AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

ISSN: 0033-295X

https://doi.org/10.1037/rev0000645

Dynamics of Variability and Bias in Working Memory

Ivan Tomić[1, 2], Sebastian Schneegans[2], and Paul M. Bays[2]

[1] Department of Psychology, Faculty of Humanities and Social Sciences, University of Zagreb

[2] Department of Psychology, University of Cambridge

Visual working memory performance declines with the quantity of material stored and the duration of its maintenance, but it is disputed whether these constraints are independent or interact. Specifically, different studies have reached contradictory conclusions on whether greater memory load accelerates deterioration of the stored information during retention, a question with implications for storage mechanisms. Meanwhile, prominent models of visual working memory based on stable attractors predict that recall should become increasingly biased toward a fixed set of canonical features with longer retention, but empirical support is mixed. Uncertainty is further exacerbated by limited data, methodological variation, and inconsistent analytical approaches across studies. Moreover, previous studies have not always distinguished response variability from stimulus-specific biases nor considered the possibility of swap errors (intrusions). To address these issues, we conducted six new experiments and reanalyzed data from seven published experiments using consistent analytical methods to examine the effects of set size and delay on variability and bias. Our results show that recall variability increased with longer delays and higher set sizes, consistent with prior findings, and importantly, that these effects did not operate independently: The effect of delay was amplified at higher set sizes. We also observed consistent evidence for systematic biases across stimuli, but unlike variability, bias amplitude did not change with either set size or retention interval. These results support diffusion accounts of visual working memory maintenance, in which variability grows through the accumulation of random error, while challenging stable attractor accounts that predict systematic drift toward canonical values during maintenance.

Keywords: working memory, forgetting, recall error, bias, stimulus-specific effects

Supplemental materials: https://doi.org/10.1037/rev0000645.supp

Visual working memory (VWM) refers to the brain's ability to temporarily maintain visual information so it can continue to guide behavior when the sensory input is no longer available (Bays et al., 2024; D'Esposito & Postle, 2015). Despite its central role in everyday behavior, the process of encoding information in VWM is imperfect, limiting the fidelity with which information can be reconstructed and leading to errors in recall. These imperfections are further amplified with the passage of time since encoding and as the number of items held in memory increases, highlighting two key constraints on VWM recall.

Effects of Memory Load on Response Variability

The observation that recall precision decreases as the number of objects held in memory increases is considered a hallmark finding in VWM research. In studies using analogue report tasks (Prinzmetal et al., 1998; Wilken & Ma, 2004; see Tomić et al., 2024, for analysis), this effect is observed as a broadening of response error distributions and a corresponding increase in measures of dispersion as more objects are committed to memory. In addition to this, error distributions exhibit excess kurtosis relative to a circular normal distribution, with disproportionately high density in the tails—the latter effect also intensifies with increasing set size. This finding has proven to be robust and reproducible across numerous visual feature dimensions (Bays, 2014, 2016; Bays et al., 2009; Salmela et al., 2019; Schurgin et al., 2020; Tomić & Bays, 2024b; van den Berg et al., 2012, 2014).

The mechanistic basis of the set size effect on precision and long-tailed error distributions has been widely debated, with differing

Kara J. Blacker served as action editor.

Ivan Tomić https://orcid.org/0000-0001-6046-9642

Paul M. Bays https://orcid.org/0000-0003-4684-4893

All data and analysis code are available at https://doi.org/10.17863/CAM.121175. This research was funded by the Wellcome Trust (Grant 106926 to awarded Paul M. Bays). The authors thank all researchers who made their data publicly available.

Ivan Tomić played a lead role in data curation, visualization, and writing–original draft and an equal role in conceptualization, formal analysis, investigation, methodology, software, and writing–review and editing. Sebastian Schneegans played a supporting role in conceptualization, methodology, and writing–review and editing and an equal role in data curation and investigation. Paul M. Bays played a lead role in funding acquisition, resources, supervision, validation, and writing–review and editing, a supporting role in formal analysis, visualization, and writing–original draft, and an equal role in conceptualization and methodology.

Correspondence concerning this article should be addressed to Paul M. Bays, Department of Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, United Kingdom. Email: pmb20@cam.ac.uk

accounts attributing them to random responding, variability in encoding precision, or psychological scaling (Fougnie et al., 2012; Schurgin et al., 2020; van den Berg et al., 2012; Zhang & Luck, 2008). Several theoretical frameworks inspired by principles of neural population coding have been proposed to explain these patterns, drawing on complementary principles such as stochastic sampling (Schneegans et al., 2020), representational geometry (X.-X. Wei & Woodford, 2025), and object confusability (Schurgin et al., 2020). In the population coding framework (Bays, 2014; Schneegans et al., 2020), the inherent stochasticity in neural firing causes decoded population signals to rarely reflect the exact stimulus, producing errors that vary in magnitude. In this account, the spiking activity associated with each memory item is constrained by normalization (Carandini & Heeger, 2012) such that the total activity is fixed, implementing a form of limited memory resource.

Nontarget Responses in Working Memory

In addition to inherent variability in recall of target items, a subset of erroneous responses in multi-item analogue report tasks are typically clustered around the feature values of nontarget items, that is, memory items other than the one cued for report (Bays et al., 2009; Oberauer & Lin, 2017; Schneegans & Bays, 2016). Such responses are commonly referred to as "swap" errors, reflecting the substitution, or swapping in, of a nontarget feature in place of the target feature. These swap errors are uniformly distributed with respect to targets and so have sometimes been mistaken for random responses or confounded with low-precision representations of the target item. Empirical studies have shown that these swap errors occur more frequently at higher set sizes, after longer delays, and when the confusability of items' cue-dimension features at encoding is high (Bays et al., 2009; Emrich & Ferber, 2012; Rerko et al., 2014; Tomić & Bays, 2024a). These observations are consistent with the hypothesis that swap errors arise from variability in the recall of cue-dimension features, causing an error in the cue-to-item matching process that leads to reporting of an uncued item (McMaster et al., 2022; Schneegans & Bays, 2017).

Because swap errors are random relative to the target, failing to differentiate them from other trials will lead to inflated estimates of memory variability. To prevent this, contamination models can be applied that probabilistically identify reports based on nontargets and pool their errors with those of target-based responses when estimating variability (Taylor & Bays, 2018; Taylor et al., 2023; Tomić et al., 2024).

Effects of Retention Interval on Response Variability

A key challenge for working memory (WM) lies in maintaining information over time. As the time since encoding progresses, internal representations deteriorate, eventually becoming indistinguishable from noise. Psychophysical tasks investigate this by varying the retention interval during which stimuli must be held in memory before they can be reported. Multiple studies using this paradigm have consistently found that extending the retention interval impairs recall performance, with effects emerging shortly after stimuli disappear (Tomić & Bays, 2024a) and accuracy continuing to degrade with further delay, albeit at a slower rate (e.g., Barrouillet et al., 2012; McKeown & Mercer, 2012; Phillips & Baddeley, 1971; Rademaker et al., 2018). With the growing use of analogue report tasks, this temporal degradation has been demonstrated across multiple visual features, including orientation (Shin et al., 2017), angular location (Schneegans & Bays, 2018), color (Schurgin et al., 2020), motion (Zokaei et al., 2011), artificial shapes (Zhang & Luck, 2009), and faces (Krill et al., 2018).

However, the mechanisms by which time affects the degradation of representations have not yet been clearly delineated, and the observed decline in precision is consistent with several competing explanations. Early proposals suggested that memory traces passively decay with the mere passage of time (Ricker et al., 2014), potentially due to a shift in attention away from memoranda and an inability to actively refresh the relevant traces (Barrouillet & Camos, 2012). Alternatively, interference accounts argue that multiple representations overlap and interfere with each other over time, impairing the ability to recall the target information (Souza & Oberauer, 2015). In contrast to the notion of continuous deterioration, another theory holds that representations may suffer a "sudden death," abruptly disappearing from memory without affecting other representations (Zhang & Luck, 2009).

Recent advances have converged on a mechanism of diffusion, as implemented in attractor networks, as a prominent explanation for memory deterioration. In this account, stimulus information is maintained as a sustained pattern of activity within a population of interconnected neurons, whose tuning functions cover the range of possible feature values. This activity can be visualized as a "bump" centered on the stored value. Over time, random noise pushes this activity pattern in different directions, resulting in a random walk or diffusion through the feature space. These stochastic changes in the encoded feature value correspond to a gradual loss of information (Burak & Fiete, 2012). Critically, the diffusion can occur without a decay in signal strength. Diffusion is typically implemented in these models as Wiener diffusion, although other types of diffusion such as Ornstein–Uhlenbeck diffusion can also be used. A growing body of empirical support, both at the behavioral (Panichello et al., 2019; Schneegans & Bays, 2018; Tomić & Bays, 2024a) and neural levels (Lim et al., 2019; Wimmer et al., 2014; Wolff et al., 2020), supports diffusion as a key mechanism of memory deterioration. Since the impact of accumulated noise depends on the strength of the underlying signal, one prediction is that representations will degrade more rapidly when more items share a fixed pool of mnemonic resources. However, existing evidence has not definitively confirmed or refuted this prediction.

Stimulus-Specific Biases in Recall

Most investigations of human VWM have focused on the distribution of response errors aggregated across many trials with different target features. Typically, summary statistics from these distributions—such as circular standard deviation—are used to quantify the representational fidelity of stored objects. This approach implicitly assumes that perceptual and memory fidelity are uniform across all stimulus values. However, human visual processing is known to exhibit systematic errors that vary across feature dimensions. These "stimulus-specific biases" have been observed across a range of visual features, with the strongest evidence available for color (Bae et al., 2014; Hardman et al., 2017), where recall is commonly biased toward canonical colors, and orientation (Pratte et al., 2017), where recall is typically biased away from the cardinal axes.

Recently, a class of models based on discrete attractor dynamics has sought to explain both response variability and stimulus-specific biases in WM (Eissa & Kilpatrick, 2023; Gu et al., 2025; Panichello et al., 2019; J. Yang et al., 2024; Ye et al., 2025). In these models, discrete attractors refer to a small number of stable neural population states, corresponding to specific stimulus features. While neural noise causes stimulus representations to diffuse, attractor dynamics pull them toward these stable states, simultaneously counteracting random deviations (i.e., variability) and introducing systematic biases (Panichello et al., 2019). As the accumulation of noise in the network increases—for example, with longer retention interval or higher set size—the influence of attractors is predicted to strengthen, increasing the bias toward the discrete states. Predictions of these discrete attractor models can be contrasted with those of continuous attractor models, in which noise primarily leads to increases in unsystematic error without concomitant changes in bias (see the Discussion section for a detailed comparison). However, empirical tests of these predictions are limited, leaving open questions about the extent to which WM constraints influence response variability and stimulus-specific biases.

The Present Study

In this study, we addressed this gap by systematically investigating how variations in the number of memory items and the duration of the retention interval jointly affect unsystematic error variability and stimulus-specific biases. By examining the interaction between these factors, we aimed to provide new insights into the mechanisms underlying memory fidelity and elucidate the roles that intrinsic memory limitations play in shaping human VWM recall.

Method

We analyzed six previously unpublished experiments conducted in our lab along with data from five published studies (Panichello et al., 2019; Pertzov et al., 2017; Schneegans & Bays, 2018; Schurgin et al., 2020; Shin et al., 2017), comprising a total of 13 experiments, 565 observers, and more than 123,000 trials. All experiments used an analogue report task with the following general protocol: Observers were asked to memorize a study array containing a variable number of visual stimuli and, following a variable delay, were then asked to identify the target feature of the item indicated by a cue. Details of all studies are shown in Table 1. Below, we describe the methodology of the experiments collected in our lab.

New Experiments

Participants

A total of 415 naive observers (263 females, 149 males, three preferred not to say; aged 18–35) took part in the study after giving informed consent in accordance with the Declaration of Helsinki. All observers reported normal or corrected-to-normal visual acuity and normal color vision and were remunerated for their participation (£10 per hour in the lab experiment, £5 per hour in the online experiments). Observers for the experiments conducted online were recruited using Prolific (https://www.prolific.com). In total, 101 observers participated in Experiment 1, 10 in Experiment 2, 10 in Experiment 3, 100 in Experiment 4, 97 in Experiment 5, and 97 in Experiment 6. The sample sizes for the lab and online experiments were based on typical sample sizes used in studies involving analogous report tasks (e.g., Bays, 2014; van den Berg et al., 2012; Zhang & Luck, 2008; Tomić & Bays, 2024b). Procedures were approved by the University of Cambridge Psychology Research Ethics Committee.

Orientation Experiments

The first three experiments (i.e., Experiment 1, Experiment 2, and Experiment 3) used orientation as a memory feature. All experiments used the same set sizes (1 and 3) and retention intervals (1 and 7 s), but differed in two key ways: Experiment 1 used color as a cue feature and was conducted online (Figure 1A), whereas Experiment 2 and Experiment 3 used location as a cue feature and were conducted in a laboratory (Figure 1B).

