

A dynamic neural resource model bridges sensory and working memory

Ivan Tomić , Paul M. Bays

University of Cambridge, Department of Psychology, Cambridge, UK • University of Zagreb, Department of Psychology, Zagreb, CRO

 https://en.wikipedia.org/wiki/Open_access Copyright information

Reviewed Preprint

Revised by authors after peer review.

[About eLife's process](#)

Reviewed preprint version 2

March 21, 2024 (this version)

Reviewed preprint version 1

November 1, 2023

Sent for peer review

August 18, 2023

Posted to preprint server

July 14, 2023

Abstract

Probing memory of a complex visual image within a few hundred milliseconds after its disappearance reveals significantly greater fidelity of recall than if the probe is delayed by as little as a second. Classically interpreted, the former taps into a detailed but rapidly decaying visual sensory or “iconic” memory (IM), while the latter relies on capacity-limited but comparatively stable visual working memory (VWM). While iconic decay and VWM capacity have been extensively studied independently, currently no single framework quantitatively accounts for the dynamics of memory fidelity over these timescales. Here we extend a stationary neural population model of VWM with a temporal dimension, incorporating rapid sensory-driven accumulation of activity encoding each visual feature in memory, and a slower accumulation of internal error that causes memorized features to randomly drift over time. Instead of facilitating read-out from an independent sensory store, an early cue benefits recall by lifting the effective limit on VWM signal strength imposed when multiple items compete for representation, allowing memory for the cued item to be supplemented with information from the decaying sensory trace. Empirical measurements of human recall dynamics validate these predictions while excluding alternative model architectures. A key conclusion is that differences in capacity classically thought to distinguish IM and VWM are in fact contingent upon a single resource-limited WM store.

eLife assessment

This study presents **important** insights into the dynamical process whereby sensory information is converted from stimulus-driven activity to a working memory representation from which the information can be recalled later. The evidence supporting the claims is **convincing**, using detailed fits and model-comparison techniques applied to new and existing psychophysical data sets to evaluate a wide variety of potential mechanisms. The overall conclusion, that iconic memory and working memory are not distinct mechanisms but rather two slightly different regimes of the same circuitry, will be of interest to neuroscientists and psychologists studying sensory systems and/or working memory.

Introduction

Keeping relevant information in an easily accessible state is vital for adaptive behavior in dynamic environments. In the primate visual system, this requirement is met by visual working memory (VWM), the capacity to actively maintain visual information from milliseconds to seconds after a stimulus disappears from view [1–4]. While the contents of VWM are frequently updated to reflect changes in the environment and in behavioral priorities, the visual processing hierarchy itself introduces additional layers of dynamism [5, 6]. The fidelity of representations therefore evolves from the moment VWM starts accumulating evidence [7, 8] throughout the maintenance period until the information is used for action [9–11].

Nonetheless, within most theoretical frameworks, VWM is treated as a stationary process whereby representations are measured and modelled as fixed states of the system. One such model of working memory is based on principles of neural population coding [12, 13]. In the Neural Resource model, visual information is encoded in the activity of a population of noisy feature-selective neurons [14, 15]. The spiking activity of the neural population is constrained by normalization [16, 17], such that the total activity is fixed but flexibly distributed between memoranda, implementing a form of limited memory resource. At retrieval, encoded stimulus values are reconstructed from the noisy spiking activity. This model has provided a quantitative account of patterns of recall error across a range of tasks and stimulus dimensions [18–22]. However, despite its grounding in principles of neural coding, the basic architecture of the model lacks a temporal dimension to describe the dynamics of memory representations during encoding and maintenance.

Research on prolonged memory maintenance has demonstrated that the precision of stored representations gradually deteriorates over time (e.g., 23, 24). Computational models attempting to account for these dynamics have often relied on principles of diffusion within an attractor network. In such a network, information is maintained in a sustained pattern of activity, which can be visualized as a “bump” of activity centered on the stored value. Over time, the bump diffuses along the feature dimension due to random fluctuations in neural activity, leading to stochastic changes in the encoded feature value and a gradual loss of information [25, 26]. Critically, the neural code diffuses without decay in signal strength. A growing body of empirical support, both at the behavioral [9] and neural level [27, 28], identifies diffusion as a key mechanism of memory deterioration.

In contrast to such gradual deterioration over longer retention intervals, studies that probed memory within a few hundred milliseconds of stimulus offset revealed a precipitous decrease in memory fidelity immediately after a stimulus disappears [29–32]. This early superior recall was attributed to a high-capacity but short-lived form of storage termed iconic memory (IM) [33]. An implicit assumption has often been that the behavioral advantage of early cues derives from reading out information directly from IM and circumventing capacity limitations imposed by VWM, however, this idea has not been formally modelled or tested. At the neural level, IM is thought to be supported by a brief period of decaying neural activity in early visual areas following the response elicited by the visible stimulus [34–37]. In contrast to later memory dynamics arising due to noise accumulation, early changes in memory fidelity were supported by modulation of the neural signal strength. However, little is known about the read-out of this sensory memory buffer.

Finally, memory fidelity changes during encoding while the evidence is extracted from the visible stimulus. Previous studies revealed that longer stimulus exposures have a favorable effect on the subsequent recall, but that this effect is modulated by the number of simultaneously encoded objects [38–40], providing evidence for a processing or encoding limitation of VWM. As stimulus presentation duration increases, more information may be extracted from the sensory signal into VWM, increasing the fidelity of the representation. Critically, with prolonged exposure,

VWM fidelity approaches a stable level that depends on the number of encoded items, suggesting that a ceiling is imposed on evidence accumulation by a shared limit on VWM resources. However, a computational framework describing information accumulation from sensory areas into VWM is lacking, and the observed encoding limit may reflect dynamics in sensory areas registering visible objects as well as VWM accumulating this sensory evidence.

Here, we investigated the temporal dynamics in the fidelity of VWM from information encoding until its recall. To map human recall fidelity to the time domain, we conducted psychophysical experiments in which we probed memory representations at different time points relative to stimulus onset and offset while simultaneously manipulating set size. To isolate memory dynamics due to changes in the representational signal, we advanced an analogue reproduction task with a novel response method specifically adapted to minimize the time cost of motor (i.e., response) processes and capture the momentary state of memory representations. This allowed us to precisely measure the time course of fidelity dynamics during representation formation (i.e., encoding) and retention (i.e., maintenance). A major conclusion is that the enhanced precision seen at very brief retention intervals depends on integration of information from the sensory store into VWM following the cue, with the result that retrieval from IM of even the simplest stimulus is subject to the temporal and capacity limitations of working memory.

To explain the neural computations underlying the observed time courses, we devised a comprehensive neural model of memory dynamics whose core architecture is rooted in the Neural Resource model of VWM [12, 13]. The Dynamic Neural Resource (DyNR) model assumes that changes in memory fidelity reflect temporal dynamics in the sensory population registering the stimuli and from signal and noise accumulation processes of resource-limited VWM (Fig. 1). In particular, the model prescribes how time-dependent gain control mechanisms in sensory areas produce a smooth neural response following abrupt changes in stimulus presence. As this sensory signal provides feed-forward input to VWM, the dynamics in VWM activity in the temporal vicinity of stimulus presentation (i.e., onset and offset) strongly reflect not only limits in VWM, but also the dynamics of the sensory signal. Finally, once accumulated into VWM, the neural signal is subject to perturbations due to noise accumulation, resulting in degradation of internal representations with time. The DyNR model accurately reproduced the detailed empirical patterns of human recall errors in the psychophysical experiments. Based on these results, we argue that changes in memory fidelity on short time scales reflect dynamics in the gain or signal strength in neural populations representing the stimulus, while changes on longer time scales are dominated by corruption of the representation by accumulated noise.

Dynamic Neural Resource (DyNR) Model

The Dynamic Neural Resource model generalizes an established neural population account of VWM, originally proposed by Bays [12] and inspired by similar models of attention and perceptual decision-making [41–43]. In the original model, memorization and recall of visual stimuli is achieved by encoding and decoding of spiking activity in idealized feature-tuned neurons. The limited capacity of VWM to hold multiple object features simultaneously is reproduced by a global divisive normalization that constrains total spiking activity, implementing a continuous memory resource [16, 12]. The DyNR model (illustrated in Fig. 1) extends this stationary encoding-decoding model with a temporal dimension. First, to capture encoding dynamics, stimulus information enters the VWM population (Fig. 1, bottom) indirectly, by accumulation of neural signal from a separate sensory population (top), which receives the visual input. The signal strength in the VWM population at any point in time jointly depends on the history of the signal in the sensory population and the number of features competing for representation in VWM. Once the sensory signal is gone, the VWM signal is maintained at its maximum attained amplitude, but the stimulus value encoded by the signal gradually diffuses due

