exclusion criteria: M = -0.36, SEM = 1.73,

$t(25) = 0.21$, $p = .835$, $d = .06$; all subjects: $M = 1.02$, SEM = 1.65, $t(31) = 0.62$, $p = .540$, $d = .13$) or the 90-objects group (subjects passed exclusion criteria: $M = 1.65$, SEM = 1.85; $t(25) = 0.89$, $p = .382$, $d = .25$; all subjects: $M = 0.58$, SEM = 2.08; $t(46) = 0.28$, $p = .782$, $d = .13$). As noted in Experiment 2, the lack of shift errors could be because the Neighboring quadrants (46–135° from the Rich color) spanned color distances that might be expected to produce repulsion (shorter distances) and attraction (longer distances) in accordance with the relational representation model (Bae & Luck, 2017; see also Golomb, 2015).

Other parameters

The absolute proportion of random guessing was similar between groups (subjects who passed exclusion criteria: $M_{40} = .271$, $M_{90} = .301$; $t(50) = 0.54$, $p = .594$, $d = .15$; all subjects: $M_{40} = .336$, $M_{90} = .430$; $t(77) = 1.57$, $p = .120$, $d = .37$). After accounting for the difference in the proportion of target responses, we did not observe unequal allocation of the remaining proportion of non-target responses to the random guessing distribution between groups, as indicated by a non-significant difference in the relative proportion of random guessing (subjects who passed exclusion criteria: $M_{40} = .499$, $M_{90} = .473$; $t(50) = 0.33$, $p = .740$, $d = .09$; all subjects: $M_{40} = .533$, $M_{90} = .553$; $t(77) = 0.30$, $p = .764$, $d = .07$). Exploratory analyses of the standard deviation of the target and swap/swap-comparison distributions revealed no significant difference between groups (40-objects, 90-objects) in the target standard deviation parameter (subjects passed exclusion criteria: $t(50) = 1.01$, $p = .315$, $d = .28$; all subjects: $t(77) = 1.56$, $p = .122$, $d = .36$) or the swap/swap-comparison standard deviation parameter (subjects passed exclusion criteria: $t(50) = 0.86$, $p = .394$, $d = .24$; all subjects: $t(77) = 1.16$, $p = .250$, $d = .27$).

Collapsed experiments analyses

Sources of swap errors (confidence range reports and post-experiment survey)

In the *Introduction*, we posed a distinction between three possible sources of swap errors: explicit biased guessing, implicit biased guessing, and false memory. Did participants make swap errors because they explicitly realized that most objects were of a certain color and that reporting the Rich color was the "best" guess when uncertain, to improve overall memory performance and reward (explicit biased guessing)? Were participants explicitly unaware of statistical regularities and yet still reported the Rich color when uncertain due to statistical learning (implicit biased guessing)? In these two cases, swap errors are occurring only when participants feel like they do not remember the correct color. A third possibility is that participants truly thought that the correct color of an object was the Rich color (i.e., false memory). One critical difference between false memory and implicit/explicit biased guessing is that only false memory should lead participants to confidently commit swap errors. As a reminder, there was no incentive or feedback provided in regard to confidence range reports, so participants had no reason report a narrow, highly confident range of colors if they were actually using biased guessing (also see the OSM for an analysis on the time-course of swap errors, which argues against swap errors emerging from feedback). To explore if a potential mixture of these sources underlies swap errors, or whether one source could fully account for swap errors in our datasets, we examined confidence range reports and the post-experiment survey results.

Participants were instructed to provide confidence ranges for every color report by highlighting the smallest portion of the color wheel they believed contained the original color and to highlight the entirety of the color wheel when they were completely guessing. When participants committed swap errors, were their confidence ranges more similar to the confidence ranges of correct reports (i.e., narrower confidence range, in support of false memory) or more similar to the confidence ranges of random guessing reports (i.e., wider confidence range, in support of implicit/explicit biased guessing)?

We did not preregister any confidence analyses except for Experiment 1, where we stated that we would separately model participants' most and least confident reports (by taking the median confidence width for each participant and grouping their trials into more and less confident groups). This analysis, in addition to other exploratory confidence analyses, is detailed in the OSM. Here we report an exploratory, "super-subject" analysis that combined data across experiments to increase statistical power (N = 148). We included memory reports for objects originally sampled 23–157° (encompassing Short, Medium, and Long color-distance segments) away from the Rich color. We separately modeled trials on which participants responded with very high confidence (confidence range < 30°) and trials on which participants responded with very low confidence (confidence range report > 90°). Note that the Rich quadrant extended 90°, so highlighting the entire rich quadrant would be classified as a low confidence response. Also note that if a participant were selecting confidence ranges by highlighting entire color categories, then this would span more than 30° (e.g., Hardman, Vergauwe, & Ricker, 2017). For instance, selecting all the "red" parts of our color wheel would span roughly 60–70°.