Experiment 1

The task was presented via internet browsers on observers' personal computers and was coded in JavaScript and HTML Canvas. At the beginning of each trial, a dark-gray fixation circle

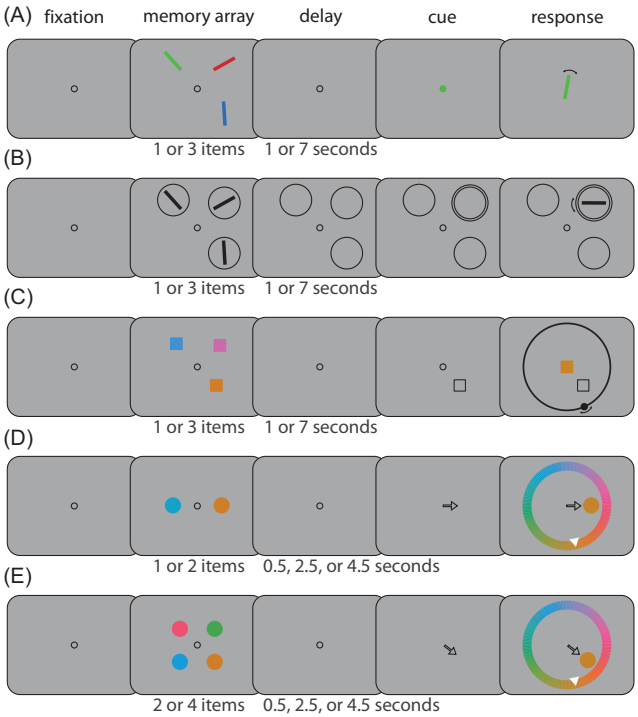
Table 1
Methodological Details of the Analyzed Experiments

No.	Study	Stimulus	Set size	Delay (seconds)	Participant	Trial[a]
1	Experiment 1	Bar	1, 3	1, 7	101	9,224
2	Experiment 2	Bar	1, 3	1, 7	10	7,000
3	Experiment 3	Bar	1, 3	1, 7	10	6,974
4	Shin et al. (2017) Experiment 1	Gabor	1, 2, 4, 6	1, 2, 3, 6	5	9,600
5	Shin et al. (2017) Experiment 2	Gabor	1, 2, 4, 6	1, 2, 3, 6	6	5,760
6	Pertzov et al. (2017)	Bar	1, 2, 4, 6	1, 2, 3[b]	10	8,035
7	Experiment 4	Color	1, 3	1, 7	100	9,948
8	Experiment 5	Color	1, 2	0.5, 2.5, 4.5	97	13,967
9	Experiment 6	Color	2, 4	0.5, 2.5, 4.5	97	13,966
10	Schurgin et al. (2020)	Color	1, 3, 6	1, 3, 5	19	17,100
11	Panichello et al. (2019) Experiment 1	Color	1, 3	1, 7	60	12,000
12	Panichello et al. (2019) Experiment 2	Color	1, 3	1, 7	30	5,437
13	Schneegans and Bays (2018)	Location	1, 2, 4	0.5, 1, 2, 4	20	4,668

[a] The total number of analyzed trials. [b] We did not analyze the 0.1-s delay condition in this experiment.

Figure 1

Task Illustration for the Newly Conducted Experiments



Note. Each trial began with the presentation of a fixation marker, followed by a memory array consisting of one or more objects (oriented bars or colored squares or discs). After a variable delay, a cue (color or location) indicated which of the previously shown objects observers had to recall. Observers then reported the cued object using a continuous response scale. (A) Experiment 1. (B) Experiment 2 and Experiment 3. (C) Experiment 4. (D) Experiment 5. (E) Experiment 6. Confidence reports from Experiment 1, Experiment 3, and Experiment 4 are omitted. Displays are not to scale. See the online article for the color version of this figure.

(radius 5 pixels) was presented at the screen center against the mid-gray background. Observers were instructed to look at the fixation circle throughout the trial (Figure 1A). After 750 ms, one or three randomly oriented bars (width 10 pixels, height 100 pixels) were presented for 500 ms, positioned on an imaginary circle (radius 175 pixels) centered on the fixation point. Bars' locations were randomly chosen for every trial while imposing a minimum distance of 1 radian between neighboring bars. Each bar was colored in one of three highly distinguishable colors: red (255,0,0), green (0,255,0), and blue (0,0,255). After the memory array disappeared, the central fixation was displayed for 1 or 7 s. This was followed by a centrally presented colored disc (radius 5 pixels), which indicated the orientation of which colored bar observers were required to report. Observers were instructed to move their mouse when they were ready to make a response. Once the mouse movement was detected, the adjustable colored bar was displayed centrally, and observers were required to adjust its orientation with the mouse. The response was finalized by pressing the spacebar. Finally, observers were presented with the horizontal slider line on the screen marked with 0% at one end and 100% at the other. They were asked to adjust the slider to best represent their confidence in the recalled orientation on that trial. We did not analyze observers' confidence reports here. Finally, they concluded the trial by clicking on the "Next trial" button. All observers performed six practice trials and 23 experimental trials per condition (combination of set size and delay) for a total of 98 trials. Trials with different set sizes and delays were randomly interleaved. Halfway through the experiment, observers had a 1-min break. Before analyzing the data, we excluded all trials in which observers confirmed their response without adjusting the mouse, as well as trials with response times faster than 200 ms (0.73% of all trials in this experiment).

Experiments 2 and 3

Experiments 2 and 3 were identical except for the distribution from which item orientations were sampled: In Experiment 2, orientations were drawn from a uniform distribution, whereas in Experiment 3, orientations of all items were sampled from a bimodal distribution centered on the cardinal axes, mirroring the orientation statistics of natural scenes. The task and all stimuli were presented on a 69-cm γ-corrected LCD monitor (resolution 2,560 × 1,440) with a refresh rate of 144 Hz. Observers were seated in a dark room and viewed the monitor at a distance of 60 cm, with their heads stabilized by a forehead and chin rest. Stimulus presentation and response registration were controlled using MATLAB (The MathWorks, Inc.) with the Psychtoolbox extensions (Brainard, 1997; Pelli, 1997). Gaze direction was monitored online at 1,000 Hz using an infrared eye tracker (SR Research). Observers were required to fixate the central fixation point from the beginning of the trial until one stimulus was cued for reproduction. Any trial on which gaze deviated by more than 2° from the fixation point, prior to the cue, was aborted and restarted with a new set of stimuli.

Each trial began with the presentation of a central annulus (radius 0.25°, thickness 0.1°), which served as a fixation point (Figure 1B). Once a stable fixation was registered, the thickness of the annulus changed (0.05°), indicating that the memory array would appear in 500 ms. The memory array was then presented for a duration of 500 ms. The array consisted of one or three oriented black bars (length 2.8°), presented along with a placeholder circle (radius 1.5°). Each stimulus was positioned in one of six predetermined locations equally distributed around the imaginary circle with a radius of 5° around the center of the screen.

Memory array presentation was followed by a retention interval (1 or 7 s) during which the fixation circle and placeholders remained visible on the screen. One of the items was then randomly selected for recall, indicated by a larger circle drawn around one of the placeholders. Observers initiated their response by rotating a physical response dial (Griffin Technology PowerMate USB). Once movement of the dial was detected, a black bar with random initial orientation appeared within the placeholder at the cued location. Observers adjusted the orientation of the bar using the dial to match their memory of the target item. To confirm their response, observers pressed the dial. In Experiment 3, observers were additionally asked to indicate their confidence in the recalled orientation by drawing a symmetric arc centered on their reported orientation angle.

Set size and delay duration varied across trials and were randomly intermixed. Every observer completed a total of 700 trials, distributed across two sessions, with each session comprising seven equal blocks of 50 trials. There was a minimum 1-day gap between the two sessions, and each session lasted approximately 1 hr. Prior to

each session, observers familiarized themselves with the task and experimental setup by completing up to 50 practice trials. In Experiment 3, 26 trials were lost for one subject because of a software malfunction (0.37% of all trials). A comparison of Experiments 2 and 3 is provided in the Supplemental Material.

Color Experiments

The next three experiments (i.e., Experiment 4, Experiment 5, and Experiment 6) used color as a memory feature.

Experiment 4

The experiment and procedures for Experiment 4 were similar to those for Experiment 1, with the following exceptions. The memory array consisted of either one or three colored squares (size 60 × 60 pixels; Figure 1C). On each trial, colors were chosen at random from a color wheel defined in Commission internationale de l'éclairage Lab space ($L = 50$, $a = 20$, $b = 20$, $r = 60$), and converted to RGB for presentation. Following a delay of either 1 or 7 s, an empty square was shown at one of the previously occupied locations. Observers were tasked with recalling the color that had been previously presented at that location. Once observers started moving the mouse, a central square and a black central ring (radius 250 pixels), along with the black dot (radius 10 pixels) randomly located on the ring, were presented on the screen. Observers could adjust the position of the black dot on the ring by moving the mouse. Since each position on the ring corresponded to a color, moving the black dot around the ring resulted in a corresponding change in the color of the central square. On each trial, the entire color-angle mapping on the response ring was rotated by a random offset, preserving the relative order of colors while varying their absolute positions, matching the rotation of the color wheel typically used in similar studies. Observers were instructed to match the hue of the remembered square to the hue of the central square as closely as possible. They confirmed their response by pressing the spacebar and then concluded each trial by indicating their confidence level. Observers completed six practice trials and a total of 100 experimental trials. Trials with different set sizes and delays were randomly interleaved. Halfway through the task, observers were given a 1-min break. Before analyzing the data, we excluded all trials in which observers confirmed their response without adjusting the mouse, as well as trials with response times faster than 200 ms (0.52% of all trials in this experiment).

Experiments 5 and 6

Experiments 5 and 6 were identical except for set sizes (Figure 1D and 1E). In both experiments, each trial started with the presentation of a central fixation circle (radius 6 pixels) against a light-gray background for 500 ms. Participants were instructed to look at the fixation circle throughout the trial. This was followed by the presentation of the memory array (500 ms). For Experiment 5, the memory array consisted of one or two colored discs (radius 50 pixels). Each disc was positioned in one of two possible locations, 150 pixels left or right from the screen center. For Experiment 6, the memory array consisted of two or four colored discs positioned in one of four possible locations arranged in a square centered on the screen center, with each location being 150 pixels from the screen center. When only two colored discs were presented in this experiment, they were always displayed on the same side of the screen, arranged vertically either left or right from the screen center.

On each trial, colors were chosen pseudorandomly from a color wheel defined in Commission internationale de l'éclairage Lab space ($L = 50$, $a = 20$, $b = 20$, $r = 60$), with a minimum distance of 90° (Experiment 5) or 45° (Experiment 6), and converted to RGB for presentation. Following a delay of 500 ms, 2,500 ms, or 4,500 ms, one item was cued with a centrally presented arrow (length 60 pixels) pointing toward the location of the previously presented disc. Observers were instructed to start moving the mouse once they were ready to respond. Following the detection of mouse movement, a randomly rotated color wheel (radius 300 pixels, thickness 50 pixels) and a disc at the cued location were presented. As observers moved their mouse pointer around the color wheel, a white triangle slid over the wheel, indicating the pointer's position. Simultaneously, the disc at the cued location changed color accordingly with every mouse movement. Observers were instructed to match the hue of the remembered disc to the hue of the presented disc. They confirmed their response by pressing the spacebar, after which feedback was presented. Feedback consisted of the color wheel, the white triangle indicating the chosen color, and a black triangle indicating the target color. To initiate the next trial, observers clicked on the central button. Trials with different set sizes and delays were randomly interleaved. Observers completed six practice trials and a total of 144 experimental trials. Halfway through the task, observers had a 1-min break. All sizes in pixels are reported for a 1920 × 1,080 resolution. When a different resolution was detected, all sizes in pixels were automatically adjusted to maintain consistency in stimuli presentation across different display settings. In Experiments 5 and 6, a total of three trials were lost due to a software malfunction.

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All data and analysis code are available at https://doi.org/10.17863/CAM.121175. Data were analyzed using MATLAB R2022b (The MathWorks, Inc.) and JASP (JASP Team, 2022). This study's design and its analysis were not preregistered.

Analysis

In all experiments, an array of stimuli with feature values defined in a circular space (e.g., oriented angles, angular locations, or hue angles) was presented on each trial, and observers responded with an estimate of the target feature corresponding to the item identified by a cue. We measured reproduction error in each condition as the angular difference between the reported and target features. For all newly collected experiments, we confirmed that no observers showed random responding. We did this by quantifying performance as mean absolute deviation and comparing it to chance level ($\pi/2$ radians). Only at higher set sizes and longer delays did error distributions become substantially noisier; however, even in these conditions, observers' performance remained reliably better than chance.

Inferential Statistics

We used Bayesian statistics to compare differences in performance across conditions, implemented in JASP (JASP Team, 2022)

using the default Jeffreys–Zellner–Siow prior on effect sizes (Liang et al., 2008). The Bayes factor compares the predictive adequacy of two competing hypotheses (e.g., alternative and null) and quantifies the change in belief that the data bring about for the hypotheses under consideration (Wagenmakers et al., 2018). For example, $BF_{10} = 5$ indicates that the data are five times more likely to occur under the alternative than the null hypothesis. Evidence for the null hypothesis is indicated by $BF_{10} < 1$, in which case the strength of evidence is indicated by $1/BF_{10}$. Alongside Bayes factors for *t* tests, we report the corresponding 95% credible interval for the mean difference. When reporting results from Bayesian analyses of variance (ANOVAs), we report the overall evidence for an effect (BF_{incl}). This is derived via Bayesian model averaging (Hinne et al., 2020), which averages evidence for an effect over all candidate models that contain the effect of interest.

To obtain full posterior distributions for the parameter of interest, we used Gibbs sampling with the Markov Chain Monte Carlo algorithm implemented in JAGS (Plummer, 2003), with weakly informative uniform priors specified for all model parameters. Three parallel chains were run, with the first 500 samples from each chain discarded as burn-in. This yielded 100,000 post-burn-in samples per chain, resulting in 300,000 samples in total for estimating the posterior distributions.

When reporting mean values of parameters in figures, we report a 95% credible interval assuming a uniform prior and normal posterior. A 95% credible interval is a range of parameter values that contains 95% of the posterior distribution. Given the observed data and a uniform prior, this interval represents the range of values within which the true mean has a 95% probability of falling.

Contamination Model

We fit a probabilistic contamination model to the data (also known as a three-component mixture model; Bays et al., 2009), assuming the distribution of responses is a mixture of three components: reports of the target item indicated by the cue, erroneous reports of an item other than the cued item (i.e., swap errors), and lapses (i.e., responses unrelated to the presented stimuli). Reports of target and nontarget items are distributed as von Mises centered on the item feature with concentration κ_θ. The model is described by the mixture probability density given by:

$$p(\hat{\theta}) = \alpha\phi_{VM}(\hat{\theta};\theta,\kappa_\theta) + \frac{\lambda}{M}\sum_{i=1}^{M}\phi_{VM}(\hat{\theta};\varphi_i,\kappa_\theta) + \frac{1-\alpha-\lambda}{2\pi}, \quad (1)$$

where $\hat{\theta}$ is the response value, θ is the target value, and $\{\varphi_1, \ldots \varphi_M\}$ are the M nontarget values. The probability of reporting the target item is given by α, and the probability of mistakenly reporting one of the remaining, noncued items is given by λ. The model assumes responses on the remaining proportion of trials $(1 - \alpha - \lambda)$ have a circular uniform distribution.