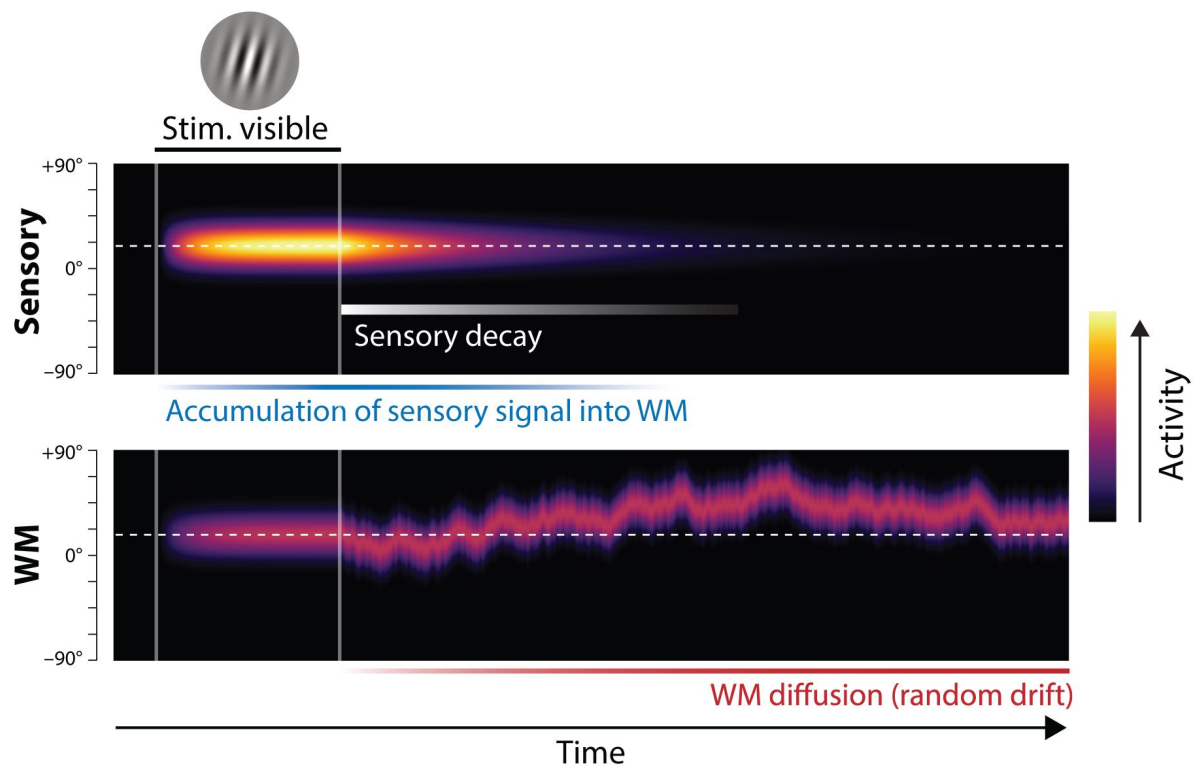


Figure 1

Proposed neural population dynamics for encoding a single orientation into VWM and maintaining it over a delay. Top: Stimulus onset is followed by a ramping increase in activity (indicated by color) of sensory neurons whose tuning (indicated on y axis) matches the stimulus orientation. Following stimulus offset, this sensory signal rapidly decays. The sensory signal, including its decaying post-stimulus component, provides input into VWM. Bottom: At stimulus onset, the VWM population begins to accumulate activity from the sensory population. This accumulation saturates at a maximum amplitude determined by global normalization. As the sensory activity decays, the activity in the VWM population is maintained at a constant amplitude, but accumulation of random errors causes the activity bump to diffuse along the feature dimension (y axis) over time, changing the orientation represented by the population. At recall, when the VWM population activity is decoded, accuracy of the recall estimate depends on both the orientation represented (centre of the activity bump) and the fidelity with which it can be retrieved (determined by activity amplitude).

to accumulation of random noise. Recall error depends on both the stimulus value represented at the time of retrieval (*what* is encoded) and the signal amplitude at that time, read out in the form of spikes (*how precisely* it can be decoded).

Dynamics of sensory signal strength

To model the temporal dynamics of human memory fidelity, we begin by defining computations of the sensory system registering the incoming signal. A particularly important computation is temporal filtering – a property of neurons to respond more sensitively to specific temporal patterns in stimuli. To model the signal represented in the cortical sensory level, we assume that the sensory response to a stimulus presentation of fixed duration (described as a step function in visual input amplitude, **Fig. 2A & B**, left) is controlled by a monophasic temporal filter having a low-pass frequency response [44]. This choice is a natural one since it is consistent with electrophysiological studies demonstrating that a large range of temporal frequencies registered by the retina and LGN [45, 46] is attenuated at higher frequencies before the signal enters the primary visual cortex [47]. Passing the stimulus through such a temporal filter attenuates the neural response to fast transients in the signal, and thereby produces a smooth rise and decay of neural activity in response to a uniform input signal (**Fig. 2C**). In particular, we assume that the activity of the sensory population after stimuli onset and offset changes exponentially towards the maximum sensory activity and baseline activity, respectively.

The choice of the filter's temporal response characteristics (i.e., its time constant) fully defines dynamics in the sensory population activity and controls the signal projected towards higher areas. The available physiological evidence suggests the temporal properties of the rising and decaying neural response are not symmetric [48–50]. In particular, the neural response typically reaches the maximum activity after the onset faster than it reaches the baseline activity after the offset. Consistent with this, we allowed the sensory signal to decay at a different rate than the rising rate. The temporal dynamics in sensory population firing activity in response to a fixed input signal of duration t_{offset} is then given by:

$$\dot{\gamma}_s(t) = \begin{cases} (\tilde{\gamma}_s - \gamma_s(t))/\tau_{\text{rise}} & \text{for } t \leq t_{\text{offset}} \\ -\gamma_s(t)/\tau_{\text{decay}} & \text{for } t > t_{\text{offset}} \end{cases} \quad (1)$$

where $\tilde{\gamma}_s$ is the maximum sensory signal, τ_{rise} and τ_{decay} are rising and decaying time constants of the temporal filter, respectively.

The temporal properties of the sensory response have been shown to depend on the physical characteristics of stimuli, such as contrast and location [48, 51]. Similarly, previous work has demonstrated that the decaying component of the sensory response is strongly influenced by the engagement of the sensory population after stimuli offset (e.g., 35). In particular, a new input signal, e.g. a backward noise mask, curtails ongoing activity related to the previous stimulus, resulting in a faster decay of activity compared to the unmasked post-stimulus period [52]. Consistent with this, here we assume that the backward mask operates by interrupting ongoing sensory processing of stimuli, limiting the access to the sensory signal (**Fig. 5**) (cf. integration mask) [53].

Dynamics of VWM signal strength

The information registered by the sensory system is subsequently accumulated into a VWM population capable of maintaining activity in the absence of further input (e.g. by self-excitation, see 54, 55, 26; although only the resulting dynamics are modelled here). The total activity of the VWM neural population is normalized, implementing a limited resource shared out between memory items [12, 13]. Consequently, if the stimuli are presented for long enough, the

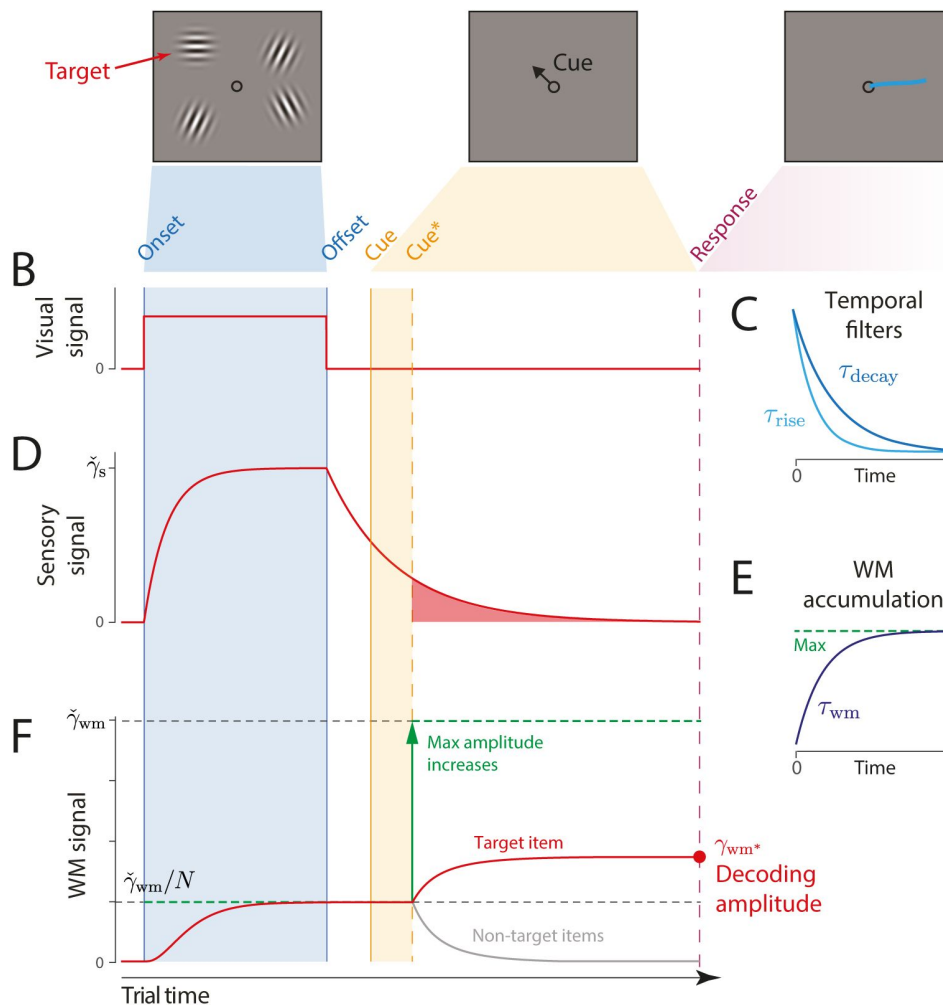


Figure 2

Schematic of signal amplitudes in the DyNR model during a cued recall trial. **(A)** Observers are presented with a memory array (left), followed after a blank delay (not shown) by an arrow cue (center) indicating the location of one item (the target) whose remembered orientation should immediately be reported (right). **(B)** The amplitude of the visual input associated with each item is modelled as a step function (left). The sensory response **(D)** is modelled as a low-pass filtering of the stimulus input, with different time constants for rise and decay **(C)**. **(F)** Amplitude of the working memory signal reflects a saturating accumulation of activity from the sensory population (illustrated in **E**). Beginning with stimulus onset, activity associated with each item is accumulated from the sensory population into the VWM population, approaching an upper bound (green dashed line) that reflects a total activity limit shared between the N items in memory. Once the cue has been presented (solid orange line) and processed (dashed orange line), uncued items can be dropped from VWM, raising the ceiling on activity available to represent the cued item (green arrow). This allows more information about the cued item to be accumulated from the decaying sensory trace (equivalent to the red shaded area in **D**). Response variability depends on the asymptotic VWM signal amplitude available for decoding (red circle) combined with the accumulated effects of diffusion (see text).

evidence accumulated from the sensory signal into VWM will saturate at a level that reflects the total number of stimuli represented (**Fig. 2D** [↗](#)). The dynamics in VWM population activity are given by:

$$\dot{\gamma}_{\text{wm}}(t) = \gamma_s(t)(\tilde{\gamma}_{\text{wm}}/M(t) - \gamma_{\text{wm}}(t))/\tau_{\text{wm}} \quad (2)$$

Where $\tilde{\gamma}_{\text{wm}}$ is the maximum VWM signal amplitude, $M(t)$ is the number of items represented in VWM at time t , τ_{wm} is the time constant of accumulation into VWM.