As an initial check of the data, we confirmed that participants were using the confidence range report as intended: there was a reliably larger proportion of correct target responses for the high-confidence (< 30°) trials compared to the low-confidence (> 90°) trials (high confidence: .411 [.399 .426], low confidence: .226 [.202 .253]; 0% overlapping). There was also a reliably smaller proportion of random guesses for the high-confidence trials compared to the low-confidence trials when using absolute proportions (high confidence: γ=.526 [.502 .545], low confidence: γ=.683 [.657 .705]; 0% overlapping) but not relative proportions (high confidence: rel γ=.894 [.859 .918], low confidence: rel γ=.883 [.857 .901]; 56% overlapping).

Critically, we observed credible swap errors in both the very-high-confidence and very-low-confidence trials, as indicated by nonoverlapping HDIs between the swap and swap-comparison parameters for both high-confidence trials (S = .052, [.041 .072], C = .010, [.002 .020]) and low-confidence trials (S = .090, [.069 .109], C = .000, [.000 .008]). However, participants made more swap errors on the low-confidence trials, in terms of both absolute and relative proportions of swap errors. The absolute proportion of swap errors (swap minus swap-comparison model parameters) was credibly different across confidence splits (high confidence: S-C = .047 [.033 .059]; low confidence: S-C = .088 [.069 .109]; 1.5% overlapping), as was the relative proportion of swap errors equating for overall performance differences (high confidence: relS – relC = .079 [.057 .101]; low confidence: relS – relC = .150 [.112 .181]; 1.5% overlapping). These results are consistent with at least some observed swap errors reflecting a type of biased guessing, where participants make relatively more of these errors on trials on which they felt uncertain of the true color. However, this confidence analysis cannot differentiate between explicit versus implicit guessing accounts, nor does it completely rule out false memory as an additional potential source of swap errors, because credible swap errors were observed even when participants were highly confident in their color reports.

To specifically investigate explicit versus implicit biased guessing, we ran another exploratory "super-subject" model that included only trials with a confidence range > 90° (i.e., trials where participants were more likely to be guessing) and compared participants who reported being explicitly aware of the statistical regularity against participants who were not explicitly aware of the statistical regularity, classified based on responses to the post-experiment survey (see Experiment 1, *Materials and procedure*). The post-experiment survey results for each question and each experiment are detailed in the OSM. When participants were asked if they believed objects often shared approximately the same color, 21/44 participants said yes in Experiment 1, 16/26 (MTurk) and 19/26 (in-lab) participants said yes in Experiment 2, and 26/32 (40-Objects) and 34/47 (90-Objects) participants said yes in Experiment 3. That is, 66.3% (116/175) of participants across all experiments responded that objects often shared approximately the same color, suggesting that this subset of participants possessed explicit knowledge of the statistical regularity by the end of the experiment.

Credible swap errors were observed in both groups (those who said "yes" vs. "no" regarding their knowledge of the regularities), as indicated by nonoverlapping HDIs between the swap and swap-comparison parameters ("yes" participants: S = .066, [.057 .097, 95% HDI], C = .000, [.000 .015]; "no" participants: S = .090, [.079 .108], C = .013, [.006 .026]). Importantly, the ratio of swap errors was similar between groups, as indicated by the lack of a credible difference between posterior probability densities for the difference of posteriors (swap minus swap-comparison) ("yes": S-C= .067 [.055 .086]; "no": S-C = .078 [.064 .090]; 40.5% overlapping). There was also no credible difference between the difference of posteriors for relative swapping ("yes": relS-relC = .095 [.068 .135]; "no": relS-relC = .158 [.104 .240]; 8.8% overlapping). We also found similar results (data not shown) when we further restricted the subset of participants who reported "yes" to those who also subsequently selected a color close to the Rich color on the second survey question. Thus, explicit knowledge did not seem to be necessary to produce swap errors on low-confidence trials because both groups of participants similarly exhibited swap errors.

Finally, we conducted one more exploratory analysis to test whether swap errors may have sometimes reflected false memory. We only included high-confidence trials (confidence range < 30°) and also only included participants who did not exhibit explicit knowledge of the Rich color in the post-experiment survey ("no" on question 1: N = 59). Despite these strict cutoffs meant to eliminate swap errors driven by implicit/explicit biased guessing, credible swap errors were still observed on these trials (S = .068, [.054 .092], C = .002, [.000 .010]).