While reports of nontarget items reflect recall of memory content, error for these responses must be calculated relative to the corresponding nontarget feature rather than the target, or estimates of error variance will be inflated. Therefore, we estimated overall recall variability as the variance (squared circular standard deviation) of the probability density given by:

$$p(\psi) = (\alpha + \lambda)\phi_{VM}(\psi;0,\kappa_\theta) + \frac{1-\alpha-\lambda}{2\pi}. \quad (2)$$

This distribution combines target and swap responses and assumes that only lapses are uniformly distributed relative to the target. We report the estimated proportions of swap errors (λ) for each analyzed data set in Supplemental Table S1. We define variance as squared circular standard deviation because this definition is appropriate for circular variables yet retains the additive property associated with variance in Euclidean space. The contamination model is used here purely as a statistical tool to reduce noise in the data by identifying swap errors. While the mechanisms underlying swap errors are well-documented (McMaster et al., 2022; Schneegans & Bays, 2017), we do not interpret the remaining model components as necessarily reflecting distinct memory states or psychological processes; this theoretical question is beyond the scope of the current work (but see discussion in Bays et al., 2024).

Finally, we applied similar reasoning for the analysis of stimulus-specific biases, but here we removed individual trials estimated to be most likely swap errors. Specifically, using the same contamination model, we classified each trial based on its posterior probability into those arising from each of the three mixture components (target recall, swaps, lapses) and excluded from analysis only those trials for which the highest posterior probability was for the swap classification, that is, the nontarget component of the model. Because biases vary over the stimulus space, including swap errors in this analysis would pollute estimates of the bias associated with the target value; excluding them reduces overall variability, thereby facilitating more precise estimation of the bias function. One participant in Schneegans and Bays (2018) and one participant in Experiment 2 each had only one nonswap trial in the conditions with the highest set size and longest delay and were therefore excluded from the bias analysis. A MATLAB toolbox implementing the mixture model is available to download from https://bayslab.org/toolbox.

Quantifying the Rate of Forgetting With Time

For each subject in each experiment, we fit a linear regression model of variance against delay interval, separately for each set size:

$$\sigma^2 = \beta_0 + \beta_{delay} \times \text{delay}. \quad (3)$$

Consistent with previous literature, we refer to the estimated slopes (β_{delay}) of the fitted regression lines as "forgetting slopes" (Pertzov et al., 2017; Shin et al., 2017). The β_{delay} parameter quantifies the rate of increase in recall variance over time, expressed in units of variance per second.

Using the obtained forgetting slopes, we next performed another linear regression of the forgetting slopes against set sizes:

$$\beta_{delay} = \beta_0 + \beta_{set\ size} \times \text{set size}. \quad (4)$$

Here, $\beta_{set\ size}$ represents the change in the forgetting slope per unit increase in set size, with units of variance per second per item. We test the hypothesis that, if the rate of forgetting changes with set size, $\beta_{set\ size}$ is larger than zero (i.e., one-tailed test).

In addition, we applied two conventional methods to quantify the joint effect of set size and delay on recall variance. Specifically, we performed a two-way repeated-measures ANOVA with set size,

delay, and their interaction as factors. We also performed a mixed-effects linear regression using set size, delay, and their interaction to predict recall variance:

$$\sigma^2 = \beta_0 + \beta_{\text{delay}} \times \text{delay} + \beta_{\text{set size}} \times \text{set size} + \beta_{\text{int}} \times \text{delay} \times \text{set size}. \quad (5)$$

The interaction term in this regression captures variability in recall variance after accounting for the main effects of delay and set size. Positive values indicate that variance increases more for larger set sizes than for smaller ones. For this analysis, we mean-centered delay and set size and included observers as random effects. All regressions were performed by using the *polyfit* and *fitlm* functions in MATLAB.

While all three methods address the same effect, we present the two-stage slope analysis as the main analysis for didactic reasons: Its stepwise design more clearly illustrates the effects of interest. Specifically, the two-stage model decomposes the interaction effect into forgetting slopes (the change in variance with delay for each set size, rather than the overall effect of delay across all set sizes as in regression) and the effect of set size on these slopes. In addition, a limitation of the mixed-effects linear model is that individual parameters are jointly estimated and therefore dependent (i.e., shrunken toward the fixed-effect estimate), which prevents us from performing a *t* test across observers on these estimates.

Additionally, we performed the same analyses on recall errors after correcting them for observed biases. Calculating the variance of response errors while ignoring the fact that different stimuli are, on average, reported with fixed and systematic error can inflate variance estimates and conflate changes in recall variance and bias across delay and set size. We performed this correction by subtracting the estimated bias (see below for details of our nonparametric bias estimation method) from responses relative to both target and nontarget stimuli. We then calculated recall variance on trial-by-trial bias-corrected responses (analogous to recall variance computed from raw response errors) and submitted these corrected errors to the contamination model, which further accounts for swap errors (analogous to our primary measure of recall variance reported in the Results section). Finally, we performed the same analyses on recall variance estimates computed with and without bias correction.

Nonparametric Method of Bias Estimation

To estimate systematic distortions in memory recalls, we computed the bias—defined as the expected signed error—as a function of target value (Figure 2). In general, this procedure assumes that all observers within an experiment share the same shape of bias function (which is specific to the stimulus feature space), but that the amplitude of the bias varies between observers and experimental conditions. Our approach is broadly based on two steps. First, we estimate the shape of the bias (i.e., the group bias) using all available data from each experiment except trials from a single observer in a particular experimental condition (i.e., held-out trials). Second, we estimate the bias amplitude in those held-out trials, assuming that the shape is described by the group bias.

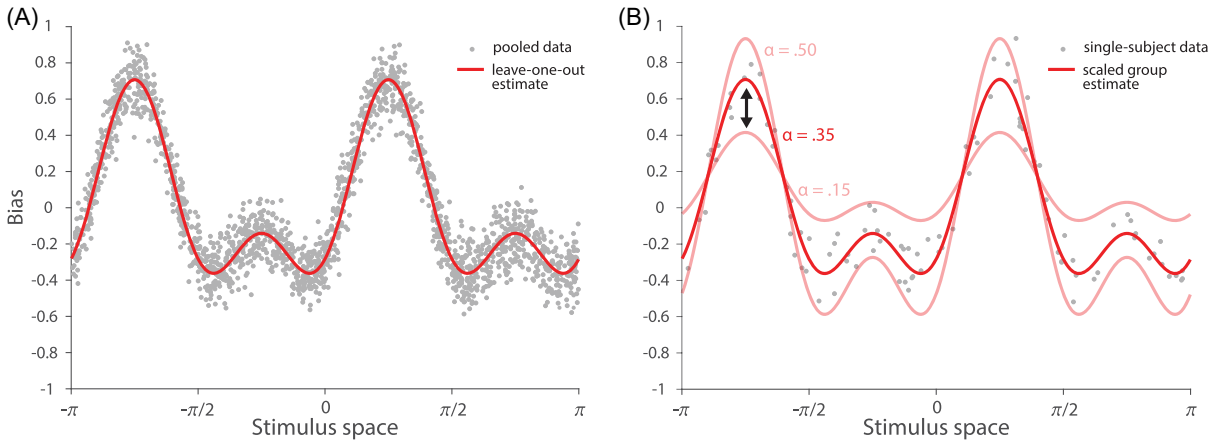
Specifically, given responses $\hat{\theta}$ and target values θ, the bias for any arbitrary stimulus ψ was estimated using a circular variant of the kernel conditional density estimator (Rosenblatt, 1969):

$$\hat{b}(\psi) = \frac{\sum_i (\hat{\theta}_i \ominus \theta_i)\phi_{\text{VM}}(\psi - \theta_i; \kappa)}{\sum_i \phi_{\text{VM}}(\psi - \theta_i; \kappa)}, \quad (6)$$

where $\phi_{\text{VM}}(\psi; \kappa) = \exp(\kappa \cos(\psi))$ denotes the von Mises kernel and $\ominus$ refers to modular (circular) subtraction returning values in the interval $[-\pi, \pi)$. This approach smooths the signed error across nearby stimulus values, providing a continuous estimated bias

Figure 2

Illustration of the Nonparametric Method for Estimating Bias



Note. (A) Simulated recall biases (gray points) plotted as a function of stimulus value, pooled across all simulated subjects. The red curve shows the estimated continuous bias function using a leave-one-out procedure, where the estimate excludes data from the individual subject being evaluated. (B) Simulated data from a single subject (gray points), overlaid with scaled versions of the group-level bias estimate from Panel A. The bias function is shown at different amplitudes (red curves) to illustrate how individual bias strength is assessed via projection onto the normalized group-level bias shape. The data were simulated from an arbitrary bias-generating function to illustrate the method. See the online article for the color version of this figure.

function $\hat{b}(\psi)$. We used a kernel width of $\kappa = 10$ for all reported analyses, but confirmed that our conclusions were robust across a range of smoothing values (e.g., $\kappa = 40$).

Assuming that all data share a common bias shape but may differ in amplitude across observers or experimental conditions, we normalized the estimated bias function to isolate its amplitude. First, we removed any constant offset by demeaning:

$$\bar{b} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{b}(\psi)d\psi, \quad \hat{b}_0(\psi) = \hat{b}(\psi) \ominus \bar{b}. \tag{7}$$

We then normalized the demeaned function, resulting in a function that preserves the shape of the bias pattern while removing differences in overall magnitude:

$$\hat{f}(\psi) = \frac{\hat{b}_0(\psi)}{\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{b}_0(\psi)^2 d\psi}}. \tag{8}$$

To quantify the bias amplitude for a given subject and experimental condition (i.e., a specific combination of set size and delay), denoted as subset c, we projected the observed errors $(\hat{\theta}_i^c \ominus \theta_i^c)$ onto the normalized bias shape $\hat{f}^{-c}(\psi)$ estimated from the remaining data:

$$\hat{\alpha}^c = \frac{1}{n_c}\sum_i (\hat{\theta}_i^c \ominus \theta_i^c)\hat{f}^{-c}(\theta_i^c), \tag{9}$$

where $\hat{f}^{-c}(\psi)$ is calculated according to Equations 6–8 in a leave-one-out manner, using all data except that from condition c. This leave-one-out procedure was repeated for each subject and experimental condition, providing cross-validated estimates of bias amplitude that reflect the strength of systematic bias while controlling for overfitting and shared structure across conditions. For completeness, in the orientation experiments, we also obtained a parametric estimate of bias amplitude by assuming bias followed a sine function of stimulus orientation, that is, replacing the estimated normalized bias function $\hat{f}(\psi)$ with a fixed function $f(\psi) = \sqrt{2}\sin(2\psi)$.

Because we assumed that the shape of the bias is shared across observers and experimental conditions within each study, we estimated the bias shape using all aggregated trials except those from the condition for which the bias amplitude was being estimated. In this way, data from all other observers and conditions contributed to a reliable estimate of the bias shape, while the bias amplitude was estimated separately for each observer. This approach is different from a typical cross-validation for a predictive model, where model parameters are fitted to one partition of the data and then assessed on the remainder. The present method was optimized for precise estimation of individual bias amplitudes (the variable of primary interest), given there were insufficient data to reliably estimate both bias shape and amplitude for each observer separately.

Results

Effects of Memory Load and Retention Interval on Response Variability

Our primary question here is whether forgetting in human observers depends on the number of items they attempt to memorize over extended periods. To illustrate the raw data patterns, Supplemental Figure S1 shows response error distributions across set sizes and delays for the six newly collected experiments. We begin addressing this question by calculating linear slopes for recall error variance (Figure 3) with respect to delay for every set size separately (Figure 4A). The estimated forgetting slopes for each set size and study, expressed in variance units per second, are displayed in Figure 4B, while the fits of this linear model to each experiment are presented in Supplemental Figure S3. As anticipated, error variance increased with delay, as indicated by the positive average values of forgetting slope estimates. These findings align with a substantial body of research that has consistently demonstrated the degradation of memory representations over time (for a review, see Bays et al., 2024). Qualitatively, the observed pattern of changes in forgetting slopes with set size within each study suggests that the rate of forgetting increases with set size.

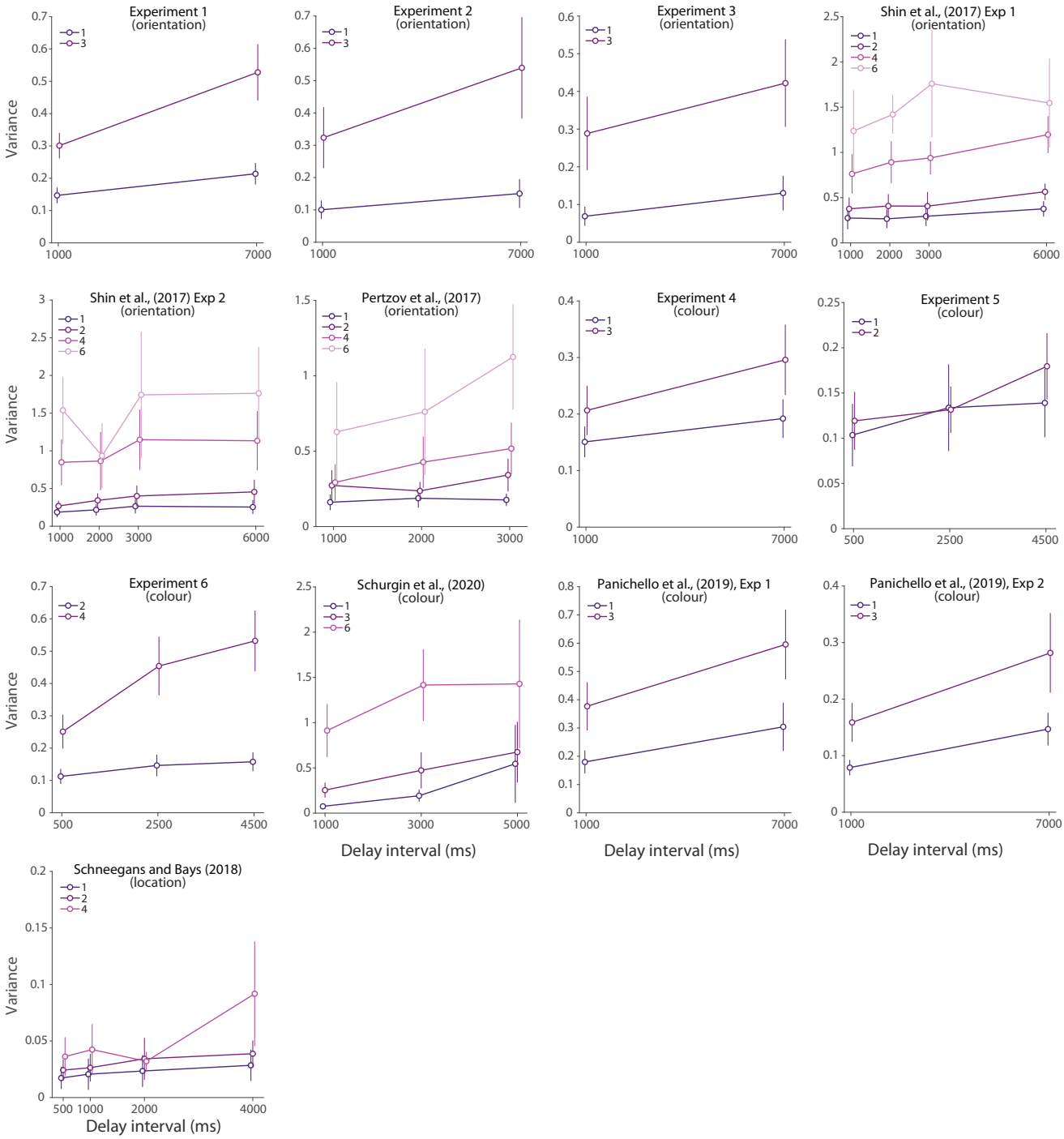
To formally test this hypothesis, we compared forgetting slopes for the lowest and highest set size in each study using a one-tailed Bayesian t test. Only one experiment showed moderate evidence against a difference (Schurgin et al., 2020: $BF_{10} = 0.27$), whereas four experiments provided strong to extreme evidence for forgetting slopes being larger with larger set sizes (Experiment 1: $BF_{10} = 892$; Experiment 2: $BF_{10} = 10.6$; Pertzov et al., 2017: $BF_{10} = 12$; Experiment 6: $BF_{10} = 4.36 \times 10^4$). The remaining experiments showed only anecdotal evidence in favor of alternative hypothesis (Experiment 3: $BF_{10} = 2.13$; Panichello et al., 2019, Experiment 1: $BF_{10} = 1.97$; Panichello et al., 2019, Experiment 2: $BF_{10} = 1.35$; Schneegans & Bays, 2018: $BF_{10} = 2.53$) or null hypothesis (Shin et al., 2017, Experiment 1: $BF_{10} = 0.66$; Shin et al., 2017, Experiment 2: $BF_{10} = 0.99$; Experiment 4: $BF_{10} = 0.57$; Experiment 5: $BF_{10} = 0.47$).