A common assumption of VWM models is that the strength of the representational signal remains stable after encoding from a visible stimulus. This stationary view has been reinforced by typically measuring VWM sufficiently long after the stimulus disappears (~ 1 second) and at a single time-point. In contrast, work on IM demonstrated that recall fidelity in a brief period after stimulus offset typically surpasses and then precipitously decays towards VWM fidelity level [\[56\]](#). Consistent with that, we consider how the normalized representational signal in VWM formed during encoding can be boosted in the absence of the physical stimulus. In particular, we assume a representation stored in VWM can be strengthened as long as the sensory population provides feed-forward input and VWM activity is not saturated at the normalized level. Such a scenario can be achieved by cueing an item for recall in the temporal vicinity of stimulus offset, i.e. before sensory activity decays to zero. By cueing an item for recall, the remaining contents of VWM becomes obsolete and can be removed from memory [\[57\]](#). In the model,

$$M(t) = \begin{cases} N & \text{for } t \leq t_{\text{cue}^*} \\ 1 & \text{for } t > t_{\text{cue}^*} \end{cases} \quad (3)$$

where t_{cue^*} is the time when the item is identified for a recall and the readout of stimulus value begins. This “dismounting” of resource from uncued items makes it available for storing additional information about the cued item, which is extracted from the residual sensory representation, increasing the representation fidelity beyond that granted by equal distribution of neural signal between items. Critically, as sensory information quickly decays, there will be less signal remaining to supplement the VWM representation of a cued item if the cue is delivered later, and at the longest cue intervals the cue will confer no advantage over the fidelity attained when all items compete equally for VWM representation (**Fig. 2D** [↗](#)).

We note that removal of uncued items cannot occur until the cue has been processed to the point of identifying one of the N items in the memory array. We follow Hick [\[58\]](#) in modelling this cue processing time as logarithmic in the number of alternatives:

$$t_{\text{cue}^*} = t_{\text{cue}} + b \log_2(N) \quad (4)$$

where b is a scaling parameter. Previous work demonstrated that estimation of temporal dynamics in attention and memory could be confounded with the time needed to interpret the cue and start acting on it [\[59\]](#). This is especially significant when trying to accurately capture quickly changing processes, such as decay of the sensory residual. Although the cue processing time likely fluctuates on a trial-by-trial basis due to changes in, e.g. attention, arousal, or motivation, here we focus on the influence of set size arising from a limited information processing capacity.

Diffusion of VWM encoded values

So far we have described only changes in the strength of the neural signal encoding features in memory. However, feature representations maintained over time in neural activity will accumulate noise in the absence of external input. We model this process of noise-driven diffusion as Brownian motion in feature space throughout the retention interval (**Fig. 1** [↗](#)), contributing to

variability in the decoded feature value [25, 9]. The resulting variability is described by a wrapped normal distribution with variance σ^2 that increases linearly with time from stimulus offset, so that at time t the encoded feature corresponding to a true stimulus feature θ is

$$\theta(t) \sim \mathcal{WN}(\theta, \sigma^2(t)) \quad (5)$$

$$\sigma^2(t) = (t - t_{\text{offset}})\sigma_{\text{diff}}^2 \quad (6)$$

where σ_{diff}^2 specifies the base diffusion rate. While the fast decay of sensory activity after stimuli offset accounts for early dynamics in VWM fidelity, diffusion becomes prominent over longer delays, accounting for more gradual deterioration of precision with time.

Such a diffusion account has support in the available neural evidence as well as in theoretical work. At the neural level, an electrophysiological study in monkeys performing a spatial working memory task demonstrated that shifts of neural tuning curves during a memory delay predicted behavioral response errors [26]. A similar finding was observed in humans where drift in the fMRI activity patterns relative to the target predicted errors in an orientation discrimination task [27]. At a theoretical level, continuous attractor models explain diffusion as a consequence of neural variability in networks where excitatory and inhibitory connections constrain population activity to a sub-space or manifold corresponding to the encoded feature space [25, 60, 55].

Retrieval

To model the process that leads to a response we first consider that in some trials observers may erroneously identify a non-target item as being cued. Previous work indicates these “swap” errors occur due to uncertainty in memory for the cue features of the stimuli, in this case their locations [20, 61]. We assume that changes in variability in the cue features mirror those of the memory features, leading swap frequency to decrease exponentially as a function of presentation duration and increase linearly with retention interval (Fig. S2):

$$p_{\text{swap}} = (N - 1) \left[\left(\frac{1}{N} - r_{\text{spatial}} t_{\text{cue}^*} \right) e^{-\frac{t_{\text{offset}}}{\tau_{\text{spatial}}}} + r_{\text{spatial}} t_{\text{cue}^*} \right] \quad (7)$$

where τ_{spatial} is the time constant related to presentation duration, and r_{spatial} is the rate constant related to the retention interval.

If θ is the true feature value of the item identified as the target (i.e. the cued item with probability $1 - p_{\text{swap}}$, a randomly selected non-cued item with probability p_{swap}), then due to diffusion (Eq. 5) the value encoded in the VWM population at the time of retrieval is given by

$$\theta^* \sim \mathcal{WN}(\theta, \sigma^2(t_{\text{cue}^*})) \quad (8)$$

We model retrieval as estimation of θ^* based on spiking activity in the VWM population that encodes the selected item. For this purpose we assume an idealized set of tuning functions, where the mean response of neuron i encoding orientation θ with population gain γ is described by

$$f_i(\theta, \gamma) = \frac{\gamma}{n} \exp(\kappa(\cos(\theta - \phi_i) - 1)) \quad (9)$$

where n is the number of neurons, and κ determines the tuning width. The preferred orientations of the neurons, ϕ_i , are evenly distributed throughout the circular space to provide uniform coverage. The spike count produced by each neuron is drawn from a Poisson distribution,

$$r_i \sim \text{Poisson}(f_i(\theta^*, \gamma_{\text{wm}^*})) \quad (10)$$

and the decoded orientation estimate is obtained by maximum likelihood estimation based on the spike counts:

$$\hat{\theta} = \arg \max_{\theta} p(\mathbf{r}|\theta). \quad (11)$$

Additional assumptions

To fit the model to behavioral data, we make several further simplifying assumptions. We assume that the exponential decay of the sensory signal is rapid enough that there is effectively no information remaining by the time the VWM population is decoded to generate a response. This allows us to approximate the VWM activity at the time of decoding by the asymptotic VWM activity were the sensory decay to continue indefinitely:

$$\gamma_{\text{wm}^*} \approx \gamma_{\text{wm}}(\infty) \quad (12)$$

Next, we identify diffusion in the encoded value at the time of retrieval with diffusion at the time of target item identification, justifying the use of t_{cue}^* in [Eq. 8](#). We reason that the rate of diffusion is slow enough relative to the rate of sensory decay, that any additional diffusion in the brief period of post-cue sensory accumulation is negligible.

In Experiment 1 (see below), a task with a fixed 200 ms exposure period, we assume that the initial encoding of all items into VWM is complete by the time of stimulus offset, i.e. that VWM activity at this time can be approximated by its asymptotic level reflecting normalization:

$$\gamma_{\text{wm}}(t_{\text{offset}}) \approx \tilde{\gamma}_{\text{wm}}/N \quad (13)$$

Finally, in the condition of Experiment 1 where memory array and cue are presented simultaneously, we assume that only the cued feature is encoded in VWM, reaching the maximum amplitude, $\tilde{\gamma}_{\text{wm}}$, irrespective of set size. Maximum likelihood fits were obtained via the Nelder-Mead simplex method (function *fminsearch* in Matlab). All parameters and variables used to describe the DyNR model are listed in [Table 1](#).

Overview of Experiments

We tested predictions of the Dynamic Neural Resource model against empirical data collected in continuous report tasks. In Experiment 1 ([Fig. 3A & B](#)), observers were presented with an array of oriented stimuli for a fixed duration followed after a variable delay by a visual cue identifying one of the preceding stimuli whose orientation should be reported. This experiment was designed to investigate the contribution of decaying sensory representations following stimulus offset to the dynamics of recall fidelity. Experiment 2 ([Fig. 3C](#)) was aimed at expanding the results of the first experiment to now also assess the accumulation of information during the time the stimuli were visible. In this case, the exposure duration was varied while the delay before the visual cue was held constant. In both experiments we varied the number of stimuli in the array (set size) to assess capacity limitations affecting encoding and maintenance.

To provide additional validation of the DyNR model, we also tested its predictions against data from a previously published continuous report experiment (Experiment 1 in [12](#)) and one additional dataset collected as part of a separate study [[62](#)]. A detailed description of all experiments is provided in Supplementary Information.

No.	Parameter/variable	Description
1	$\tilde{\gamma}_{wm}$	Maximum VWM signal amplitude
2	κ	Tuning curve width
3	τ_{rise}	Rise constant of the sensory temporal filter
4	τ_{decay}	Decay constant of the sensory temporal filter
5	τ_{wm}	Time constant of accumulation into VWM
6	σ_{diff}^2	Base diffusion rate
7	$\tau_{spatial}$	Time constant for spatial encoding
8	$r_{spatial}$	Rate constant for spatial diffusion
9	b	Scaling parameter for Hick's law
10	t	Time, relative to stimulus onset ($t = 0$)
11	t_{offset}	Time of stimulus offset
12	t_{cue}	Time of cue onset
13	t_{cue*}	Time an item is identified for report
14	N	Number of items in stimulus array
15	$M(t)$	Number of items in memory at time t
16	$\tilde{\gamma}_s$	Maximum sensory signal amplitude
17	$\gamma_s(t)$	Sensory signal amplitude at time t
18	$\gamma_{wm}(t)$	VWM signal amplitude at time t
19	γ_{wm*}	VWM signal amplitude at the time of decoding
20	$\sigma^2(t)$	Accumulated diffusion at time t
21	n	Number of neurons
22	θ	True stimulus feature value
23	θ^*	Encoded stimulus feature value at the time of decoding
24	$\hat{\theta}$	Decoded stimulus feature value

Table 1

DyNR model parameters (1–9) and other variables (10–24) used in model description.