Shift errors as a function of color distance

We began this study with the hypothesis that both swap errors and shift errors might emerge in long-term memory due to statistical learning. While we found robust evidence for swap errors across all of our datasets and analyses, the shift errors may be more condition-dependent. Specifically, for Experiment 1, we hypothesized that there might be attraction for the Long color-distance segment and repulsion for the Short segment. While we observed a trend for attraction and repulsion effects in Experiment 1, we did not find significant shift effects in Experiments 2 and 3 (possibly due to binning across a large Neighboring quadrant, which could obscure repulsion/attraction depending on how shift effects varied as a function of color distance). Given that all experiments involved a similar paradigm and showed similar mixture-proportion results, we performed an exploratory analysis where we collapsed data across all experiments and fit our modified swap model to memory errors, binned into five color distances (angular distance from the memory item's color to the participant's Rich color: bin 1: <23°, bin 2 [23-67°, i.e. "Short"], bin 3 [68-112°, i.e. "Medium"], bin 4 [113-157°, i.e. "Long"], and bin 5: >157 deg). We utilized the aggregate "super-subject" approach described in Experiment 1 and calculated 95% HDIs for the μ parameter of the target distribution for each bin (Fig. 6).

The direction of shift errors followed the predictions from the relational representation model (see Fig. 6; Bae & Luck, 2017; see also Golomb, 2015). Credible repulsion bias was observed for bins 1 and 2 (95% HDI not overlapping with the critical threshold of zero° of error), no credible shift errors were observed for bin 3, and credible attraction bias was observed for bin 4. Bin 5 showed attraction bias overall but had a 95% HDI that was not credibly above the critical threshold; note that recent work suggests that color distances are not psychologically scaled in a linear fashion and are less meaningful at these large distances (Schurgin et al., 2018), which would predict that bin 5 would be less likely to demonstrate reliable bias. These findings mirror the relational representation model described in visual working memory literature: similar features between the memory item and a reference point are more likely to produce repulsion (representations become more distinct) and dissimilar features between the memory item and a reference point are more likely to produce attraction (representations become generalized).

General discussion

The present three experiments consistently demonstrated swap errors in long-term memory due to statistical regularities present during object encoding. While less robust than swap errors, subtle shift errors were also observed in some cases, where the direction of errors was dependent on how far apart the memory item's color and the Rich color were in feature