To further quantify this effect, we fit a linear slope to the estimated forgetting slopes with set size (Figures 4C and 5). Estimated slopes, expressed in variance units per second per item, in all experiments exhibited a bias toward positive values, suggesting that set size indeed detrimentally affects forgetting rates (Figure 4D; see Supplemental Figure S4 for full posterior distributions). To test the hypothesis that these slopes are greater than zero, we conducted a one-sample Bayesian t test against zero (one-tailed) for each study. The results of this analysis are shown in Table 2. In summary, in eight analyzed experiments, evidence remained largely weak for either a difference or its absence ($0.27 \leq$ all $BF_{10} \geq 2.1$). However, in the remaining five experiments, we found clear evidence that forgetting slopes increase with set size, specifically as follows: Experiment 1 ($BF_{10} = 892.7$), Experiment 2 ($BF_{10} = 10.51$), Pertzov et al. (2017; $BF_{10} = 19.42$), Experiment 6 ($BF_{10} = 4.3 \times 10^4$), and Schneegans and Bays (2018; $BF_{10} = 5.43$). Three of those studies used orientation as a memory feature, one used color, and one used angular location, suggesting that the effect does not depend on the specific memory feature.

To assess the overall effect of set size on forgetting, we aggregated all individual estimates across studies ($N = 565$) and tested them against zero (indicating no effect of set size, i.e., forgetting slopes do not change with set size). This analysis revealed extremely strong evidence for a difference from zero ($\delta = 0.012$, 95% CI [0.009, 0.015]; $BF_{10} = 5.82 \times 10^9$; posterior median over effect size $Mdn = 0.302$, 95% CI [0.217, 0.386]). To demonstrate that this effect was not solely influenced by the five studies in which we

Figure 3

Recall Variance Across Set Size and Delay

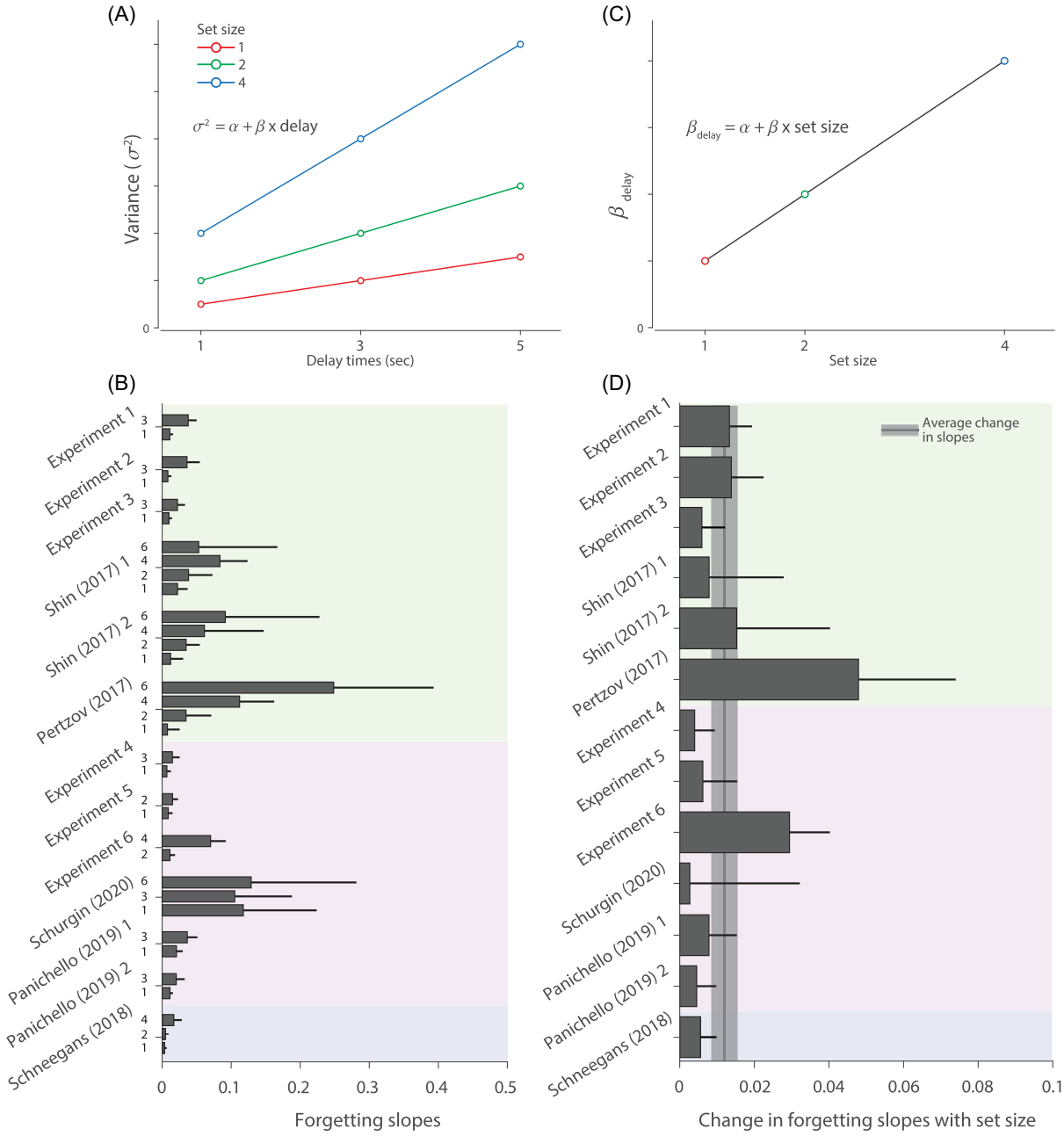


Note. Recall variance (corrected for swap errors, Equation 2) as a function of set size and delay duration, shown separately for each experiment. Each panel corresponds to one experiment. Points and error bars indicate means and 95% credible intervals. To improve visibility, data points for different set sizes are horizontally offset along the *x*-axis. See the online article for the color version of this figure.

observed strong evidence for a difference, we aggregated all estimates while excluding those five studies (N = 327). Although the effect was now expectedly smaller (δ = 0.006, 95% CI [0.002, 0.010]), we again observed moderate evidence for a difference from zero (BF_{10} = 6.5; posterior median over effect size Mdn = 0.156, 95% CI [0.049, 0.264]). Performing the same analyses on recall variance corrected for bias yielded the same conclusions (see the Supplemental Material for detailed analyses).

Figure 4

Schematic Illustration and Empirical Results of the Forgetting Slopes Analysis

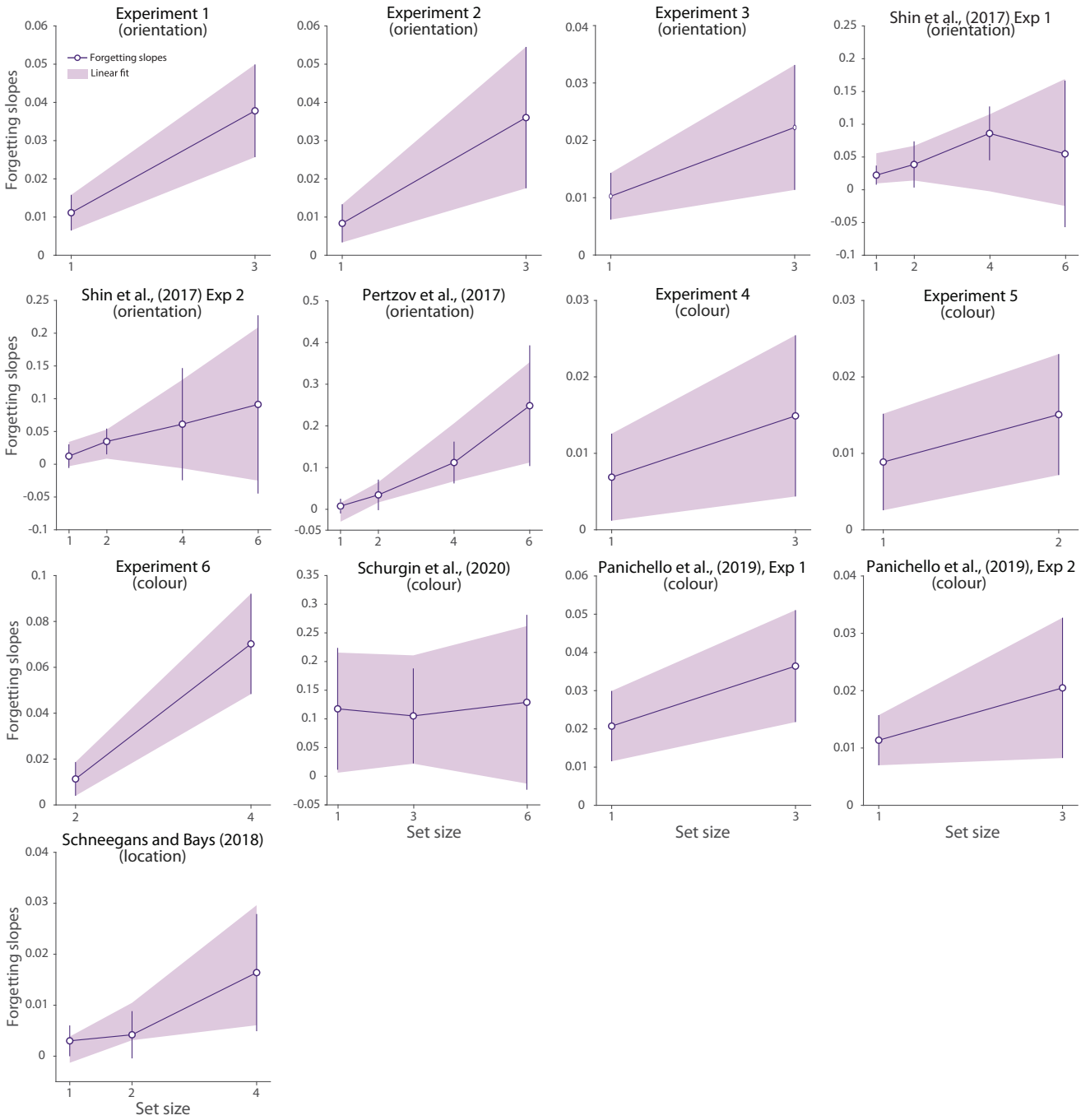


Note. Panels A and C show the schematic illustrations of the analytical approach using simulated data, whereas Panels B and D show the empirical results based on psychophysical data. (A) Illustration of linear fits to recall variance as a function of delay for different set sizes. Slopes are simulated and shown for illustrative purposes only. (B) Estimated forgetting slopes (mean and 95% credible interval) across set sizes and studies, reflecting how rapidly memory degrades over time. (C) Illustration of the second-stage model, in which forgetting slopes from Panel A are regressed against set size. (D) Results of the model described in Panel C. Bars show mean and 95% credible interval, reflecting how the rate of memory deterioration changes as a function of set size. Black vertical line shows the mean change in slope, and the gray shaded region indicates the 95% credible interval. Background colors in Panels B and D denote the memory feature tested (orientation = green, color = purple, location = blue). See the online article for the color version of this figure.