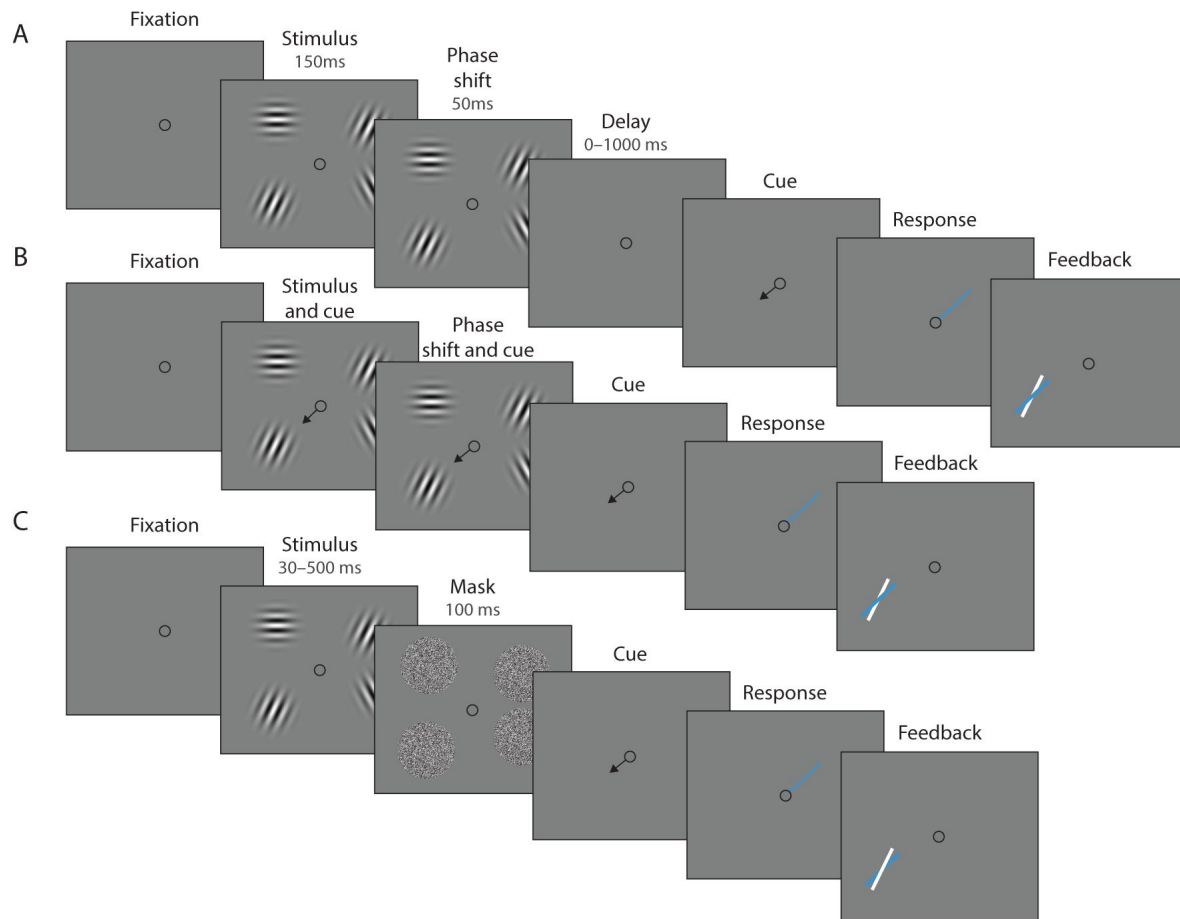


Figure 3

Experimental procedure. **(A)** Experiment 1. On each trial, a memory array was presented consisting of 1, 4, or 10 randomly oriented Gabor stimuli. In 50% of all trials, the stimuli underwent a change of phase and contrast towards the end of the exposure period intended to minimize retinal aftereffects. After a variable delay, an arrow cue was shown pointing towards the location of one stimulus from the preceding array. Observers reported the remembered orientation of the cued stimulus by swiping their index finger on the touchpad. The response was followed by feedback showing the true orientation. **(B)** In a proportion of trials, the cue was presented simultaneously with the stimuli. **(C)** Experiment 2. On each trial a memory array consisting of 1, 4, or 10 randomly oriented Gabors was presented for a variable duration, and followed by a white noise flickering mask. The mask was replaced by an arrow cue pointing towards the location of one stimulus from the preceding array. Observers reported its remembered orientation and received feedback as in Experiment 1. Stimuli are not drawn to scale.

Results

Experiment 1: Delay duration

In Experiment 1, we evaluated the time course of VWM fidelity over brief memory intervals. Previous work has demonstrated that immediately after a stimulus physically disappears, its representation briefly persists in the sensory system in the form of residual neural activity [36]. Accumulation of this lingering sensory activity into VWM could enable superior recall of information [56] within the constraints of a finite VWM resource that strongly limits representational fidelity [3]. To describe these dynamics, we examined human recall of orientation stimuli presented in arrays of varying sizes and probed after a variable delay ranging from 0 ms to 1000 ms. Here we focus on an experimental condition in which retinal afterimages were suppressed by a phase shift towards the end of stimuli presentation. Validation of this method and results from the condition without a phase shift are provided in the Supplementary Information.

Experimental data

Recall error distributions and mean performance in Experiment 1 are plotted in **Figs. 4A and B**. Response error (measured as RMSE) increased with both set size and delay duration. A repeated measures ANOVA revealed a significant effect of set size ($F_{(2,18)} = 117.8, p < .001, \eta^2 = .44$), delay time ($F_{(5,45)} = 52, p < .001, \eta^2 = .23$), and their interaction ($F_{(10,90)} = 26.7, p < .001, \eta^2 = .13$) on response error. We further explored this interaction, first finding response error in the 1 item condition (red in **Fig. 4**) did not change with delay ($F_{(5,45)} = 1.32, p = .27, \eta^2 = .07$). This was supported by Bayesian analysis ($BF_{10} = 0.34$) which found weak to moderate evidence against modulation of 1 item recall by memory delay. In contrast, response error increased with delay for the remaining two set sizes (4 items, green; 10 items, blue; main effect: $F_{(5,45)} = 55, p < .001, \eta^2 = .48$). This increase in response error consisted of an initial rapid rise (over the first 200 ms), followed by a more gradual increase as the delay between stimulus and cue increased. Next, we found a modulating effect of delay on recall for the remaining two set sizes (interaction: $F_{(5,45)} = 10.1, p < .001, \eta^2 = .05$). The direct comparison revealed that the increase in response error with delay ($\Delta\text{RMSE} = \text{RMSE}_{1000\text{ms}} - \text{RMSE}_{\text{Simult}}$) was greater when observers memorized more items ($t_{(9)} = 9.1, p < .001, d = 2.88$).

One surprising result was the observed set size effect in the 0 ms delay condition ($F_{(2,18)} = 23.7, p < .001, \eta^2 = .53$) consistent with a stepwise increase in recall error with set size (pairwise comparison, $t_{(9)} \geq 2.88, p \leq .036, d \geq 0.91$, Bonferroni correction applied). Importantly, this effect was a consequence of responding based on a memory of the stimulus, since orientation reproduction was comparable across set sizes in the perceptual condition (simultaneous presentation; $F_{(2,18)} = 1.26, p = .3, \eta^2 = .04, BF_{10} = 0.47$). Previous studies have characterized iconic memory as an effectively unlimited store, capable of holding any number of items without a consequent loss of fidelity [63, 30]. While our modelling ultimately affirmed this conception of IM, we nonetheless show that recall of information is contingent on the number of objects concurrently in memory from the moment stimuli physically disappear (see below).

Taken together, these results provide evidence that the fidelity of stored representations changes dramatically over the first few moments after stimuli offset. We next aimed to explain the neural computations supporting these dynamics. In summary, behavioral data displayed three key characteristics we aimed to explain, all visible in **Fig. 4B**. First, recall fidelity for a single item remained relatively stable across changes in delay, and was the same as perceptual fidelity. Second, recall fidelity for higher set sizes showed substantial, non-linear temporal dynamics. Lastly, recall fidelity was contingent on the number of stored items from the moment stimuli disappeared.

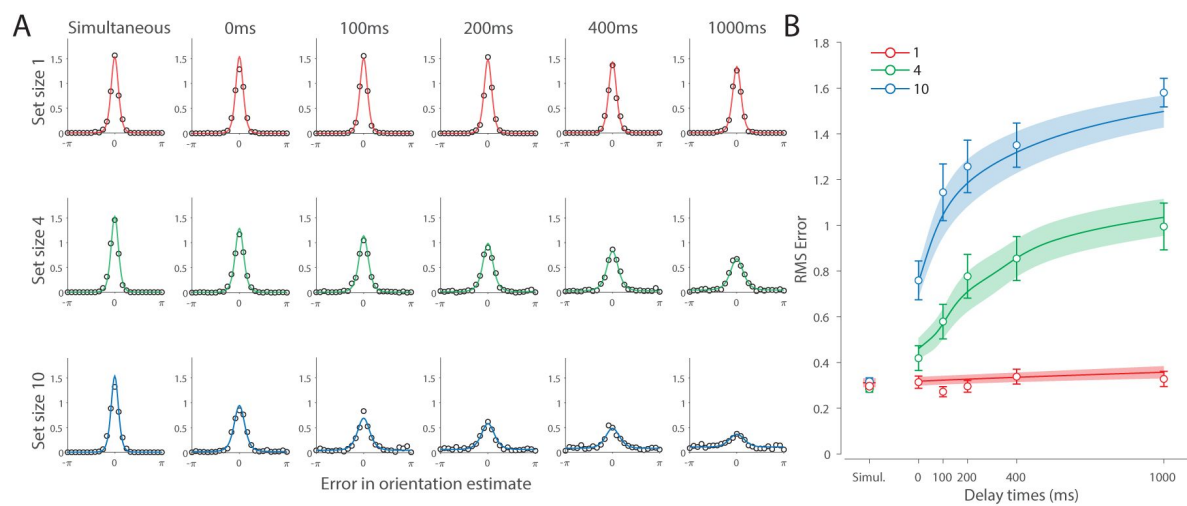


Figure 4

Experiment 1 data and model fits show the consequences of varying set size and delay duration on WM reproduction error. **(A)** Empirical recall error distributions (black circles) and the DyNR model fits (colored curves). Different panels correspond to different set sizes (rows) and delays (columns). **(B)** Corresponding RMS errors from experimental data (circles and errorbars) and the DyNR model fits (curves and error patches). Error bars and patches indicate ± 1 SEM.

Dynamic Neural Resource model

Curves in **Figs. 4A and B** show fits of the model with maximum likelihood (ML) parameters (mean $\pm$ SE: population gain $\gamma = 59.8 \pm 3.3$, tuning width $\kappa = 3.21 \pm 0.2$, sensory decay time constant $\tau_{\text{decay}} = 0.21 \pm 0.052$, VWM accumulation time constant $\tau_{\text{VWM}} = 0.096 \pm 0.045$, cue processing constant $b = 0.171 \text{ s} \pm .055 \text{ s}$, base diffusion $\sigma^2 = 0.03 \pm 0.017$, swap probability $p = 0.027 \pm 0.009$). The model provided a close fit to response error distributions (**Fig. 4A**) and summary statistics (**Fig. 4B**; see also **Fig. S2** for reproduction of swap error frequencies), successfully reproducing the pattern of changes with set size and delay. In particular, the model accounted for the three key observations identified above.