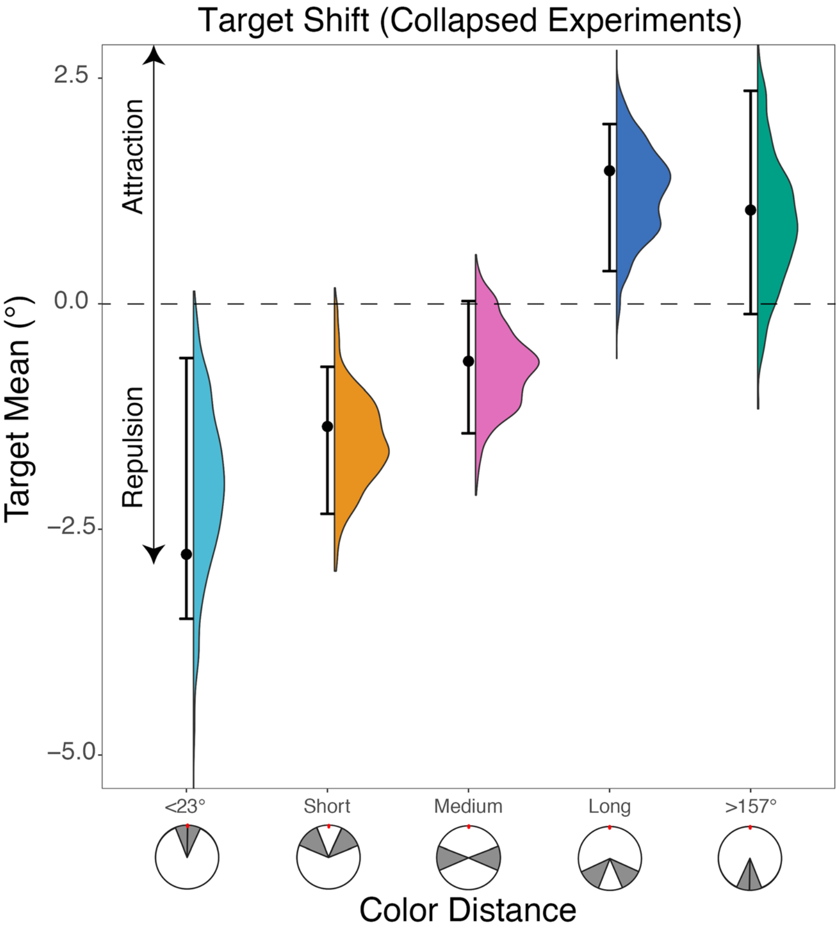

Fig. 6 Parameter estimates of the mean of the target distribution (μ) demonstrate repulsion (negative shift) and attraction (positive shift), dependent on the distance in color space between the original color of the memory item and the Rich color. Circles displayed under the x-axis represent the analyzed color-distance segment, with red dots depicting the location of the aligned Rich color. We binned trials according to their color distance (angular distance from the memory item's color to the participant's Rich color): bin 1: <23°, bin 2 (23–67°, i.e. "Short"], bin 3 (68–112°, i.e. "Medium"), bin 4 (113–157°, i.e. "Long"), and bin 5: >157°. Split-violin plots illustrate the posterior probability distributions of the mean shift parameter for the target distribution over 15,000 post-convergence samples. Black dots mark maximum a posteriori estimates, with whiskers representing 95% highest density intervals. Models were fit separately for each color-distance segment, collapsed across subjects. Memory responses for trials where the object was originally colored with the Rich color (0° color difference) or the color 180° away from the Rich color were not included

space. We elaborate on different possible sources of swap errors (explicit biased guessing, implicit biased guessing, and false memory) and their underlying mechanisms in the following paragraphs. We then relate the subtle attraction/repulsion effects observed across experiments to visual working memory literature and discuss limitations and overall implications for our findings.

To our knowledge, no previous studies have demonstrated swap and shift errors in long-term memory due to imposed statistical regularities. In visual working memory paradigms, however, the influence of prior experience on biased recall has been repeatedly observed. Prior experience is a useful cue for the memory system because it can trigger reinstatement and integration processes that allow for strengthened memories due to associative knowledge (van Kesteren, Brown, & Wagner, 2016). When a participant cannot accurately recall a memory item, learned regularities can (explicitly or implicitly) bias memory reports towards previously experienced stimuli. For example, in a visual working memory task, Pratte (2018) showed participants an array of colored squares and then asked them to report the location of a probed color. When the color probed had not originally appeared in the study array, participants' reports were still centered around study locations. This suggests that participants were using information about where memory items could appear to bias their guesses. Even during a perception task, Fan, Hutchinson, and Turk-Browne (2016) demonstrated that color reports can be biased by task-irrelevant past experience: when asked to report the color of a stimulus that was never presented, participants often chose the color associated with the stimulus in a previous, irrelevant portion of the experiment. The results from both of these examples may be due to participants using prior experience to maximize performance when memory certainty is low. This pattern of biased guessing could have been due to a conscious (explicit) decision on behalf of the participant to use prior experience to try to improve performance (explicit biased guessing), and/or it could have been due to the memory system automatically and implicitly incorporating prior experience to subsequently optimize behavior (implicit biased guessing; e.g., Perruchet & Pacton, 2006; Saffran et al., 1996).

False memory is a third potential source of our observed swap errors. False memory has been repeatedly observed and shown to be remarkably easy to induce (e.g., Deese, 1959; Loftus, 2003; Roediger & McDermott, 1995; Wixted et al., 2018). For example, Loftus, Miller, and Burns (1978) demonstrated that people are likely to mistakenly remember a stop sign for a previously encountered yield sign. The Deese-Roediger-McDermott paradigm is a reliable method for inducing false memories, where participants study a list of related words and later recall a related but non-presented word at the same frequency as actual studied words (Deese, 1959; Roediger & McDermott, 1995). While previous work has demonstrated that people are generally good at reporting their own memory quality (Fougnie, Suchow, & Alvarez, 2012; Rademaker, Tredway, & Tong, 2012; Suchow, Fougnie, Brady, & Alvarez, 2014), false memories are nonetheless present in spite of high confidence. Swap errors could reflect a type of false memory where participants recall the wrong color-pair association for an object with high confidence (Chen et al., 2019). While explicit/implicit biased guessing uses prior information to optimize behavior when memory strength is low, false memories occur when someone truly misremembers an item, offering potential insight to how items are represented in memory.