To illustrate the relative contribution of delay, set size, and their interaction on recall variance beyond reporting raw parameter estimates, we generated a model-based visualization of their combined effects (Figure 6). Specifically, we substituted Equation 4 into Equation 3 to express variance directly as a function of delay and set size, such that the slope of the delay–variance relationship (β_{delay})

Figure 5

Forgetting Slopes as a Function of Set Size



Note. Each panel displays data from a single experiment. Points and error bars show mean estimates and 95% credible intervals of forgetting slopes at each set size (shown in Figure 4B). Shaded patches (means and 95% credible intervals) show the fit of the linear model relating forgetting slopes to set size (slope estimates shown in Figure 4D). See the online article for the color version of this figure.

was modeled as a linear function of set size. We used the posterior mean estimates from Equation 4 to compute predicted values of β_{delay} and used these to calculate predicted changes in variance across a grid of delay and set size values. Figure 6 displays the predicted changes in variance derived from this procedure. Because the intercept from Equation 3 varies substantially across studies due to differences in stimuli and baseline memory precision, we focused on relative changes in variance, omitting the intercept. As a result, the absolute values on the *y*-axis (i.e., recall variance) are not directly interpretable, but the relative differences between slopes—reflecting

Table 2
The Effect of Set Size on Forgetting Rates

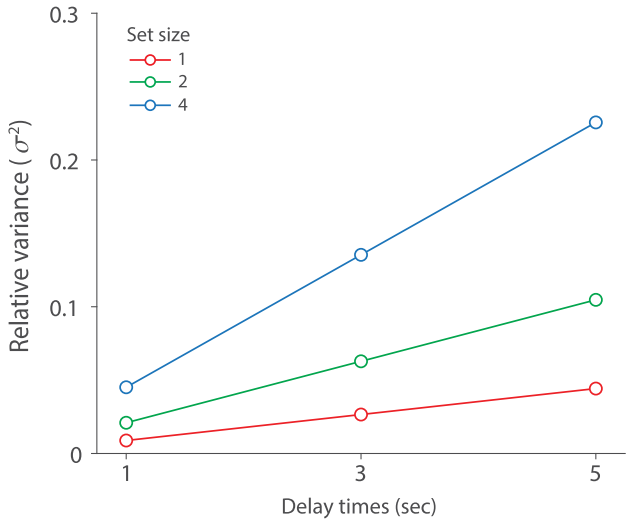
No.	Study	δ	95% CI	BF_{10}
1	Experiment 1	0.013	[0.007, 0.019]	892.7
2	Experiment 2	0.014	[0.004, 0.024]	10.51
3	Experiment 3	0.006	[−0.001, 0.013]	2.1
4	Shin et al. (2017) Experiment 1	0.008	[−0.020, 0.036]	0.74
5	Shin et al. (2017) Experiment 2	0.015	[−0.018, 0.048]	1.05
6	Pertzov et al. (2017)	0.048	[0.018, 0.078]	19.42
7	Experiment 4	0.004	[−0.001, 0.009]	0.57
8	Experiment 5	0.006	[−0.003, 0.016]	0.47
9	Experiment 6	0.029	[0.018, 0.04]	4.3×10^4
10	Schurgin et al. (2020)	0.003	[−0.029, 0.034]	0.27
11	Panichello et al. (2019) Experiment 1	0.008	[0.0002, 0.015]	1.96
12	Panichello et al. (2019) Experiment 2	0.005	[−0.001, 0.01]	1.34
13	Schneegans and Bays (2018)	0.006	[0.0009, 0.010]	5.43

Note. δ = the mean of the posterior distribution over slope; CI = credible interval; BF = Bayes factor.

how variance increases with delay at different set sizes—are informative and allow meaningful comparison of effect magnitudes. This analysis showed that, while the effect of delay alone on variance is relatively modest (as reflected by the shallow slope of the red line in Figure 6), its influence increases markedly with set size—being approximately seven times greater at Set Size 6 than at Set Size 1, which spans the range of set sizes examined in this study.

In conclusion, our extensive analysis of memory recall across multiple studies demonstrates that set size impacts forgetting rates. We observed robust evidence supporting the claim that larger set sizes lead to increased forgetting. These findings clarify the relationship between memory capacity and recall performance, highlighting the adverse effect of memory load on memory retention over extended periods.

Figure 6
Predicted Changes in Relative Recall Variance as a Function of Delay and Set Size



Note. Predictions were computed using posterior mean parameter estimates and represent relative changes in variance with the intercept from Equation 3 omitted, while the intercept from Equation 4 was retained. The slope of the delay–variance relationship (β_{delay}) increases linearly with set size. Different colored lines indicate that the effect of delay on variance is substantially amplified at larger set sizes. See the online article for the color version of this figure.

Mixed-Effects Linear Regression

For completeness, we conducted a more conventional analysis using repeated-measures ANOVA on recall variance and mixed-effects linear regression. The results of the former are presented in the Supplemental Material, while the results of the latter are briefly described here. In both cases, the findings closely align with those of our primary analysis.

We used mixed-effects linear regression to predict recall variance based on delay, set size, and their interaction. Faster forgetting with larger set sizes would be indicated with positive values of estimated interactions. To assess this, we first tested, for each study, whether 95% credible intervals of the estimated random effects of the interaction effect contain zero. Due to the interdependence of these random effects for individual observers within each experiment, we cannot test whether they differ from zero (i.e., no interaction) by using a *t* test.

These results align closely with the analysis of forgetting slope changes with set size in the previous section. Specifically, in seven experiments, we obtained 95% credible intervals that did not contain zero, indicating that the random effects estimates differed from zero. Specifically, we obtained the following results: Experiment 1, δ = 0.013, 95% CI [0.008, 0.019]; Experiment 2, δ = 0.014, 95% CI [0.006, 0.022]; Pertzov et al. (2017), δ = 0.048, 95% CI [0.032, 0.063]; Experiment 5, δ = 0.006, 95% CI [0.004, 0.009]; Experiment 6, δ = 0.029, 95% CI [0.023, 0.036]; Panichello et al. (2019) Experiment 1, δ = 0.008, 95% CI [0.002, 0.014]; and Schneegans and Bays (2018), δ = 0.005, 95% CI [0.001, 0.008].

In the remaining six experiments, 95% CI did include zero, indicating that interaction effects did not differ from zero: Experiment 3, δ = 0.006, 95% CI [−0.001, 0.013]; Shin et al. (2017) Experiment 1, δ = 0.008, 95% CI [−0.008, 0.023]; Shin et al. Experiment 2, δ = 0.015, 95% CI [−0.003, 0.034]; Experiment 4, δ = 0.004, 95% CI [−0.001, 0.009]; Schurgin et al. (2020), δ = 0.003, 95% CI [−0.017, 0.023]; and Panichello et al. (2019) Experiment 2, δ = 0.005, 95% CI [−0.0002, 0.009].

Finally, we aggregated all estimated fixed effects of the interaction (one for each study) and tested them against zero. We observed strong evidence for a difference from zero ($\delta = 0.013$, 95% CI [0.005, 0.020]; $BF_{10} = 24.3$), reaffirming our earlier analysis indicating that forgetting increases with set size.

Effects of Memory Load and Retention Interval on Response Bias

After analyzing overall recall variability, we focused on systematic distortions—that is, biases—in memory recall. We estimated bias as a continuous function of stimulus value using our nonparametric leave-one-out approach (Figure 2). As a first step, we visualized individual responses as a function of target angle for the six newly collected experiments (Figure 7). Departures from the identity line in these plots indicate the presence of systematic response biases in the data. In experiments using orientation (Figure 7A and 7B), these plots revealed a characteristic S-shaped pattern, consistent with previous findings.

To illustrate bias across conditions, we plotted the full (nonparametric) bias estimates for the smallest and largest set sizes as well as for the shortest and longest retention intervals in each study (Figure 8). When orientation was used as the memory feature, the pattern of estimated biases aligned with the widely reported sine-wave function, displaying systematic deviations away from the nearest cardinal axis. In contrast, for studies using color, the shape of the bias function is less interpretable because angular values do not correspond to a single mapping onto hues. Additionally, variations in the specific color wheels used across studies led to differences in the shape of the resulting bias functions, further limiting direct comparisons between experiments.

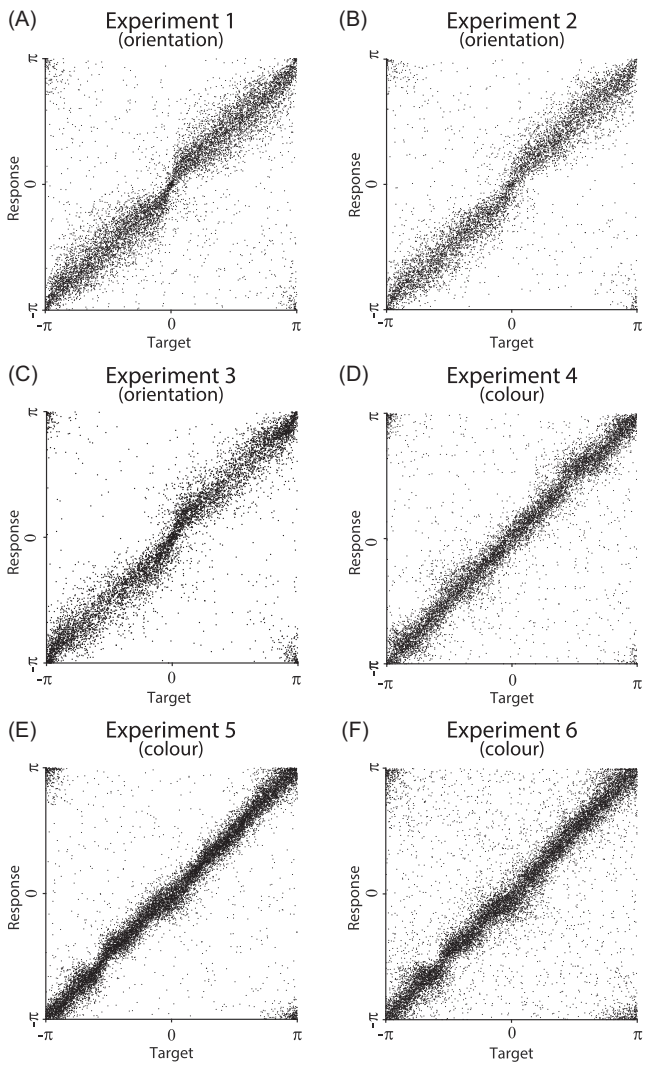
To test how bias magnitude changes with memory load and delay, we analyzed these estimates using repeated-measures ANOVA, examining the effects of set size, delay, and their interaction. Results of this analysis are shown in Table 3 and estimated bias amplitudes in Figure 9. When analyzing the effects of set size on bias amplitude, two studies demonstrated anecdotal evidence for such an effect, specifically: Shin et al. (2017) Experiment 1, $BF_{incl} = 2.2$, and Schurgin et al. (2020), $BF_{incl} = 1.13$. Two experiments exhibited moderate evidence for the effect of set size (Schneegans & Bays, 2018: $BF_{incl} = 4.84$; Panichello et al., 2019, Experiment 2: $BF_{incl} = 3.22$), with one showing a decrease and the other an increase in bias amplitude. Finally, one experiment revealed extreme evidence for the effect of set size (Pertzov et al., 2017: $BF_{incl} = 2657.1$), with bias amplitudes being larger at smaller set sizes (one item vs. six items $BF_{10} = 795$; two items vs. six items $BF_{10} = 61$), but no evidence for a difference between four and six items ($BF_{10} = 0.8$). In the remaining nine experiments, we found evidence ranging from weak to strong ($0.686 \geq$ all $BF_{incl} \geq 0.087$) against the effect of set size on bias amplitudes. Taken together, although some studies show weak and anecdotal support for an effect of set size on bias amplitude, the majority of experiments provide evidence supporting the absence of such an effect.

When examining the effects of delay on bias amplitude, the results were notably consistent across studies. Specifically, all experiments except one (Panichello et al., 2019, Experiment 2: $BF_{incl} = 1.126$) provided evidence against the effect of delay on bias amplitude, indicating that bias amplitude remains stable over retention time. While two studies provided only weak evidence for the absence of effect (Experiment 3: $BF_{incl} = 0.340$; Experiment 5: $BF_{incl} = 0.441$), four studies revealed moderate evidence ($0.243 \geq$ all $BF_{incl} \geq 0.105$), and six studies revealed strong evidence ($0.082 \geq$ all $BF_{incl} \geq 0.025$) for the absence of this effect. This consistency suggests that bias amplitude is largely unaffected by the duration of the retention interval.

Finally, when analyzing an interaction of set size and delay on bias amplitude, we found compelling evidence against the influence of this interaction on bias amplitudes. While one study provided only weak evidence against the interaction (Panichello et al., 2019, Experiment 2: $BF_{incl} = 0.55$), the remaining 12 experiments revealed moderate (Experiment 2: $BF_{incl} = 0.186$; Experiment 3: $BF_{incl} = 0.144$) to strong ($0.049 \geq$ all $BF_{incl} \geq 0.002$) evidence

Figure 7

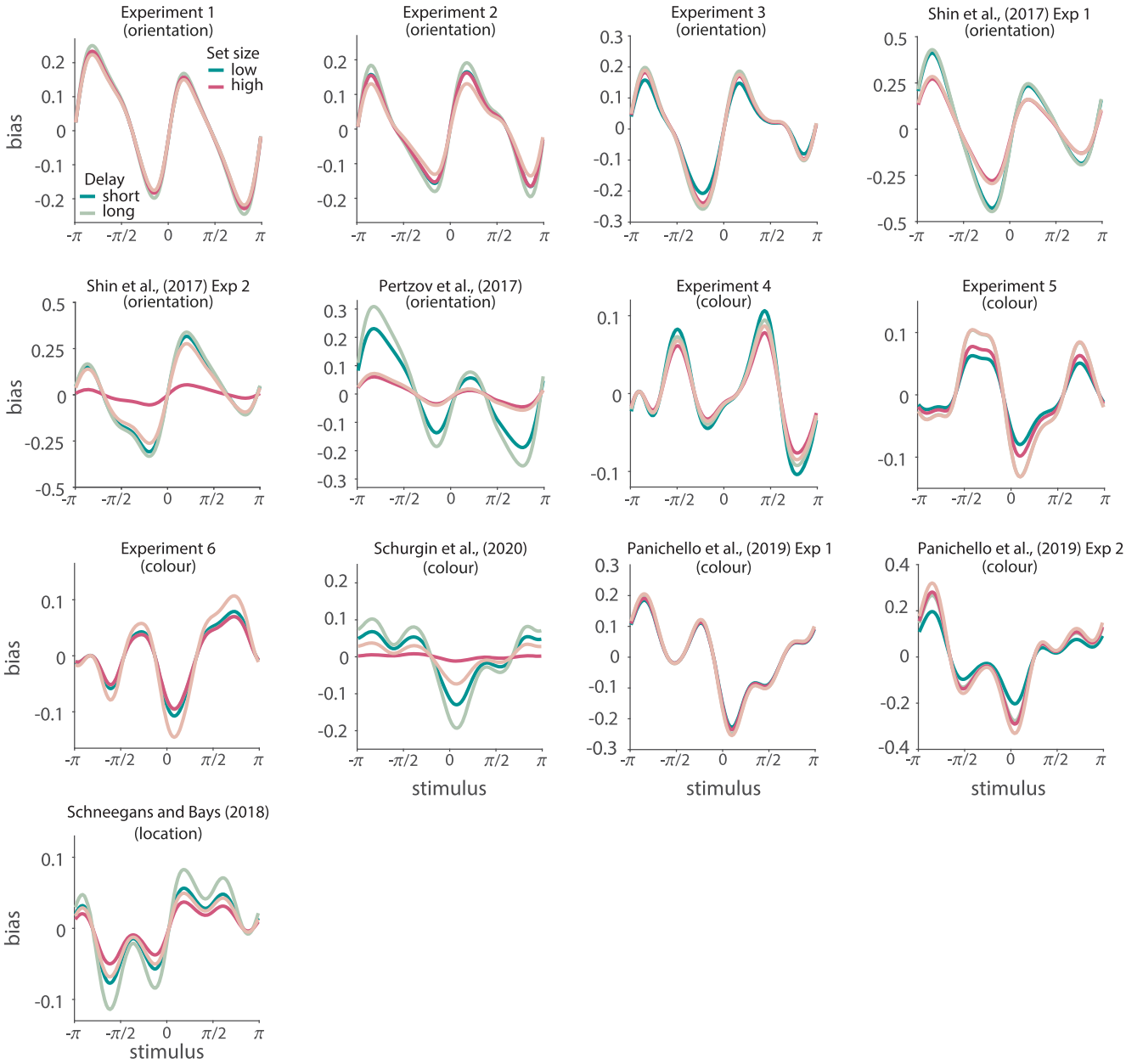
Response Biases Revealed by Plotting Reported Angles Against Target Angles



Note. Scatterplot of reported angles versus target angles. Data are pooled across all experimental conditions and observers. (A) Experiment 1. (B) Experiment 2. (C) Experiment 3. (D) Experiment 4. (E) Experiment 5. (F) Experiment 6.