First, the model predicted the near-constant recall fidelity observed for a single item across these short retention intervals. The neural signal associated with the target object at recall depends on the normalized signal in VWM at offset supplemented by the available sensory signal post-cue. The sensory signal is integrated into VWM after the cue to fill any unallocated neural resource that arose by discarding uncued items. In the case of a single item, the entirety of VWM resources are allocated to one object during encoding, so no resource is freed by the cue that would allow the signal to be further strengthened based on the decaying sensory representation.

Importantly, this prediction contradicts the classical view of direct read-out from IM, according to which representational fidelity should be enhanced with very short delays irrespective of VWM limitations (see *Alternative accounts* below for a formal test of such a model). Note that the DyNR model nonetheless predicts some deterioration in fidelity over time even for a single item, due to noise-driven diffusion of the stored value. However, based on previous reports, we expected this process to be substantially slower and the impact on single item precision relatively small on this ($\leq 1 \text{ s}$) timescale. The fitted diffusion parameters and resulting shallow slope of fitted RMS error (red curve in **Fig. 4B**) confirmed this.

Second, the neural model predicts the specific pattern of dynamics observed in trials with multiple items (set sizes 4, green, and 10, blue curves). Once the cue is presented, resources encoding uncued items are freed and the decaying sensory signal representing the target item is further integrated into VWM, still subject to limited total VWM resources but now without competition from other items. Due to exponential decay of the sensory signal, the increase in fidelity thus accrued changes rapidly with retention interval over the first few hundred milliseconds. At longer delays, the cue identifies the target only after the sensory signal has effectively disappeared, so the VWM signal representing the target item remains at the normalized level reflecting equal distribution between all items in the memory array, and memory dynamics consist only of the more gradual deterioration of fidelity due to accumulated noise in the encoded value.

Finally, the DyNR model predicts the presence of a set size effect on fidelity throughout the entire memory period, including the no delay (0 ms) condition in which the cue onset was coincident with stimulus offset, but not in the simultaneous cue condition. In the model, this behavior emerges as a consequence of two independent processes. First, at the end of stimulus presentation, items within smaller (lower set size) arrays are encoded in VWM with higher signal amplitude, reflecting normalization. This signal strength represents a baseline that can be supplemented by further integration of the sensory signal after an early cue. However, if the sensory decay is sufficiently rapid, then even if the cue is presented immediately the target representation will not attain the maximum amplitude (equivalent to set size of one) starting from a lower baseline. Second, as described by Hick's Law [58] it takes longer to identify the target item based on the cue as the number of alternatives increases (see *Alternative models* below for a formal test of this assumption). As a result, for higher set sizes, less sensory signal encoding the target item remains to be integrated into VWM once it has been identified.

Model variants

We next focused on alternative explanations for the temporal dynamics observed in Experiment 1. Specifically, we examined whether the observed dynamics could be accounted for either solely by post-stimulus changes in neural signal amplitude or solely by noise-driven diffusion of stored values. To pre-empt our conclusions, we demonstrate that both components are needed to explain the observed dynamics in memory fidelity. Moreover, to more closely examine the role of diffusion in WM dynamics, we fit our neural model to an additional dataset collected in our lab ([62]; see Additional dataset 1 in Supplementary Information). This experiment used longer delays compared to those used in Experiment 1, and therefore precluded any beneficial effect of post-stimulus sensory information, while at the same time allowing the diffusion to operate over a longer period. This experiment allowed us to test whether diffusion is sufficient to account for human recall errors with longer memory delays.

Fixed neural signal

A recent computational study on forgetting in VWM proposed that diffusion is sufficient to explain memory dynamics over delay [10]. To test for this, we developed two reduced versions of the DyNR model in which the diffusion process was solely responsible for memory fidelity dynamics. In both variants, the sensory signal terminated abruptly with stimuli offset, so the VWM signal encoding the stimuli was independent of the delay duration and equal to the limit imposed by normalization ($\tilde{\gamma}_{wm}/N$). In the first variant, the diffusion rate was constant across set sizes, as in the full model. The formal model comparison demonstrated that the full DyNR model performed better than this simplified alternative ($\Delta AIC = 609.5$).

In the second variant, we allowed the diffusion rate to increase proportionally with set size (for a similar proposal see [64]). This model was again outperformed by the full DyNR model ($\Delta AIC = 666.4$). Critically, both models tested here failed to qualitatively reproduce the observed non-linear pattern of changes in recall error with time, notably overestimating recall error at the shortest delays by assuming no modulation in the representational signal (Fig. S3).

Diffusion

We developed two variants of the proposed neural model to test the role of diffusion. In the first variant, we completely omitted the diffusion process from the model to test whether the sensory signal modulation during the retention period is sufficient to explain temporal dynamics in recall fidelity. It could be argued that diffusion accounts for only minor changes in precision over brief delays as used here, and therefore adds unnecessary complexity to the proposed model without improving the fit substantially. However, the formal model comparison revealed that the full DyNR model provides a better fit to human recall error compared to the matching model without diffusion ($\Delta AIC = 17.9$).

The second variant was identical to the proposed model, except that we replaced the constant diffusion rate with a set-size-scaled diffusion rate by multiplying the right side of Eq 6 by N . The model comparison showed that the full DyNR model also outperformed this variant ($\Delta AIC = 29.8$). While both model variants qualitatively reproduced the increase in memory error with delay and set size, the pattern of variability was better explained by the model with a constant diffusion rate across set sizes. Although a more substantial diffusion effect could become apparent with longer delays than those used here, previous work demonstrated that noise-driven diffusion causes representations to deteriorate throughout the entire retention period [60].

Finally, we examined the role of diffusion with longer memory intervals in a separate experiment using variable set sizes and memory intervals (1 and 7 seconds) (for full details see Additional dataset 1 in Supplementary Information). We demonstrated that, once sensory information

decayed completely, an accumulation of error during retention interval accounted for continuing memory deterioration. Together, the results presented here corroborate findings on the role of diffusion in temporal dynamics of recall fidelity [9].

Experiment 2: Exposure duration

In Experiment 2, we evaluated the encoding phase of VWM, by testing recall of orientation stimuli displayed in arrays of variable size presented for variable durations. In the DyNR model, increasing the sensory evidence by prolonging stimulus presentation has a favorable effect on later recall of stimulus, as more of that evidence can be accumulated into VWM. Importantly, this accumulation is also capped by the VWM resources available to store it (Fig. 5).

Experimental data

Figure 6 shows the response error for different presentation durations and set sizes. Consistent with previous findings, response error can be seen to decrease with prolonged presentation duration, but increase as the number of items in memory increases. This was confirmed with a significant effect of display duration ($F_{(6,72)} = 29.01, p < .001, \eta^2 = .21$), set size ($F_{(2,24)} = 112.51, p < .001, \eta^2 = .54$), and their interaction ($F_{(12,144)} = 2.58, p = .004, \eta^2 = .019$).

We further explored this interaction by first confirming that response error decreased with display duration within each set size ($F_{(6,72)} \geq 10.24, p < .001, \eta^2 \geq .26$). A consistent pattern was observed across set sizes, comprising an initial rapid decrease in response error over the briefest presentation times (first 200 ms), followed by saturation at prolonged exposure durations. Next, we calculated the change in recall error between the longest and the shortest display exposure within each set size, revealing that response error decreased more rapidly with display time as the number of items in memory decreased (ANOVA: $F_{(2,24)} = 7.79, p = .002, \eta^2 = .21$; corrected pairwise comparisons: $t_{1-4} = 3.65, p = .016, d = 0.87, t_{4-10} = 0.96, p = .72, d = .27$).

These results reveal the time course of information accumulation into VWM and forming of stable representations. We again identified several key characteristics of the dynamics of recall fidelity in the data (Fig. 6B) to test against the DyNR model. Consistent with previous studies, we found recall fidelity changed with both presentation duration and the number of presented stimuli [38–40]. Specifically, as display duration increased from the shortest exposure, recall error showed an initial rapid decrease followed by a gradual levelling-off. As set size increased, the initial slope became shallower and the plateau occurred at a higher level of error.

Dynamic Neural Resource model

Curves in Figs. 6A & B show fits of the model with maximum likelihood (ML) parameters (mean $\pm$ SE: population gain $\gamma = 188.5 \pm 109.6$, tuning width $\kappa = 10.2 \pm 6.08$, sensory rise time constant $\tau_{\text{rise}} = 0.33 \pm 0.18$, sensory decay time constant $\tau_{\text{decay}} = 0.61 \pm 0.19$, VWM accumulation time constant $\tau_{\text{WVM}} = 0.8 \pm 0.34$, cue processing constant $b = 0.2 \text{ s} \pm .09 \text{ s}$, base diffusion $\sigma^2 = 0.28 \pm 0.08$, spatial uncertainty time constant $\tau_{\text{spatial}} = 0.013 \pm 0.004$, swap probability $p = 0.053 \pm 0.01$). The model provided an excellent quantitative fit to response distributions (Fig. 6A) and RMSE (Fig. 6B), successfully reproducing the pattern of changes with set size and presentation duration.