Although we cannot conclusively establish the source of our swap errors, the pattern of evidence suggests that the swap errors we observed due to our imposed statistical regularities (i.e., Rich color) likely reflected a mixture of implicit biased guessing, explicit biased guessing, and false memory. We reasoned that implicit and explicit biased guessing should occur most frequently when overall performance was low (such that guessing was more frequent) and when trials were reported with low confidence, and that our post-experiment survey results could be used to further help to delineate explicit and implicit biased guessing. Meanwhile, we reasoned that false memory should be associated with high confidence trials and should be observed even across participants who were explicitly unaware of statistical regularities. To summarize our results (including the collapsed experiments analyses that explored confidence and post-experiment survey reports, the results of Experiment 2 where MTurk participants showed overall worse performance than in-lab participants, and the results of Experiment 3 meant to explore swap errors as a function of memory load), we found the following: First, credible swap errors were similarly observed regardless of overall performance differences between groups, memory load, and online versus in-person task context. Second, swap errors were credibly observed across both low- and high-confidence trials but were more frequently observed on low-confidence trials, suggesting that implicit/explicit biased guessing accounted for swap errors more than false memory; however, false memory likely accounted for at least some portion of swap errors because swaps were credibly observed even across very high-confidence trials with explicitly unaware participants. Finally, swap errors were found regardless of post-experiment survey responses meant to gauge explicit awareness of statistical regularities. Thus, we were not able to delineate explicit and implicit biased guessing, but at the very least, explicit biased guessing could not fully account for the proportion of swap errors in our data by itself, and some portion of swap errors may have reflected true false memories. Regardless of their source, the core observation of swap errors is still important because it shows how statistical regularities can induce a reference point for subsequent long-term memory reports. Previous studies showing swap errors in long-term memory have not involved statistical regularities (e.g., Lew et al., 2016; Utochkin & Brady, 2019). Moreover, previous visual working memory papers have generally not attempted to investigate these different sources of memory distortions, so these analyses might also be of use for future experiments linking errors in visual working memory and long-term memory contexts.

In regard to shift errors, several visual working memory studies have observed attraction and repulsion effects for actively maintained memory items dependent on the distance in feature space between the memory item and a reference point (Bae & Luck, 2017; Golomb, 2015; Golomb et al., 2014). Bae and Luck (2017) described a relational representation model in which working memory items close in feature space to the reference point are easily confused and therefore need to be distinguished from each other, resulting in a negative shift (repulsion). Meanwhile, working memory items farther apart in feature space to the reference point may be generalized in order to decrease memory noise, resulting in a positive shift (attraction). These visual working memory findings are also conceptually related to the direction illusion and the tilt illusion, as well as their respective visual aftereffects (e.g., Gibson, 1937; Gibson & Radner, 1937; Hiris & Blake, 1996; Wenderoth & Johnstone, 1988; Wenderoth & Wiese, 2008; Wiese & Wenderoth, 2007). These are perceptual phenomena that similarly demonstrate attraction and repulsion dependent on the distance in feature space between perceptual stimuli (although the shift errors observed in working memory cannot be explained by the same perceptual mechanism; see Scotti, Hong, Leber, & Golomb, 2020). Another related perceptual phenomenon, serial dependence, refers to how items presented in close temporal succession may be perceived as biased towards or away from recently seen items (e.g., Fischer & Whitney, 2014; Kiyonaga, Scimeca, Bliss, & Whitney, 2017). Linking these findings from working memory and perception to long-term memory, long-term memory researchers have outlined two distinct computational processes in the hippocampus, pattern completion and pattern separation, that would predict attraction and repulsion respectively (see Yassa & Stark, 2011, for review). Therefore, the perceptual, working, and long-term memory systems can all be adapted to a similar relational framework, and this experiment demonstrates how it is possible to translate the relational memory distortions previously only observed in the visual working memory (and perception) domain over to the visual long-term memory domain.

It is important to note that while shift errors observed in the present study were consistent with the relational representation model, effects were small and only observed when experiments were collapsed. In addition, aggregate data can sometimes lead to results that are not representative of any single individual, which is one downside to the "super-subject" modeling approach (Estes, 1956; Heathcote, Brown, & Mewhort, 2000). The weak effects that we observed with shift errors may imply that shift errors are less prominent and less likely to occur than swap errors, but this may be heavily dependent on the experimental design. The weak effects we observed may also be influenced by the role of categorical colors in memory: there are inherent biases in color recall where responses are sometimes attracted towards prototypical colors (see Bae, Olkkonen, Allred, & Flombaum, 2015; Hardman, Vergauwe, & Ricker, 2017), and these color category effects may have introduced noise into our data that obscured attraction/repulsion in reference to the Rich color. Importantly, color category effects would introduce noise, not systematic attraction or repulsion bias, because it was equally likely that the nearest canonical color would be towards or away from the Rich color on any given trial. Nonetheless, the present findings suggest that the relational representation model may generalize to passively stored memory items, with similar relational mechanisms being able to account for perceptual, working memory, and long-term memory systems.