Figure 8

Bias Functions Estimated Using the Nonparametric Method



Note. Estimated bias functions across stimulus space for each study. Each panel shows four curves representing combinations of set size (low = green, high = red) and delay (short = solid, long = transparent). To facilitate visual comparison, only the combinations of the lowest and highest set sizes and the shortest and longest delays are shown, while the remaining combinations are omitted. See the online article for the color version of this figure.

against the interaction of set size and delay on bias amplitude. Thus, the data consistently indicate no interactive effect of memory load and retention interval on bias amplitude.

Parametric Estimates of Bias in Orientation

For completeness, we also fitted biases from the orientation experiments using a parametric description, approximating the pattern of biases with a sine-wave function. Amplitude estimates from this parametric approach closely matched those obtained with the nonparametric method (Shin et al., 2017, Experiment 2: $r = .63$; all other correlations $r \geq .88$).

The results, summarized in Table 4, were largely consistent with the nonparametric analysis, providing evidence against an effect of retention interval and against an interaction between set size and retention interval. Effects of set size were likewise consistent across

Table 3
Results of Bayesian Repeated-Measures Analysis of Variance for Bias Amplitude

No.	Study	BF_{incl} set size	BF_{incl} delay	BF_{incl} interaction
1	Experiment 1	0.087	0.082	0.011
2	Experiment 2	0.396	0.243	0.186
3	Experiment 3	0.243	0.340	0.144
4	Shin et al. (2017) Experiment 1	2.2	0.066	0.039
5	Shin et al. (2017) Experiment 2	0.686	0.066	0.046
6	Pertzov et al. (2017)	2657.1	0.11	0.043
7	Experiment 4	0.113	0.074	0.01
8	Experiment 5	0.285	0.441	0.049
9	Experiment 6	0.136	0.025	0.002
10	Schurgin et al. (2020)	1.13	0.125	0.019
11	Panichello et al. (2019) Experiment 1	0.10	0.132	0.015
12	Panichello et al. (2019) Experiment 2	3.22	1.126	0.553
13	Schneegans and Bays (2018)	4.84	0.074	0.015

Note. BF = Bayes factor.

most cases, with two exceptions. In Shin et al. (2017) Experiment 1, the parametric analysis indicated extreme evidence for the effect of set size (BF_{incl} = 1,471), with bias amplitude reduced at larger set sizes, specifically between Set Sizes 1 and 6 (BF_{10} = 240), Set Sizes 1 and 4 (BF_{10} = 9.1), and Set Sizes 2 and 6 (BF_{10} = 19.3; other comparisons: $1.9 \geq BF_{10} \geq 0.8$). Similarly, in Experiment 2 of the same study (Shin et al., 2017), we also found extreme evidence for the effect of set size (BF_{incl} = 1,002), with reduced bias amplitudes at larger set sizes: between Set Sizes 1 and 6 (BF_{10} = 68), 2 and 6 (BF_{10} = 58), and 4 and 6 (BF_{10} = 2.75; other comparisons: $0.62 \geq BF_{10} \geq 0.25$). This pattern of decreasing bias amplitude is also evident in Figures 8 and 9. A key difference between the three experiments showing the effect of set size and the three that did not is that the former extended to higher set sizes. In a post hoc analysis, we consistently observed that bias amplitudes differed between the largest set size and the smaller ones, but not among the smaller set sizes, suggesting that stimulus-specific biases change considerably only at relatively high memory loads.

Although widely applied, the sine-wave function only very roughly approximates the true pattern of biases in orientation estimation, with informal observations often pointing to unequal peak-trough magnitudes and asymmetries in slope (i.e., one flank of the wave being steeper than the other), observations supported by our nonparametric estimates of bias shape (the first five panels of Figure 8). Nevertheless, whether bias amplitudes were estimated using parametric or nonparametric methods, the conclusions converged: Biases were largely insensitive to manipulations of set size and retention interval and, critically, showed no sensitivity to their combined effect. This strongly contrasts with response variability, which consistently increased with set size, retention interval, and their interaction.

Discussion

We analyzed data from 13 published and newly conducted experiments to examine the impact of VWM capacity and maintenance limitations on variability and bias in recall. Our findings provide robust evidence that response error variability increases more rapidly over time as the number of items held in memory increases. By contrast, response biases remain unaffected by capacity or maintenance demands. These results indicate that the two components of response error—unsystematic variability and systematic bias—are affected differently by the inherent limitations of VWM.

Set Size and Forgetting Over Time

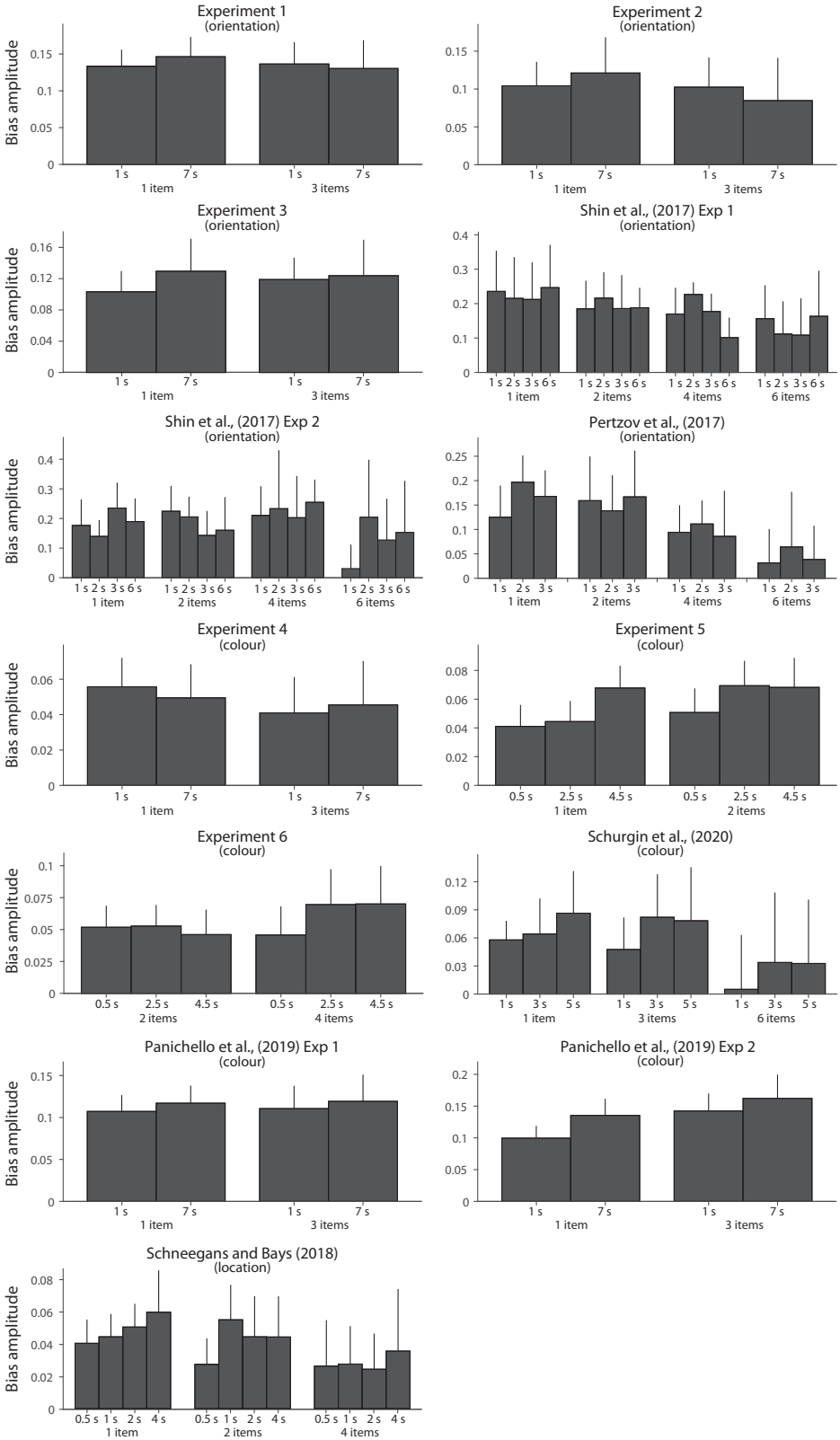
Previous studies have drawn disparate conclusions regarding the effect of set size on forgetting rate, despite employing similar experimental designs. Using an analogue report task with orientation as the report feature, Pertzov et al. (2017) argued that rapid forgetting involves an interaction between retention time and the number of stored items, with larger set sizes leading to faster deterioration of internal representations. A subsequent study (Cohen-Dallal et al., 2018) attributed this effect to competition among multiple objects held in memory during maintenance, rather than to the initial strength of encoded representations, which might also diminish as set size increases. In contrast, Shin et al. (2017), using a similar task, found weaker evidence for an effect of delay on forgetting and failed to find evidence that the rate of forgetting was modulated by set size. Reflecting these divergent findings, a recent attempt to establish robust empirical benchmarks in WM avoided including effects of (unfilled) delay on measures of forgetting (Oberauer et al., 2018).

To address the inconsistent conclusions of previous studies, we compiled a large data set of new and existing analogue report experiments covering a range of different feature dimensions and response modalities, including both online and laboratory-based testing, but which all assessed VWM recall while systematically manipulating set size and retention time. Our analysis provided robust evidence that variability increased with retention time, in addition to the well-documented increase with set size. To further examine the influence of set size on the rate of forgetting, we applied a simple linear model to estimate forgetting slopes (i.e., changes in response variability with delay) as a function of set size. The results demonstrated clear evidence for an increased rate of forgetting with larger set sizes, both within individual experiments and across pooled experiments. These findings address a long-standing debate in the field, demonstrating a detrimental impact of memory load on memory retention over extended delays.

One advantage of our study is that we applied a consistent set of measures and analysis techniques across all component data sets.

Figure 9

Bias Amplitude Across Set Size and Delay



Note. Bias amplitude estimates derived from the nonparametric bias estimation method, shown separately for each delay and set size across individual studies. Bar heights represent mean bias amplitude, and error bars indicate 95% credible intervals.

Table 4
Results of Bayesian Repeated-Measures Analysis of Variance for Parametrically Estimated Bias Amplitude

No.	Study	BF_{incl} set size	BF_{incl} delay	BF_{incl} interaction
1	Experiment 1	0.082	0.075	0.008
2	Experiment 2	0.248	0.251	0.151
3	Experiment 3	0.243	0.338	0.138
4	Shin et al. (2017) Experiment 1	1,471	0.071	0.098
5	Shin et al. (2017) Experiment 2	1,002	0.115	0.298
6	Pertzov et al. (2017)	677	0.067	0.032

Note. BF = Bayes factor.

In particular, previous studies have based their conclusions about variability on different measures of dispersion for circular variables. The interpretation of interactions like the one between set size and retention interval relies on having a meaningful measurement scale. We argue that squared circular standard deviation is the correct measure in this context, because it retains the important property of additivity within a circular space. Only using this definition does random noise added to a stored feature value at a constant rate correspond to a linear increase in variance with time. This justifies our key conclusion—that the rate at which noise accumulates during maintenance increases with the number of memory items.

Performance in the types of tasks analyzed here is known to vary both between individuals and across experimental setup, often complicating comparisons across experiments. Specifically, in WM tasks, performance depends on the stimulus feature being used (e.g., orientation vs. angular location; Tomić & Bays, 2024b) and, in some cases, even within a feature, depending on the exact type of visual object (e.g., oriented Gabors vs. bars; Alvarez & Cavanagh, 2008) or on secondary characteristics of the stimulus (e.g., the contrast of oriented Gabors; Tomić & Bays, 2018). In addition, performance varies with presentation time (Bays et al., 2011; Sewell et al., 2014), stimulus eccentricity from fixation (Staugaard et al., 2016), and cue type (e.g., color vs. location; Schneegans & Bays, 2017). The experiments presented here differed in several of these aspects, leading to variation in performance between superficially similar experiments (e.g., compare Shin et al., 2017 and Pertzov et al., 2017). Importantly, the statistical tools applied consistently across all studies focus on relative changes in recall variance with delay and set size, ensuring that differences in baseline performance across studies do not confound the effects of interest.

Distributions of errors in WM recall show sharper peaks and longer tails compared to von Mises distributions. While some studies have attributed heavy tails to random guessing, substantial work has demonstrated that they are naturally predicted by—and are a critical component of—models assuming noisy and variable memory representations (Bays, 2014; Fougnie et al., 2012; Schurgin et al., 2020; van den Berg et al., 2012). While this debate is outside the scope of the present article and is covered extensively elsewhere (Bays et al., 2024; Ma et al., 2014; Taylor & Bays, 2020), we believe that including heavy tails in estimates of recall variance is uncontroversial: Doing so reflects the state of memory as a function of set size and delay, regardless of whether those tails are interpreted as guesses or low-precision representations.