The model predicted that information from a visible stimulus accrues at a high rate immediately after the stimulus onset, consistent with observed changes in human recall error over stimulus durations up to 200 ms (Fig. 6). This initial high encoding rate emerges naturally in the model due to the joint dynamics of sensory and VWM populations. In the sensory population, a low-pass temporal filter serves as a neural gain control mechanism, attenuating neural response to transient changes in stimuli [44, 47]. As a consequence, the neural response to stimulus onset increases exponentially (Fig. 5). The information from sensory areas is accumulated into VWM, such that the accumulation rate is directly proportional to the difference between the current and

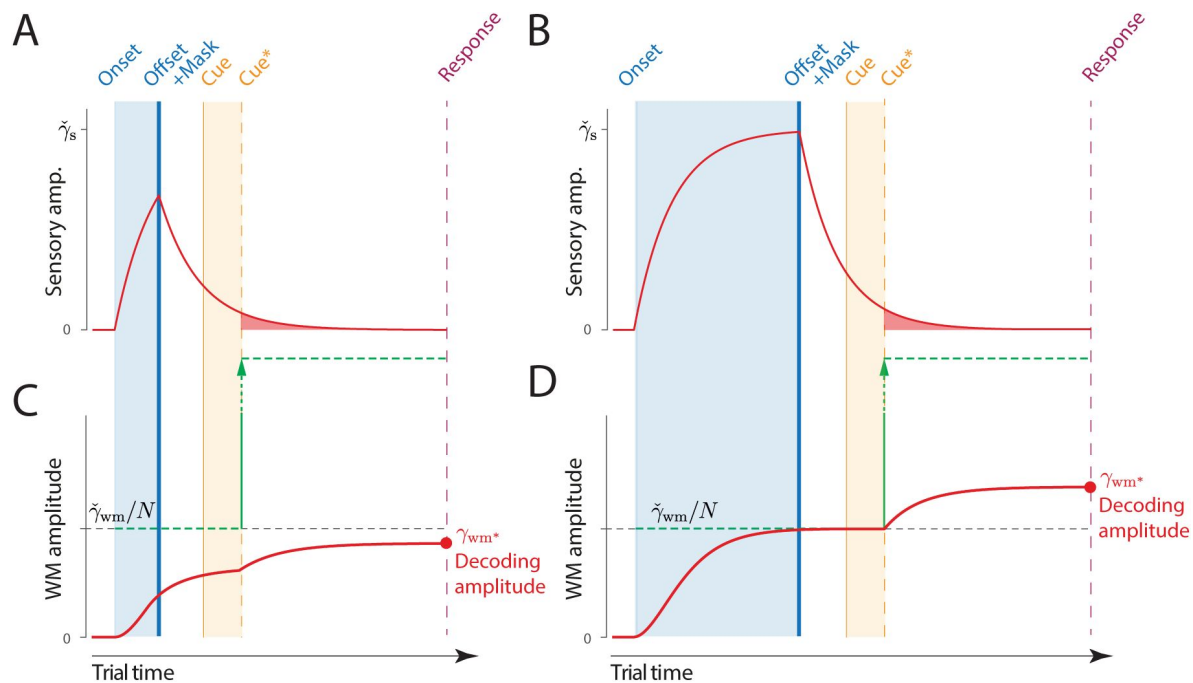


Figure 5

Time course of sensory and WM gain with variable exposure duration. **(A, B)** The signal amplitude in the sensory population increases from stimulus onset, exponentially approaching the maximum sensory activity (γ_s). For shorter presentation durations **(A)** the attained amplitude at stimulus offset is only a fraction of the maximum (compare **B**, late offset). Following offset, sensory areas produce a decaying neural response, that is curtailed (faster decay) but not abolished by a backward mask. **(C, D)** Information about the stimulus is accumulated in WM from sensory activity. A shorter presentation **(C)** provides less sensory evidence for the initial accumulation of all items into VWM (compare **D**, late offset), and subsequently less decaying sensory activity that can supplement VWM activity for the target item following the cue.

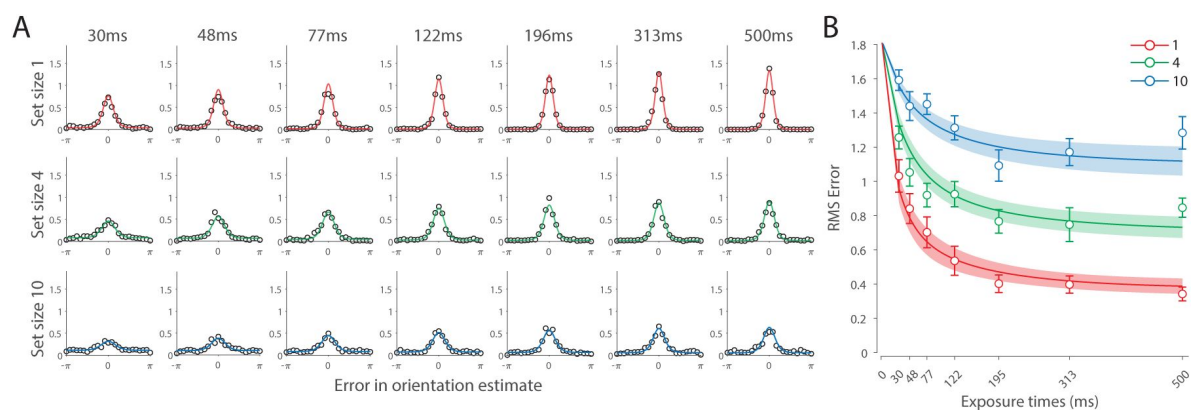


Figure 6

Experiment 2 results and modelling data show the consequences of varying set size and stimulus exposure time on VWM reproduction error. **(A)** Empirical recall error distributions (black circles) and the DyNR model fits (colored curves). Different panels correspond to different set sizes (rows) and exposure durations (columns). **(B)** Corresponding RMS errors from experimental data (circles and errorbars) and the DyNR model fits (curves and error patches). Error bars and patches indicate ± 1 SEM.

saturation state (i.e. the rate is faster when accumulated information is far from the saturation state). Therefore, dynamics in the sensory and VWM population jointly account for the initial high rate of information extraction from stimuli, and its dependence on set size.

After the initial steep change, the model predicts that recall fidelity will asymptote. This was again observed in human behavior (Fig. 6). Extending stimulus presentation beyond 200 ms had negligible impact on recall precision, consistent with previous studies [38]. The model explains this behavior by describing how sensory signal and VWM accumulation independently saturate with time (Fig. 5). Since the temporal filtering in the sensory population attenuates only high-frequency stimuli (i.e. very short presentations), with sufficient exposure, the sensory signal plateaus, resulting in a stable feed-forward input to VWM. Similarly, VWM signal strength is subject to limits determined by normalization. Once the accumulated information reaches the normalized maximum set by the number of objects in memory, further accumulation of sensory evidence is not possible. Following the cue, a portion of the resource is freed, allowing the target representation to be further strengthened. However, because the sensory signal plateaus at longer exposures, the information available for integration after the cue remains constant across the longer exposures, supplementing normalized VWM signal by the same amount. The result is a plateau in fidelity that varies with set size.

Model variants

We investigated whether post-stimulus sensory persistence contributed to the model fits in Experiment 2. We assumed that the signal persisting after stimulus offset would be impaired but not eliminated by the subsequent presentation of a noise mask in this experiment [52]. An alternative account suggests that the mask immediately terminates any stimulus-related signal. To test for this, we fit a variant of the DyNR model in which the sensory signal was terminated by the onset of the mask, providing a feed-forward signal to VWM only for the period of the stimulus presentation. We found that the proposed DyNR model, in which some sensory signal persists after the mask onset, gave a better account of the data than this model variant ($\Delta AIC = 446.67$). Although the alternative model captured the general pattern of changes in memory fidelity with exposure duration, it mispredicted fidelity at shorter exposures, in particular the effect of set size (Fig. S4A).

A testable prediction of this alternative model is that the memory fidelity at recall should obey the neural normalization principle because there was no additional signal to supplement the presentation after initial encoding. To test for this, we additionally fitted each exposure condition separately using the original Neural Resource model with only three parameters (i.e., neural gain, tuning width, and swap probability). This model failed to predict actual fidelity levels at recall (Fig. S4B), corroborating the findings of the model comparison.

Finally, to investigate the role of the post-stimulus sensory persistence on encoding dynamics, we fit the DyNR model to an additional dataset from Bays et al. [38] (for full details see Additional dataset 2 in Supplementary Information). This experiment aimed to investigate VWM dynamics during encoding, like our Experiment 2. In contrast to our Experiment 2, Bays et al. [38] used a much longer delay interval (1100 ms vs 100 ms), precluding the possibility of further accumulation of sensory evidence following the cue. We expected that the DyNR model could account for memory dynamics in this study without any post-stimulus sensory activity. This was confirmed by accurately reproducing memory dynamics with a model in which encoding into VWM relied only on sensory evidence during stimulus presentation (detailed results in Supplementary Information).

Alternative accounts

Having demonstrated the need for both post-stimulus sensory persistence and diffusion to account for empirical data, we next considered alternatives to our account of VWM accumulation and information read-out.

Direct read-out of sensory information

In the DyNR model, recall fidelity is enhanced following the cue by integrating remaining sensory activity into capacity-limited VWM. As a consequence, response precision is bounded from above by the memory limit irrespective of the available sensory signal. An alternative possibility is that the decaying sensory representation can be directly read out following the cue to inform a response, bypassing working memory limitations. To formalize this alternative model, we assumed that independent sensory and VWM representations would be optimally combined via summation of neural activity to yield population gain

$$\gamma_{\text{sum}}^* = \gamma_{\text{wm}}(t_{\text{cue}}^*) + \gamma_{\text{s}}(t_{\text{cue}}^*) \quad (14)$$

The model is otherwise identical to the proposed DyNR model. A distinctive prediction of this model is that the precision of recall changes exponentially with delay for every set size, including 1 item (**Fig. S5**). This prediction is qualitatively inconsistent with the pattern of results observed in Experiment 1, in contrast with the DyNR model which does not predict any beneficial effect of earlier cues with set size 1. This alternative model provided a worse fit to data from Experiment 1 ($\Delta\text{AIC} = 164$) and Experiment 2 ($\Delta\text{AIC} = 84.6$), for combined evidence favouring the DyNR model of $\Delta\text{AIC} = 248.6$.

Cue processing

In the DyNR model, we assumed that identifying the target stimulus based on the cue is time-consuming, and becomes more so as the number of alternatives increases. Cue processing time encompasses perceptual, attentional, and decision components needed to interpret and act on the cue. We tested the necessity of this component by fitting a model variant in which VWM started accumulating evidence about the cued item at the moment of cue presentation. This model provided a worse fit to empirical data from both Experiment 1 ($\Delta\text{AIC} = 84.5$) and Experiment 2 ($\Delta\text{AIC} = 107.5$), for total evidence in favor of the DyNR model of $\Delta\text{AIC} = 192$ (**Fig. S6**). We fit another variant in which cue processing time was constant across set sizes. This alternative provided a worse fit to the data in Experiment 1 ($\Delta\text{AIC} = 191.6$) and Experiment 2 ($\Delta\text{AIC} = 105$), for combined evidence $\Delta\text{AIC} = 296.6$ in favor of the full DyNR model that assumes cue processing time increases with set size. These results corroborate previous findings on the important role of cue processing time in models of attention [59] and IM [65].