There are several possible future directions to extend and confirm the present findings. Confidence ratings could be more complex than can be accounted for by simple cutoffs. For example, in a visual working memory experiment, Honig, Ma, and Fougnie (2018) showed attraction towards the most frequent color given decreasing memory certainty. We could not sufficiently explore the role of confidence on shift errors in the present experiments because of the small effect size we observed for shift errors. Another future direction regards when memory distortions, including swap errors and shift errors, emerged during the stages of encoding, storage, and retrieval. This is of particular interest because pattern completion is thought to occur during encoding whereas pattern separation is thought to occur during retrieval (Duncan, Sadanand, & Davachi, 2012; O'Reilly & McClelland, 1994). New experimental designs and better model fitting procedures will allow for deeper investigations into how memory distortions arise.

Conclusion

The present findings demonstrate that visual long-term memory is easily distorted by implicitly learned patterns in our environment. Statistical regularities imposed while studying object colors resulted in systematic biases during subsequent memory retrieval. Many participants were not explicitly aware of the imposed regularities, and yet a significant proportion of responses were centered around the Rich color for these participants. Many of these swap errors were also high in reported confidence, suggesting that some portion of swap errors arose from false memories where a stored memory was incorrectly represented as the Rich color. In addition, responses were sometimes shifted away from or towards the Rich color, depending on the distance in feature space between the Rich color and the original object color. These shift errors mirrored the relational representation model observed in visual working memory, suggesting that similar memory distortion mechanisms may be present in working and long-term memory systems. In real-world situations where reliable memories are paramount, future work exploring how prior experience influences memory retrieval may yield crucial insights.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13414-020-02236-3.

Acknowledgements We thank Emma Wu Dowd for help with model analyses and statistical interpretation, Ayala Allon for sharing code that was used for the filler task, and Molly McKinney, Elliot Ping, Aisha Ayub, Needa Toofanny, and Zoe Zhang for assistance with data collection.

Author Contributions P.S. and A.L. conceptualized the initial study. P.S., Y.H., J.G., and A.L. contributed to theoretical motivation, study design, analysis plan, and writing of the paper. P.S. programmed the experiments and performed data collection and analyses.

Funding This study was funded by grants from the National Science Foundation (NSF DGE-1343012 to P.S.; NSF BCS-1632296 to A.L.; NSF BCS-1848939 to J.G. & A.L.), the National Institutes of Health (R01-EY025648 to J.G.), and the Alfred P. Sloan Foundation (to J.G.). Analyses reported in this publication were supported by an allocation of computing resources from The Ohio Supercomputer Center (1987).

Compliance with ethical standards

Competing Interests The authors declare no competing interests.

References

Aimone, J. B., Deng, W., & Gage, F. H. (2011). Resolving New Memories: A Critical Look at the Dentate Gyrus, Adult Neurogenesis, and Pattern Separation. *Neuron*, *70*(4), 589–596. https://doi.org/10.1016/j.neuron.2011.05.010

Allon, A. S., & Luria, R. (2017). Compensation mechanisms that improve distractor filtering are short-lived. *Cognition*. https://doi.org/10.1016/j.cognition.2017.03.020

Bae, G.-Y., & Luck, S. J. (2017). Interactions between visual working memory representations. *Attention, Perception, & Psychophysics*, (August), 2376–2395. https://doi.org/10.3758/s13414-017-1404-8

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. https://doi.org/10.1037/xge0000076

Bartlett, F. C. (1932). Remembering: A Study in Experimental and Social Psychology. *Cambridge, Social Psychology*. https://doi.org/10.1111/j.2044-8279.1933.tb02913.x

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), 7–7. https://doi.org/10.1167/9.10.7

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, 49(6), 1622-1631.