The reliability and stability of estimates from the contamination model used to identify swap errors are known, similar to other analytical tools, to depend on the number of observations. Some estimates suggest at least 30 trials are required to obtain stable parameter estimates (https://bayslab.com/toolbox/index.php), which is comparable to the number of trials in some of our newly collected online experiments. While caution in the fitting procedure is therefore warranted, we argue that a relatively low number of trials per condition in some experiments cannot explain our results or provide an alternative account for the observed effect. First, the effects on both recall variance and bias are consistent across experiments with larger and smaller numbers of trials. Second, the experiments that collected a smaller number of trials used relatively low set sizes (e.g., one and three items), which are dominated by the target-centered mixture component, making parameter estimation more reliable. Finally, although instability in the estimation procedure could lead to some heterogeneity across experiments, it would not differentially affect conditions with different set sizes and delays in a way that could produce the systematic patterns observed here.

Theoretical Accounts of Memory Deterioration

Despite extensive behavioral (Pertzov et al., 2017; Rademaker et al., 2018; Ricker et al., 2014; Shin et al., 2017; Zhang & Luck, 2009), computational (Burak & Fiete, 2012; Panichello et al., 2019; Schneegans & Bays, 2018), and neural work (Li et al., 2026; Lim et al., 2019; Wimmer et al., 2014; Wolff et al., 2020) examining WM dynamics over time—and despite broad agreement that memory precision deteriorates with delay—the mechanisms underlying this deterioration remain debated.

The results presented here bear directly on a central theoretical debate concerning the behavioral signatures of forgetting in WM, specifically, whether increasing noise necessarily entails systematic bias. Competing theoretical accounts make divergent predictions: Some predict that increases in noise (due to delay) must be accompanied by systematic distortions in addition to increases in unsystematic error (Panichello et al., 2019; J. Yang et al., 2024; Ye et al., 2025), whereas others predict primarily increases in unsystematic error without concomitant changes in bias (Burak & Fiete, 2012; Compte et al., 2000; Koyluoglu et al., 2017). These opposing predictions can be broadly mapped onto two classes of attractor models: discrete attractor models, in which noise drives representations toward stable states and produces systematic biases in addition to increasing recall variance, and continuous attractor models, in which noise leads to unbiased diffusion and degradation of memory representations. Consistent with the latter, our results show systematic increases in variance with noise but no corresponding increase in bias, favoring continuous over discrete

attractor accounts of forgetting in WM; we return to the implications for discrete attractor theories later in the Discussion section.

The current results have direct implications for computational models that link diffusion strength to the number of items held in memory. Schneegans and Bays (2018) developed a computational model grounded in diffusion principles to test predictions of competing theories of forgetting: one postulating decay of memory signals over time (Baddeley, 1986; Barrouillet et al., 2007; Zhang & Luck, 2009) and another postulating diffusion of memory representations driven by accumulated noise (Kinchla & Smyzer, 1967; Nilsson & Nelson, 1981). Although both classes of models predict changes in recall precision with increasing delay, they make distinct predictions for response latency, which, in an evidence-accumulation framework, closely depends on the strength of the neural signal underlying the memory representation (Pearson et al., 2014). Specifically, decay-based accounts predict increased response latencies over time, analogous to the effects of increasing set size, which also reduces neural signal. In contrast, diffusion-based accounts predict increasing error in the population vector with delay, due to accumulating noise, but no changes in representational signal strength across the delay period and therefore stable response latencies. By jointly analyzing response variability and response latencies in a memory-guided saccade task, Schneegans and Bays identified behavioral signatures consistent with diffusion rather than decay, and model-based analyses further indicated that diffusion strength increased with set size. Consistent with these findings, our large-scale analysis also revealed systematic increases in forgetting rate with set size, providing converging evidence for set-size-dependent diffusion and further constraining how diffusion processes must be implemented within attractor network models.

A similar conclusion regarding the effect of set size on the rate of diffusion was reported by Koyluoglu et al. (2017). Specifically, the authors proposed an information-theoretic model in which information is efficiently encoded in a limited bandwidth system, designed to minimize the impact of downstream noise before storage in a persistent activity network. In this framework, diffusion—and consequently recall error—increases with the number of objects held in memory, aligning with our findings across a large set of experiments. Although this model provided a better fit than the competing models considered, its predictions were confined to overall response error and did not account for the presence of swap errors. Swap errors increase in frequency with both retention interval and set size (Tomić & Bays, 2024a), and because they are randomly distributed relative to the target feature, they inflate estimates of overall response error if not properly accounted for. Here, we corrected the measure of recall error for the presence of swap errors, confirming that set-size-dependent forgetting is not an artifact of set-size-dependent swap rate.

The current results also speak to neural evidence for diffusion-based dynamics in WM. Memory studies using single-unit neural recording in monkeys (Wimmer et al., 2014), as well as functional magnetic resonance imaging (Li et al., 2026; Lim et al., 2019) and electroencephalography (Wolff et al., 2020) studies in humans, have reported gradual, systematic shifts in neural representations over time that correspond to behavioral response errors, consistent with diffusion dynamics. Specifically, in a one-item memory task, population vectors encoding memorized location in the prefrontal cortex of rhesus monkeys drifted during the memory delay in the direction of subsequent behavioral response errors (Wimmer et al., 2014). Using a comparable memory-guided saccade task in humans with functional magnetic resonance imaging recording, Li et al. (2026) similarly found that decoding errors drifted over the delay from the veridical location toward the direction of eventual saccadic response. Together, these findings provide converging neural evidence for diffusion-like dynamics in WM representations, consistent with the behavioral patterns observed in the present study.

The present results could inform the design of future behavioral and neural studies. In particular, they suggest that information is stored in neural codes that facilitate competition between object representations, resulting in faster deterioration as the number of objects increases. Such competitive interactions are characteristic of divisive normalization and compatible with attractor models, where mutual excitation and inhibition among active parts of the network drive the dynamics (Almeida et al., 2015; Bouchacourt & Buschman, 2019; Johnson et al., 2009; Z. Wei et al., 2012). Future neural studies could extend their scope to multiobject and multidelay designs, testing predictions about the neural dynamics of forgetting as a function of set size.

Three newly collected studies (Experiments 1, 3, and 4) collected confidence estimates at the end of each trial in addition to memory recall. Because not all studies collected confidence data, we did not analyze or report these data here. However, based on previous work on metacognition in VWM (Li et al., 2021; van den Berg et al., 2017), which showed that humans possess knowledge of the uncertainty in their recall estimates, we would expect mean confidence to follow recall precision by decreasing with both set size and delay, with a similar interaction between these factors as reported here for recall variance. Mechanistically, population coding models have been particularly effective in capturing subjective confidence, as well as related proxies such as response times, in analogue report tasks (Bays, 2016; Schneegans & Bays, 2018; Tomić et al., 2025), and could provide a unified framework for explaining changes in both response error and confidence as noise in memory increases.

Modulation of Memory Deterioration by Cues

While memory deterioration over time may be inevitable, it can be accelerated through active removal (Taylor et al., 2023) or slowed for a prioritized subset of objects. In experimental settings, the latter is typically achieved by presenting an informative "retro-cue" that indicates which of the multiple objects held in WM is most likely to be subsequently cued for recall (Griffin & Nobre, 2003; Landman et al., 2003). The typical finding is that validly cued items show slower deterioration over time compared to noncued items (i.e., when no cue is provided), with the effect being even more pronounced when comparing validly cued items to invalidly cued ones (i.e., when the cue signals an item that is not subsequently probed; Pertzov et al., 2013). This suggests that slowing deterioration for one object may come at the cost of faster deterioration of other objects in memory. Neural models of behavior in retro-cue tasks propose that cued and noncued objects are represented with differing levels of neural activity, with the activity of noncued items being reduced (Bays & Taylor, 2018). Together, these findings support the idea that informative cues can mitigate memory deterioration over time.

Stimulus-Specific Biases and WM Limitations

In recent years, stimulus-specific response biases have received considerable attention alongside overall response variability in

VWM. While the existence of these biases is well established, their relationship with the constraints of the memory system remains unclear. Specifically, biases have been reported across both perceptual and memory tasks for a range of visual features (Andrews, 1967; Appelle, 1972; Bae et al., 2014; Girshick et al., 2011; Hardman et al., 2017; Jastrow, 1892; Pratte et al., 2017; Van Bergen et al., 2015; X.-X. Wei & Stocker, 2015, 2017). However, a systematic investigation into the impact of WM limitations on bias amplitude has been lacking. We applied a nonparametric method to estimate response biases in our extensive data set. This method makes no assumptions about the shape of the bias function, estimating it directly from the data. When applied to orientation recall responses, it recovered the widely reported, approximately sine-wave bias away from the cardinal orientations (horizontal and vertical). Importantly, our results indicate that the estimated biases remained stable under varying capacity and maintenance demands, without significant changes in magnitude over time or for different set sizes. This finding contrasts sharply with the effects of set size and delay on unsystematic response variability, suggesting that distinct computational processes with different limitations may govern recall biases and variability.

A particularly influential framework for understanding stimulus-specific biases in perception suggests that these biases arise from adaptation to environmental statistics (Mao & Stocker, 2024; X.-X. Wei & Stocker, 2015, 2017; for a review see Tomić et al., 2026). This account integrates principles of efficient coding and Bayesian decoding, wherein biases result from the redistribution of encoding resources according to the frequency distribution of stimulus features in the environment. In this framework, responses are repelled from the peak of the prior, which represents the most frequent feature values (e.g., cardinal orientations). Simultaneously, the Bayesian component partially mitigates these repulsive biases by shifting responses closer to the peak of the prior, with the net effect being an overall repulsive bias. Integrating these principles into a population coding model of VWM has been shown to accurately capture both the overall distribution of response errors and the shape of stimulus-specific biases (Taylor & Bays, 2018). However, future work is needed to investigate how WM limitations influence the encoding–decoding computations across different stages of this representational pipeline and how these processes give rise to the patterns of variability and bias observed in our analyses.

Implications for Attractor Models of WM

The results of the present study challenge the proposed link between memory maintenance and stimulus-specific biases in WM, as formalized in models specifically developed to account for an increase in bias over retention (Eissa & Kilpatrick, 2023; Panichello et al., 2019; J. Yang et al., 2024; Ye et al., 2025). These models predict that bias amplitudes should increase along with variability as set size and retention intervals grow. Specifically, biases are conceived as the outcome of corrective dynamics that draw representations toward discrete, stable states, thereby preventing them from diffusing freely throughout the feature space. A concrete instantiation of these models is the computational framework proposed by J. Yang et al. (2024), which models changes in response biases in VWM as a function of retention interval. The model relies on recursively interacting sensory and memory systems, with the sensory module reflecting efficient encoding based on environmental statistics, while the memory module introduces drift and diffusion dynamics in an attractor network. With orientation as a memory feature, the model predicts increasing repulsive biases with delay due to stronger drift of representations toward oblique orientations, resulting from delay-induced changes in connection heterogeneity within the sensory module and connectivity strengths between the two modules. Although the model is inspired by biologically plausible and commonly used computational mechanisms, it was not directly fitted to data but rather qualitatively compared to a schematic of psychophysical data. Our extensive analysis of a large set of experiments, incorporating various set size and delay manipulations as well as different visual features, contradicts the assumptions of this model.

The progenitor of many of these studies, Panichello et al. (2019), proposed a model in which random diffusion with delay is supplemented by systematic drift, steering representations toward a set of stable mnemonic states. Using color as a memory feature in an analogue report task, they reported an increase in bias with both set size and retention interval. However, focusing on the two experiments that simultaneously manipulated set size and delay, our reanalysis found moderate support for an effect of set size on bias amplitude in one experiment but strong evidence against an effect in the other, and no support in either experiment for an effect of delay on bias amplitude. Across our full compendium of studies, we found no consistent evidence for either effect. While the presence of stimulus-specific biases in VWM recall is strongly supported by our analysis and accords with previous findings (Bae, 2021; Bae et al., 2014; Langlois et al., 2021; Rademaker et al., 2018; Q. Yang & Li, 2025), the hypothesis that these biases result from accumulated drift during memory retention is not consistent with our results.

Another challenge for models based on discrete attractor dynamics is determining the principle by which attractors emerge across different feature spaces. Using a color estimation task in which the frequency of different colors was experimentally manipulated, Panichello et al. (2019) found that memory reports became biased toward frequently presented colors, and they presented a model of these results in which discrete attractor states developed at those common color locations. On this basis, the authors argued that attractors emerge around frequently encountered stimuli. However, in orientation estimation tasks, stable attractors positioned around the most common stimuli (i.e., cardinal orientations) would predict behavioral biases inconsistent with most psychophysical studies, which instead report repulsive biases away from the nearest cardinal axis. This pattern, observed across all orientation experiments analyzed here, can be naturally explained by efficient coding principles without assuming stable attractors (Taylor & Bays, 2018), as discussed in the previous section.

References

Almeida, R., Barbosa, J., & Compte, A. (2015). Neural circuit basis of visuospatial working memory precision: A computational and behavioral study. *Journal of Neurophysiology*, *114*(3), 1806–1818. https://doi.org/10.1152/jn.00362.2015

Alvarez, G. A., & Cavanagh, P. (2008). Visual short-term memory operates more efficiently on boundary features than on surface features. *Perception & Psychophysics*, *70*(2), 346–364. https://doi.org/10.3758/pp.70.2.346

Andrews, D. P. (1967). Perception of contour orientation in the central fovea Part I: Short lines. *Vision Research*, *7*(11), 975–997. https://doi.org/10.1016/0042-6989(67)90014-4

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. https://doi.org/10.1037/h0033117

Baddeley, A. D. (1986). *Working memory*. Clarendon Press, Oxford University Press.