Constant accumulation rate

In the DyNR model, the rate of accumulation into VWM is proportional to the difference between the present VWM amplitude and the maximum normalized amplitude (Eq. 2). An arguably simpler assumption is that the neural signal approaches saturation at a constant rate [66, 67]. In particular, the rate at which the signal representing an item is transferred to VWM is constant and depends only on the number of encoded items, i.e.

$$\dot{\gamma}_{\text{wm}}(t) = \begin{cases} \gamma_{\text{s}}(t)/(M(t)\tau_{\text{wm}}) & \text{if } \gamma_{\text{wm}}(t) < \tilde{\gamma}_{\text{wm}}/M(t) \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

The dependence on $M(t)$ satisfies the constraint that the neural resources in VWM are allocated at a constant rate, irrespective of the number of items. We applied this model to psychophysical data from both experiments (**Fig. S7**) and found it provides a worse fit to the data from Experiment 1

($\Delta\text{AIC} = 11.5$) and Experiment 2 ($\Delta\text{AIC} = 36.2$), for combined evidence favouring the DyNR model with exponential saturation ($\Delta\text{AIC} = 47.7$).

Discussion

In the present study, we investigated the temporal dynamics of short-term recall fidelity. We conducted two new human psychophysical experiments and analyzed two existing datasets in order to characterize how recall errors are influenced by set size, stimulus duration and retention interval. We developed a Dynamic Neural Resource (DyNR) model to provide a mechanistic explanation of the observed behavior, capturing not only changes in overall fidelity but also the distribution of errors in the stimulus space and frequencies of swaps (intrusion errors). A key finding is that the benefit to recall precision observed at very short delays is due to additional post-cue integration of sensory information into working memory, and that direct retrieval from sensory memory is unable to account for the empirical patterns of error.

Sensory and WM dynamics during delay

In the first experiment we investigated the effects of brief unfilled delays on recall fidelity. With multi-item arrays, we observed that memory performance deteriorates precipitously over the first few hundred milliseconds after stimuli disappear, followed by a gradual levelling-off of error with longer delays (**Fig. 4**). These results are consistent with previously reported patterns of memory dynamics [29–31, 33], and estimates of sensory decay ranging between 100 ms and 400 ms [68, 69]. Here, we shed new light on these results by taking a computational approach in explaining observed temporal dynamics, and asking what this superior recall's neural origin is and its relation with VWM. To answer these questions, we adapted the Neural Resource model of Bays [12] with a temporal component. The new DyNR model considers dynamics in a sensory neural population registering the stimuli and in a VWM population that stores the stimuli for later recall. Critically, our model assumes that objects encoded with limited precision into VWM can be flexibly supplemented with sensory activity following a recall cue, within a brief temporal window while the sensory population provides a feed-forward input post-stimulus. The boost in the representational VWM signal predicts a behavioral benefit of early cues that is consistent with our data and a large corpus of previous experiments [56].

A common assumption in studies of visual short-term memory is that recall over brief delays is exclusively supported by one of two memory stores, IM or VWM [31, 32]. In this account, a cue presented within the first few hundred milliseconds after stimulus offset allows observers to access high resolution but rapidly deteriorating representations in IM; once the information in IM has decayed, objects must be retrieved from the capacity-limited VWM store. Two pieces of evidence from the current study contradict this view and strongly suggest that recall depends on VWM from the moment objects disappear. First, the recall benefit of short delays was not observed for one item arrays. We propose that this behavior reflects the fact that, during encoding, the entirety of the VWM resource is allocated to a single object, leaving no free capacity for further enhancement based on the available sensory signal post-cue. Second, we found clear evidence that recall fidelity varied with set size even with no delay between stimulus offset and cue (0 ms condition). We argue that this arises from the set-size dependence of representational fidelity in VWM, which is only incompletely compensated by integration of the decaying sensory signal post-cue, resulting in lower fidelity for higher set sizes. The DyNR model provides a successful quantitative account for these findings, which are in clear contrast with the traditional view of IM.

The rapid changes in fidelity over short delays can be distinguished from dynamics over longer retention intervals. A number of recent studies have observed a slow deterioration of VWM precision over the course of prolonged retention [9, 23, 24, 70–72]. The causes of this deterioration are still contested, but growing evidence links this behavior to noise-driven

diffusion. At a mechanistic level, diffusion is considered a fundamental property of continuous attractor networks of the kind commonly associated with models of working memory [73, 74]. In such networks, memorized features are represented as persistent activity “bumps” in the network’s representational feature space. Over a memory delay, the activity bump is sustained by balanced excitatory and inhibitory connections, while stochasticity in neural activity causes shifts of the bump along the feature dimension, taking the form of a random walk. Although we did not model the network processes governing stability and diffusion within neural populations, our implementation is consistent with random (Brownian) perturbation, as assumed by attractor models (see also 9).

Our theoretical account of memory dynamics during delay differs from several existing models of forgetting, which emphasize diffusion as the dominant source of error in short-term memory (e.g., 10, 64). To solely account for the observed data in Experiment 1, diffusion would need to be strongest early in the retention period, followed by a much weaker diffusion with longer delays. However, it is unclear why the diffusion rate would change, and particularly slow down, with time. Assuming a constant neural signal encoding the stimulus, this would predict greater variability in neural activity initially compared to the later period after stimuli offset. This is inconsistent with electrophysiological data showing relatively stable levels of spiking variability throughout the memory delay period [75–77]. The results observed here are consistent with the proposal that modulation of neural signal over short memory intervals accounts for an abrupt change in response fidelity, while diffusion accounts for a slower change that grows with time.

In the present study, a model assuming a constant diffusion rate, independent of the stored number of items, was preferred to one in which diffusion rate increases linearly with set size. This is consistent with results of Shin et al. [71] who did not find a significant effect of set size on the rate of memory deterioration. In contrast to that, Koyluoglu et al. [64] recently proposed that the rate of diffusion scales with set size. However, this study did not account for the presence of swap errors, which we found to increase with retention interval as well as set size. To draw strong conclusions about the dependence of diffusion on set size would require a future study to disentangle the different sources of error that could, in principle, increase with delay.

Sensory and WM dynamics during encoding

Having investigated memory degradation during the retention interval, in Experiment 2 we focused on the dynamics arising from accumulation of information during stimulus presentation. Using new psychophysical data, we showed that encoding of information into VWM is contingent on both presentation duration and the number of memorized stimuli. The observed patterns of data indicate that VWM encoding of elementary stimuli is mostly completed within the first 200 ms of presentation even at the largest set sizes, with minimal benefit of longer exposures, extending previous work [38–40]. This fast encoding process may have an adaptive role: with a key function of VWM to store and accumulate information across saccadic eye movements, an efficient system should deploy its resources within the duration of a typical gaze fixation [78, 79].

Our aim was again to move beyond the description of the encoding dynamics and to provide a biologically plausible neurocomputational account of these dynamics. To achieve that, we applied the same VWM accumulation process that operates post-cue to the sensory information during stimulus presentation. Using previously published and newly collected data, we show that a model in which VWM accumulates dynamical sensory input up to a fidelity limit can successfully account for patterns of human recall errors with variable set size and stimulus presentation. An important result of our modelling is that the accumulated information in VWM increases with a rate proportional to unfilled capacity. In particular, the model with such exponential accumulation provided a better fit than a model assuming a constant encoding rate. This parallels previous observations that models based on exponential-like extraction of information successfully characterize attention [80, 81], working memory encoding [38, 82], memory updating

[83], and broader cognitive processes [84, 85]. We hypothesize that this pattern represents an approach to an equilibrium state of balanced excitation from the sensory input and lateral inhibition within the VWM population, which is the basis for capacity of the memory system.

In Experiment 2, the longest presentation duration shows an upward trend in error at set sizes 4 and 10. While this falls within the range of measurement error, it is also possible that this is a meaningful pattern arising from visual adaptation of the sensory signal, whereby neural populations reduce their activity after prolonged stimulation. This would mean less residual sensory signal would be available after the cue to supplement VWM activity, predicting a decline in fidelity at higher set sizes. Visual adaptation has previously been successfully accounted for by a type of delayed normalization model in which the sensory signal undergoes a series of linear and nonlinear transformations [86]. Such a model could in future be incorporated into DyNR and validated against psychophysical and neural data.

Our computational account of VWM encoding dynamics differs from several existing modelling frameworks aiming to explain similar data. For example, the Theory of Visual Attention (TVA; 80) assumes that visual stimuli participate in a parallel exponential race towards limited VWM. Like the DyNR model, TVA assumes a form of normalization in the sense that the speed with which items race towards VWM depends on the number of items in the visual field. Unlike our dynamic model, TVA is not a theory of VWM, and it considers VWM only as a storage for categorizations of visual objects. In particular, TVA takes into account the limits of VWM but does not specify why or how these limitations arise. Finally, TVA considers whether an object was selected for entry into VWM in an all-or-none fashion; our dynamic model is mostly concerned with the fidelity of representations. A somewhat alternative account of VWM encoding is provided by the Competitive Interaction Theory (CIT; 87), which is similarly based on the Signal Detection theory and principles of normalization [43]. Like TVA, CIT is mostly focused on item selection and merely incorporates a concept of VWM capacity derived from object-based models of VWM. Although CIT had success in accounting for behavioral data from a two-alternative orthogonal discrimination task using up to four items and a limited range of encoding times, it remains an open question whether this model can account for error distributions as measured in a continuous report task, and a larger range of set sizes and stimulus exposures. Importantly, compared to both TVA and CIT, the DyNR model is strongly rooted in and inspired by findings from neuroscience. This not only adds to the biological plausibility of our model but also allows future studies to test the model's predictions using physiological methods.

Neural mechanisms

The theory presented here generalizes the Neural Resource model of Bays [12], a simple encoding-decoding model in which visual features are represented in the noisy spiking activity of neural populations [15], and where the activity representing each feature scales inversely with the total number of representations, consistent with the prevalence of normalization mechanisms in the brain and observations from single-neuron recording [88] and fMRI decoding [89] studies. The population coding in the model is based on an abstract idealization of neural response functions. Nevertheless, it has recently been shown that more realistic population coding schemes that allow for heterogeneity in neural tuning curves and correlated spiking activity as observed in visual cortex, maintain the key predictions of the idealized model [90, 13]. This may be seen as a consequence of the different population codes inducing a common representational geometry [91].

We adapted the stationary VWM model by first incorporating a sensory population that provides an input drive to the VWM population. In parallel with neurophysiological observations, a common approach is to model these dynamics with a low-pass filter which acts like a neural gain modulation mechanism [47]. As a consequence, the sensory response to stimulus onset and offset is an exponential rise and decay in activity, respectively. The decaying component of the response has been recognized as a neural substrate of visual persistence and IM [37, 36].

Here, we modelled sensory decay with an exponential function [92], although other forms of decay have been proposed. For example, Loftus et al. [68] showed that iconic decay could be better captured using a gamma survival function, a generalization of exponential decay that could simply be implemented in our neural model by replacing a single filter with a cascade of exponential low-pass filters.