Brady, T., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*. https://doi.org/10.1177/0956797610397956

Brady, T., Konkle, T., Alvarez, G., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Brady, T., Konkle, T., Alvarez, G., & Oliva, A. (2013). Real-world objects are not represented as bound units: Independent forgetting of different object details from visual memory. *Journal of Experimental Psychology: General*, *142*(3), 791–808. https://doi.org/10.1037/a0029649

Brady, T., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. (2013). Visual Long-Term Memory Has the Same Limit on Fidelity as Visual Working Memory. *Psychological Science*, *24*(6), 981–990. https://doi.org/10.1177/0956797612465439

Brady, T. F., Schacter, D. L., & Alvarez, G. (2018, August 16). The adaptive nature of false memories is revealed by gist-based distortion of true memories. *PsyArXiv*. https://doi.org/10.31234/osf.io/zeg95

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*. https://doi.org/10.1163/156856897X00357

Brainerd, C. J., & Reyna, V. F. (2008). *The Science of False Memory*. https://doi.org/10.1093/acprof:oso/9780195154054.001.0001

Brodeur, M. B., Dionne-dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a New Set of 480 Normative Photos of Objects to Be Used as Visual Stimuli in Cognitive Research. *PLoS ONE, 5*(5). https://doi.org/10.1371/journal.pone.0010773

Carpenter, A. C., & Schacter, D. L. (2017). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000340

Chen, J., Leber, A. B., & Golomb, J. D. (2019). Attentional capture alters feature perception. *Journal of experimental psychology. Human perception and performance*.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate free recall. *Journal of Experimental Psychology*, *58*, 17–22. https://doi.org/10.1037/h0046671

Dowd, E. W., & Golomb, J. D. (2019). Object-feature binding survives dynamic shifts of spatial attention. *Psychological Science*, *30*(3), 343-361.

Duncan, K., Sadanand, A., & Davachi, L. (2012). Memory's Penumbra: Episodic Memory Decisions Induce Lingering Mnemonic Biases. *Science*, *337*(6093), 485–487. https://doi.org/10.1126/science.1221936

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological bulletin*, 53(2), 134.

Fan, J. E., Hutchinson, J. B., & Turk-Browne, N. B. (2016). When past is present: Substitutions of long-term memory for sensory evidence in perceptual judgments. *Journal of Vision*, *16*(8), 1. https://doi.org/10.1167/16.8.1

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(175). https://doi.org/10.3758/BF03193146

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature neuroscience*, *17*(5), 738-743.

Fougnie, D., Suchow, J., & Alvarez, G. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1228–1229. https://doi.org/10.1038/ncomms2237

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457-472. doi: https://doi.org/10.1214/ss/1177011136

Geng, J. J., & Behrmann, M. (2002). Probability cuing of target location facilitates visual search implicitly in normal participants and patients with hemispatial neglect. *Psychological Science*, 13(6), 520-525.

Gibson, J. J. (1937). Adaptation, after-effect, and contrast in the perception of tilted lines. II. Simultaneous contrast and the areal restriction of the after-effect. *Journal of Experimental Psychology*, 20(6), 553-569.

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of Experimental Psychology*, 20(5), 453-467.

Golomb, J. D. (2015). Divided spatial attention and feature-mixing errors. *Attention, Perception, and Psychophysics*, *77*(8), 2562–2569. https://doi.org/10.3758/s13414-015-0951-0

Golomb, J. D., L'Heureux, Z. E., & Kanwisher, N. (2014). Feature-Binding Errors After Eye Movements and Shifts of Attention. *Psychological Science, 25*(5), 1067–1078. https://doi.org/10.1177/0956797614522068

Guerin, S. A., Robbins, C. A., Gilmore, A. W., & Schacter, D. L. (2012). Retrieval failure contributes to gist-based false recognition. *Journal of Memory and Language*. https://doi.org/10.1016/j.jml.2011.07.002

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology: Human Perception and Performance*, 43(1), 30. https://doi.org/10.1037/xhp0000290

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, 7(2), 185-207.

Hemmer, P., & Steyvers, M. (2009). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin and Review*, *16*(1), 80–87. https://doi.org/10.3758/PBR.16.1.80

Hiris, E., & Blake, R. (1996). Direction repulsion in motion transparency. *Visual Neuroscience*, 13(1), 187-197.

Honig, M., Ma, W. J., & Fougnie, D. (2018). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. *bioRxiv,* 1–20. https://doi.org/10.1101/306225

Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision*, *10*(2), 1–27. https://doi.org/10.1167/10.2.24

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, 129(2), 220.

Jiang, Y. V., Swallow, K. M., Rosenbaum, G. M., & Herzig, C. (2013). Rapid acquisition but slow extinction of an attentional bias in space. *Journal of Experimental Psychology: Human Perception and Performance*, 39(1), 87.

Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial dependence across perception, attention, and memory. *Trends in Cognitive Sciences*, *21*(7), 493-497.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691611406925

Leutgeb, S., & Leutgeb, J. K. (2007). Pattern separation, pattern completion, and new neuronal codes within a continuous CA3 map. *Learning & Memory, 14*(11), 745-757. https://doi.org/10.1101/lm.703907

Lew, T. F., Pashler, H. E., & Vul, E. (2016). Fragile associations coexist with robust memories for precise details in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(3), 379.

Loftus, E. (2003). Our changeable memories: Legal and practical implications. *Nature Reviews Neuroscience*, *4*(3), 231–234. https://doi.org/10.1038/nrn1054

Loftus, E., & Hoffman, H. (1989). Misinformation and Memory: The Creation of New Memories. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/0096-3445.118.1.100

Loftus, E., Miller, D., & Burns, H. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology. Human Learning and Memory*. https://doi.org/10.1037/0278-7393.4.1.19

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature,* 390, 279-281. https://doi.org/10.1038/36846

Newman, E. J., & Lindsay, D. S. (2009). False memories: What the hell are they for? *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.1613

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*. https://doi.org/10.1002/hipo.450040605

Ohio Supercomputer Center. (1987). Ohio Supercomputer Center. Columbus OH: Ohio Supercomputer Center. http://osc.edu/ark:/19495/f5s1ph73.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006

Pratte, M. S. (2018). Swap errors in spatial working memory are guesses. *Psychonomic Bulletin and Review,* 1-9. https://doi.org/10.3758/s13423-018-1524-8

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13), 21–21. https://doi.org/10.1167/12.13.21

Roediger, H. L., & McDermott, K. B. (1995). Creating False Memories: Remembering Words Not Presented in Lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/0278-7393.21.4.803

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-months-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences*, *15*(10), 467-474. https://doi.org/10.1016/j.tics.2011.08.004

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2018). Psychophysical scaling reveals a unified theory of visual memory strength. *bioRxiv*, *325472*. https://doi.org/10.1101/325472

Scotti, P. S., Hong, Y., Leber, A. B., & Golomb, J. (2020, May 9). Active, not passive, visual working memory maintenance produces repulsion. *PsyArXiv.* https://doi.org/10.31234/osf.io/md5h4

Suchow, J., Fougnie, D., Brady, T., & Alvarez, G. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception, and Psychophysics*. https://doi.org/10.3758/s13414-014-0690-7

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, 13, 9-9. https://doi.org/10.1167/13.10.9

Treisman, A. (1988). Features and Objects: The Fourteenth Bartlett Memorial Lecture. *The Quarterly Journal of Experimental Psychology Section A*. https://doi.org/10.1080/02724988843000104

Treisman, A. (1998). Feature-binding, attention and object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.1998.0284

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

Utochkin, I. S., & Brady, T. F. (2019). Independent storage of different features of real-world objects in long-term memory. *Journal of Experimental Psychology: General*, *149*(3), 530-549. https://doi.org/10.1037/xge0000664

van Kesteren, M. T., Brown, T. I., & Wagner, A. D. (2016). Interactions between Memory and New Learning: Insights from fMRI Multivoxel Pattern Analysis. *Frontiers in Systems Neuroscience*, *10*(May), 1–5. https://doi.org/10.3389/fnsys.2016.00046

Wenderoth, P., & Johnstone, S. (1988). The different mechanisms of the direct and indirect tilt illusions. *Vision Research*, 28(2), 301-312.

Wenderoth, P., & Wiese, M. (2008). Retinotopic encoding of the direction aftereffect. *Vision Research*, 48(19), 1949-1954.

Wiese, M., & Wenderoth, P. (2007). The different mechanisms of the motion direction illusion and aftereffect. *Vision Research*, 47(14), 1963-1967.

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the Reliability of Eyewitness Memory. *Perspectives on Psychological Science*, *13*(3), 324–335. https://doi.org/10.1177/1745691617734878

Yassa, M. A., & Stark, C. E. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Yoo, A. H., Klyszejko, Z., Curtis, C. E., & Ma, W. J. (2018). Strategic allocation of working memory resource. *Scientific Reports, 8*(1), 1–8. https://doi.org/10.1038/s41598-018-34282-1

Zokaei, N., Heider, M., & Husain, M. (2014). Attention is required for maintenance of feature-binding in visual working memory. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1191–1213. https://doi.org/10.1080/17470218.2013.852232

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.