Bae, G.-Y. (2021). Neural evidence for categorical biases in location and orientation representations in a working memory task. *NeuroImage*, *240*, Article 118366. https://doi.org/10.1016/j.neuroimage.2021.118366

Bae, G.-Y., Olkkonen, M., Allred, S. R., Wilson, C., & Flombaum, J. I. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision*, *14*(4), Article 7. https://doi.org/10.1167/14.4.7

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 570–585. https://doi.org/10.1037/0278-7393.33.3.570

Barrouillet, P., & Camos, V. (2012). As time goes by: Temporal constraints in working memory. *Current Directions in Psychological Science*, *21*(6), 413–419. https://doi.org/10.1177/0963721412459513

Barrouillet, P., De Paepe, A., & Langerock, N. (2012). Time causes forgetting from working memory. *Psychonomic Bulletin & Review*, *19*(1), 87–92. https://doi.org/10.3758/s13423-011-0192-8

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, *34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P. M. (2016). A signature of neural coding at human perceptual limits. *Journal of Vision*, *16*(11), Article 4. https://doi.org/10.1167/16.11.4

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

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, *11*(10), Article 6. https://doi.org/10.1167/11.10.6

Bays, P. M., Schneegans, S., Ma, W. J., & Brady, T. F. (2024). Representation and computation in visual working memory. *Nature Human Behaviour*, *8*(6), 1016–1034. https://doi.org/10.1038/s41562-024-01871-2

Bays, P. M., & Taylor, R. (2018). A neural model of retrospective attention in visual working memory. *Cognitive Psychology*, *100*, 43–52. https://doi.org/10.1016/j.cogpsych.2017.12.001

Bouchacourt, F., & Buschman, T. J. (2019). A flexible model of working memory. *Neuron*, *103*(1), 147–160.e8. https://doi.org/10.1016/j.neuron.2019.04.020

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Burak, Y., & Fiete, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(43), 17645–17650. https://doi.org/10.1073/pnas.1117386109

Carandini, M., & Heeger, D. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Cohen-Dallal, H., Fradkin, I., & Pertzov, Y. (2018). Are stronger memories forgotten more slowly? No evidence that memory strength influences the rate of forgetting. *PLOS ONE*, *13*(7), Article e0200292. https://doi.org/10.1371/journal.pone.0200292

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*(9), 910–923. https://doi.org/10.1093/cercor/10.9.910

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*(1), 115–142. https://doi.org/10.1146/annurev-psych-010814-015031

Eissa, T. L., & Kilpatrick, Z. P. (2023). Learning efficient representations of environmental priors in working memory. *PLOS Computational Biology*, *19*(11), Article e1011622. https://doi.org/10.1371/journal.pcbi.1011622

Emrich, S. M., & Ferber, S. (2012). Competition increases binding errors in visual working memory. *Journal of Vision*, *12*(4), Article 12. https://doi.org/10.1167/12.4.12

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

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

Gu, H., Lee, J., Kim, S., Lim, J., Lee, H.-J., Lee, H., Choe, M. J., Yoo, D.-G., Ryu, J. H. J., Lim, S., & Lee, S.-H. (2025). Attractor dynamics of working memory explain a concurrent evolution of stimulus-specific and decision-consistent biases in visual estimation. *Neuron*, *113*(20), 3476–3490.e9. https://doi.org/10.1016/j.neuron.2025.07.003

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), 30–54. https://doi.org/10.1037/xhp0000290

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A Conceptual Introduction to Bayesian Model Averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. https://doi.org/10.1177/2515245919898657

JASP Team. (2022). *JASP* (Version 0.16.1) [Computer software].

Jastrow, J. (1892). Studies from the university of Wisconsin: On the judgment of angles and positions of lines. *The American Journal of Psychology*, *5*(2), 214–248. https://doi.org/10.2307/1410867

Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(5), 568–577. https://doi.org/10.1111/j.1467-9280.2009.02329.x

Kinchla, R. A., & Smyzer, F. (1967). A diffusion model of perceptual memory. *Perception & Psychophysics*, *2*(6), 219–229. https://doi.org/10.3758/BF03212471

Koyluoglu, O. O., Pertzov, Y., Manohar, S., Husain, M., & Fiete, I. R. (2017). Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity. *eLife*, *6*, Article e22225. https://doi.org/10.7554/eLife.22225

Krill, D., Avidan, G., & Pertzov, Y. (2018). The rapid forgetting of faces. *Frontiers in Psychology*, *9*, Article 1319. https://doi.org/10.3389/fpsyg.2018.01319

Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, *43*(2), 149–164. https://doi.org/10.1016/s0042-6989(02)00402-9

Langlois, T. A., Jacoby, N., Suchow, J. W., & Griffiths, T. L. (2021). Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(13), Article e2012938118. https://doi.org/10.1073/pnas.2012938118

Li, H.-H., Ma, W. J., & Curtis, C. E. (2026). Dynamics of working memory drift and information flow across the cortical hierarchy. *Proceedings of the National Academy of Sciences of the United States of America*, *123*(4), Article e2518110123. https://doi.org/10.1073/pnas.2518110123

Li, H.-H., Sprague, T. C., Yoo, A. H., Ma, W. J., & Curtis, C. E. (2021). Joint representation of working memory and uncertainty in human cortex. *Neuron*, *109*(22), 3699–3712.e6. https://doi.org/10.1016/j.neuron.2021.08.022

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the*

American Statistical Association, *103*(481), 410–423. https://doi.org/10.1198/016214507000001337

Lim, P. C., Ward, E. J., Vickery, T. J., & Johnson, M. R. (2019). Not-so-working memory: Drift in functional magnetic resonance imaging pattern representations during maintenance predicts errors in a visual working memory task. *Journal of Cognitive Neuroscience*, *31*(10), 1520–1534. https://doi.org/10.1162/jocn_a_01427

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Mao, J., & Stocker, A. A. (2024). Sensory perception is a holistic inference process. *Psychological Review*, *131*(4), 858–890. https://doi.org/10.1037/rev0000457

McKeown, D., & Mercer, T. (2012). Short-term forgetting without interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 1057–1068. https://doi.org/10.1037/a0027749

McMaster, J. M. V., Tomić, I., Schneegans, S., & Bays, P. M. (2022). Swap errors in visual working memory are fully explained by cue-feature variability. *Cognitive Psychology*, *137*, Article 101493. https://doi.org/10.1016/j.cogpsych.2022.101493

Nilsson, T. H., & Nelson, T. M. (1981). Delayed monochromatic hue matches indicate characteristics of visual memory. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(1), 141–150. https://doi.org/10.1037//0096-1523.7.1.141

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Panichello, M. F., DePasquale, B., Pillow, J. W., & Buschman, T. J. (2019). Error-correcting dynamics in visual working memory. *Nature Communications*, *10*(1), Article 3366. https://doi.org/10.1038/s41467-019-11298-3

Pearson, B., Raskevicius, J., Bays, P. M., Pertzov, Y., & Husain, M. (2014). Working memory retrieval as a decision process. *Journal of Vision*, *14*(2), Article 2. https://doi.org/10.1167/14.2.2

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(5), 1224–1231. https://doi.org/10.1037/a0030947

Pertzov, Y., Manohar, S., & Husain, M. (2017). Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 528–536. https://doi.org/10.1037/xlm0000328

Phillips, W. A., & Baddeley, A. D. (1971). Reaction time and short-term visual memory. *Psychonomic Science*, *22*(2), 73–74. https://doi.org/10.3758/BF03332500

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing* (pp. 1–10). Technische Universität Wien.

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 6–17. https://doi.org/10.1037/xhp0000302

Prinzmetal, W., Amiri, H., Allen, K., & Edwards, T. (1998). Phenomenology of attention: I. Color, location, orientation, and spatial frequency. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(1), 261–282. https://doi.org/10.1037/0096-1523.24.1.261

Rademaker, R. L., Park, Y. E., Sack, A. T., & Tong, F. (2018). Evidence of gradual loss of precision for simple features and complex objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 925–940. https://doi.org/10.1037/xhp0000491

Rerko, L., Oberauer, K., & Lin, H.-Y. (2014). Spatial transposition gradients in visual working memory. *The Quarterly Journal of Experimental Psychology*, *67*(1), 3–15. https://doi.org/10.1080/17470218.2013.789543

Ricker, T. J., Spiegel, L. R., & Cowan, N. (2014). Time-based loss in visual short-term memory is from trace decay, not temporal distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1510–1523. https://doi.org/10.1037/xlm0000018

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In P. R. Krishnaiah (Ed.), *Multivariate analysis II* (pp. 25–31). Academic Press.

Salmela, V. R., Ölander, K., Muukkonen, I., & Bays, P. M. (2019). Recall of facial expressions and simple orientations reveals competition for resources at multiple levels of the visual hierarchy. *Journal of Vision*, *19*(3), Article 8. https://doi.org/10.1167/19.3.8

Schneegans, S., & Bays, P. M. (2016). No fixed item limit in visuospatial working memory. *Cortex*, *83*, 181–193. https://doi.org/10.1016/j.cortex.2016.07.021

Schneegans, S., & Bays, P. M. (2017). Neural architecture for feature binding in visual working memory. *Journal of Neuroscience*, *37*(14), 3913–3925. https://doi.org/10.1523/JNEUROSCI.3493-16.2017

Schneegans, S., & Bays, P. M. (2018). Drift in neural population activity causes working memory to deteriorate over time. *Journal of Neuroscience*, *38*(21), 4859–4869. https://doi.org/10.1523/JNEUROSCI.3440-17.2018

Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(34), 20959–20968. https://doi.org/10.1073/pnas.2004306117

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

Sewell, D. K., Lilburn, S. D., & Smith, P. L. (2014). An information capacity limitation of visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(6), 2214–2242. https://doi.org/10.1037/a0037744

Shin, H., Zou, Q., & Ma, W. J. (2017). The effects of delay duration on visual working memory for orientation. *Journal of Vision*, *17*(14), Article 10. https://doi.org/10.1167/17.14.10

Souza, A. S., & Oberauer, K. (2015). Time-based forgetting in visual working memory reflects temporal distinctiveness, not decay. *Psychonomic Bulletin & Review*, *22*(1), 156–162. https://doi.org/10.3758/s13423-014-0652-z

Staugaard, C. F., Petersen, A., & Vangkilde, S. (2016). Eccentricity effects in vision and attention. *Neuropsychologia*, *92*, 69–78. https://doi.org/10.1016/j.neuropsychologia.2016.06.020

Taylor, R., & Bays, P. M. (2018). Efficient coding in visual working memory accounts for stimulus-specific variations in recall. *Journal of Neuroscience*, *38*(32), 7132–7142. https://doi.org/10.1523/JNEUROSCI.1018-18.2018

Taylor, R., & Bays, P. M. (2020). Theory of neural coding predicts an upper bound on estimates of memory variability. *Psychological Review*, *127*(5), 700–718. https://doi.org/10.1037/rev0000189

Taylor, R., Tomić, I., Aagten-Murphy, D., & Bays, P. M. (2023). Working memory is updated by reallocation of resources from obsolete to new items. *Attention, Perception, & Psychophysics*, *85*(5), 1437–1451. https://doi.org/10.3758/s13414-022-02584-2

Tomić, I., Adamcová, D., Fehér, M., & Bays, P. M. (2024). Dissecting the components of error in analogue report tasks. *Behavior Research Methods*, *56*(8), 8196–8213. https://doi.org/10.3758/s13428-024-02453-w

Tomić, I., & Bays, P. M. (2018). Internal but not external noise frees working memory resources. *PLOS Computational Biology*, *14*(10), Article e1006488. https://doi.org/10.1371/journal.pcbi.1006488

Tomić, I., & Bays, P. M. (2024a). A dynamic neural resource model bridges sensory and working memory. *eLife*, *12*, Article RP91034. https://doi.org/10.7554/eLife.91034

Tomić, I., & Bays, P. M. (2024b). Perceptual similarity judgments do not predict the distribution of errors in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(4), 535–549. https://doi.org/10.1037/xlm0001172

Tomić, I., Lengyel, M., & Bays, P. (2026). Bayesian efficient coding as a theory of perception: Progress, controversies, and prospects. *Trends in Cognitive Sciences*. Advance online publication. https://doi.org/10.1016/j.tics.2026.05.001

Tomić, I., Raimundo, R., & Bays, P. M. (2025). *Intrinsic rewards guide visual resource allocation via reinforcement learning*. bioRxiv.

Van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730. https://doi.org/10.1038/nn.4150

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124–149. https://doi.org/10.1037/a0035234

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review*, *124*(2), 197–214. https://doi.org/10.1037/rev0000060

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wei, X.-X., & Stocker, A. A. (2015). A bayesian observer model constrained by efficient coding can explain 'anti-bayesian' percepts. *Nature Neuroscience*, *18*(10), 1509–1517. https://doi.org/10.1038/nn.4105

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(38), 10244–10249. https://doi.org/10.1073/pnas.1619153114

Wei, X.-X., & Woodford, M. (2025). Representational geometry explains puzzling error distributions in behavioral tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(4), Article e2407540122. https://doi.org/10.1073/pnas.2407540122

Wei, Z., Wang, X.-J., & Wang, D.-H. (2012). From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. *Journal of Neuroscience*, *32*(33), 11228–11240. https://doi.org/10.1523/JNEUROSCI.0735-12.2012

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), Article 11. https://doi.org/10.1167/4.12.11

Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, *17*(3), 431–439. https://doi.org/10.1038/nn.3645

Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2020). Drifting codes within a stable coding scheme for working memory. *PLOS Biology*, *18*(3), Article e3000625. https://doi.org/10.1371/journal.pbio.3000625

Yang, J., Zhang, H., & Lim, S. (2024). Sensory-memory interactions via modular structure explain errors in visual working memory. *eLife*, *13*, Article RP95160. https://doi.org/10.7554/eLife.95160

Yang, Q., & Li, H.-H. (2025). Efficient allocation of working memory resource for utility maximization in humans and recurrent neural networks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, & N. Chen (Eds.), *Advances in neural information processing systems* (Vol. 38, pp. 57504–57528). Curran Associates.

Ye, Z., Li, H., Tian, L., & Zhou, C. (2025). The effects of the post-delay epochs on working memory error reduction. *PLOS Computational Biology*, *21*(5), Article e1013083. https://doi.org/10.1371/journal.pcbi.1013083

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science*, *20*(4), 423–428. https://doi.org/10.1111/j.1467-9280.2009.02322.x

Zokaei, N., Gorgoraptis, N., Bahrami, B., Bays, P. M., & Husain, M. (2011). Precision of working memory for visual motion sequences and transparent motion surfaces. *Journal of Vision*, *11*(14), Article 2. https://doi.org/10.1167/11.14.2

Received September 24, 2025
Revision received March 6, 2026
Accepted July 1, 2026 ■