In addition to the dynamics in the sensory population, two features of VWM introduce additional dynamics in representation fidelity: the accumulation of information (discussed above) and the diffusion of representations owing to accumulated noise. Although we did not aim to model the neural processes behind diffusion, our implementation is consistent with the consequences of neural variability in attractor networks [25, 74]. Converging neural evidence demonstrating such diffusion has been observed using single-unit neural recording in monkeys [26], as well as EEG [28] and fMRI [27, 93] studies in humans.

As well as being implicated in higher cognitive processes including VWM [88, 89], divisive normalization has been shown to be widespread in basic sensory processing [94–96]. The DyNR model presently incorporates the former but not the latter type of normalization. While the data observed in our experiments do not provide evidence for normalization of sensory signals (note comparable recall errors across set size in the simultaneous cue condition of Experiment 1), this may be because sensory suppressive effects are localized and our stimuli were relatively widely separated in the visual field: future research could explore the consequences of sensory normalization for recall from VWM using, e.g., centre-surround stimuli [97].

Following onset of a stimulus, the visual signal ascends through visual areas via a cascade of feedforward connections. This feedforward sweep conveys sensory information that persists during stimulus presentation and briefly after it disappears [98]. Simultaneously, reciprocal feedback connections carry higher-order information back towards antecedent cortical areas [99]. In our psychophysical task, feedback connections likely play a critical role in orienting attention towards the cued item, facilitating the extraction of persisting sensory signals, and potentially signalling continuous information on the available resources for VWM encoding. While our computational study does not address the nature of these feedforward and feedback signals, a challenge for future research is to describe the relative contributions of these signals in mediating transmission of information between sensory and working memory [100].

Our model makes a clear distinction between dynamics in sensory and VWM populations, however, it remains agnostic as to whether the populations have the same or different anatomical locus [101]. Albeit inspired by the properties of orientation-selective neurons in area V1, population tuning of this kind is a common coding motif across the brain [15]. While it could be considered efficient to use already specialized circuits to maintain as well as process visual information, it is still debated whether sensory areas are a feasible candidate for memory storage [102, 103]. While some studies have focused on prefrontal [104], parietal [105] or occipital [106] cortices as the primary locus of VWM, others argue for distributed storage by demonstrating that VWM contents can be decoded from imaging signals originating in multiple brain areas [107].

Representational dynamics of cue-dimension features

Memory retrieval failures in which a non-cued item is reported in place of the intended target represent an important source of error in VWM recall. These swap errors occur more often at higher set sizes and when spatial confusability is high [108–111], as predicted by models in which they arise from uncertainty in the recall of cue-dimension features leading to incorrect selection of an item in memory [20, 61]. In the current study, we assumed memory for spatial location (the cue feature) undergoes similar dynamics to memory for orientation (the report feature), and in particular that spatial information degrades with retention time [9], leading to changes in swap error frequency with delay interval. Similarly, during encoding the fidelity of

spatial representation increases with the accumulation of sensory evidence [112], reducing the uncertainty at retrieval and consequently swap errors at longer stimulus exposure. Although we did not explicitly model the neural signals representing location, the modelled dynamics in the probability of swap errors were consistent with those of the primary memory feature. We provided a more detailed neural account of swap errors in our earlier works that is theoretically compatible with the DyNR model [20, 61].

The DyNR model successfully captured the observed pattern of swap frequencies (intrusion errors). The only notable discrepancy between DyNR and the three-component mixture model (Fig. S2) arises with the largest set size and longest delay, although with considerable interindividual variability. As the variability in report-dimension increases, the estimates of swap frequency become more variable due to the growing overlap between the probability distributions of swap and non-swap responses. This may explain apparent deviations from the modelled swap frequencies with the highest set size and longest delay where orientation response variability was greatest.

Removal of information from WM

In the DyNR model, taking advantage of early cues requires rapid removal of the VWM signal associated with uncued items, to admit further accumulation of activity encoding the cued item. To achieve this, an active process of selective content elimination may be required [57], as opposed to a passive decay of uncued representations during the post-cue interval. Mounting evidence for such active removal has been provided at the behavioral [113] and neural [114] level. Importantly, studies show that a functional role of such active removal is to release resources allocated to the uncued representations, facilitating the encoding of new information [115, 116]. The fast reallocation of neural resources assumed by the DyNR model is consistent with such a description of active removal.

Methods

Participants

A total of twenty-three naive observers (12 females, 11 males; aged 18–34) took part in the study after giving informed consent in accordance with the Declaration of Helsinki. Ten observers participated in Experiment 1 and thirteen observers participated in Experiment 2. Volunteers were recruited through the Cambridge Psychology research sign-up system. All observers reported normal color vision and normal or corrected-to-normal visual acuity, and were remunerated £10/hr for their participation.

General methods

Experimental setup

Stimuli were presented on a 69 cm gamma-corrected LCD monitor with a refresh rate of 144 Hz. Participants were seated in a dark room and viewed the monitor at a distance of 60 cm, with their head supported by a forehead and chin rest. Responses were collected using Magic Trackpad 2, a pointing device (16 x 11.5 cm) with a tactile sensor operating at ~90 Hz (Apple Inc.). Eye position was monitored online at 1000 Hz using an infrared eye tracker (SR Research). Stimulus presentation and response registration were controlled by a script written in Psychtoolbox and run using Matlab (The Mathworks Inc.).

Stimuli

Memory stimuli consisted of randomly oriented Gabor patches (wavelength of the sinusoid, 0.65° of visual angle; s.d. of Gaussian envelope, 0.5°) presented on a uniform mid-grey background. The contrast of Gabor patches varied between experiments (see below). Memory stimulus positions were randomly chosen from a set of ten equidistant locations on the perimeter of an invisible circle with radius 6° centered at fixation. At the start of each trial, a black fixation annulus was shown ($r = 0.15^\circ$ and $R = 0.25^\circ$) in the display center. Once steady fixation was registered, the size of the inner radius increased ($r = 0.2^\circ$). Observers perceived this change as the annulus becoming thinner. The fixation annulus then stayed visible throughout the trial. Items were cued for recall by displaying a black arrow (2° length) extending from the center of the display and pointing to one of the previously occupied locations without overlapping with it.

Procedure

Each trial started with presentation of the central fixation annulus. Observers were required to maintain gaze fixation for 500 ms within a radius of 2° around the central annulus in order for a trial to proceed. Following stable fixation, the appearance of the fixation annulus changed, indicating that the memory array would appear in 500 ms. The memory sample array consisting of 1, 4, or 10 randomly oriented Gabor patches was then presented. This was followed by a delay period and finally a cue display, indicating to observers to report the memorized orientation of an item previously displayed at the indicated location.

Observers were instructed to reproduce the remembered orientation as accurately and as quickly as possible by executing a single movement of their index fingertip over the surface of the touchpad located centrally in front of them. Simultaneously with the observer's movement, a blue line appeared on the screen, extending from the center of the screen and mimicking the observer's response in real-time. The response was terminated if one of the following conditions was satisfied: the observer stopped movement for 500 ms; the observer lifted their finger from the touchpad; or the response line reached the edge of the display. This was followed by a feedback display, consisting of the actual orientation (shown with a white line) and reported orientation (shown with a blue line) overlaid at the location of the cued item. The recalled orientation was calculated as the angle of the line connecting a starting point and an endpoint of hand movement on the touchpad.

Observers were required to maintain central fixation during the stimulus presentation and delay phase. If gaze position deviated by more than 2° a message appeared on the screen, and the trial was aborted and restarted with newly randomized orientations. Participants completed the task in blocks of 50 trials, and each block corresponded to one experimental condition. The order of blocks was randomized for every observer. At the beginning of the testing session observers familiarized themselves with the task and experimental setup by doing at most 50 practice trials.

Experiment 1

In Experiment 1 we investigated the temporal dynamics of VWM fidelity over short delays by presenting observers with sets of stimuli of variable size and then cueing one of them for recall after a variable delay relative to the stimuli offset. A typical trial sequence is shown in **Figure 3A**. The memory sample array (Michelson contrast = 0.5) was presented for 200 ms. In 50% of trials, the stimuli changed phase (by 180°) and contrast (Michelson contrast = 1) for the last 50 ms of presentation, while remaining at the same orientation. This manipulation was intended to minimize retinal after-effects (see e.g. 117 for similar techniques). The stimuli offset was followed by a variable blank delay of 0, 100, 200, 400, or 1000 ms, after which one item was cued for recall. In one additional condition, the cue was instead presented simultaneously with the memory sample array, indicating an item while it was still visible on the screen (**Fig. 3B**).

Each observer completed a total of 1800 trials, split into 36 blocks. The experiment was organized such that half of the observers first completed 18 blocks with phase shift (see above), and the other half first completed blocks without phase shift. Except for this constraint, block order was randomized for every observer. The testing was divided into four equal testing sessions, each lasting approximately 1.5 hours, with a separation of at least one day between sessions.

Experiment 2

In Experiment 2 we investigated the temporal dynamics of VWM fidelity during encoding. To this end, we displayed oriented stimuli for a variable duration and in sets of variable size. The experiment was similar to the previous experiment with a few exceptions (**Fig. 3C**). Each trial started with a presentation of a fixation annulus, followed by a memory array (Michelson contrast = 0.3). The stimuli stayed on the screen for a variable duration of 30, 48, 77, 122, 196, 313, or 500 ms, and were then replaced by noise masks (100 ms). Mask stimuli consisted of white noise at full contrast, windowed with a Gaussian envelope (0.5° s.d.) and flickering at 35 Hz. At the offset of the masking stimuli, one memory item was cued for recall. Each observer completed 21 blocks, for a total of 1050 trials. Blocks were spread over two testing sessions, each lasting approximately 1.5 hours, and taking place on different days. Observers completed 10 blocks in the first, and the remaining 11 blocks in the second session.

Data Availability

Data and code related to this study will be made available at <https://doi.org/10.17863/CAM.95223>.

Acknowledgements

We thank George Sperling and Sebastian Schneegans for helpful discussion, Robert Taylor for help with Bayesian hierarchical modelling, and Jessica McMaster for help with data collection. We used resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service. This research was supported by the Wellcome Trust (grant 106926 to PMB).

Supplementary information

Minimizing retinal after-effects

We assessed the method of minimizing retinal afterimages by repeating all measurements, with the exception of not using phase shift of stimuli (**Fig 3A**). We predicted retinal afterimages could serve as an additional source of information, but only for a brief period after stimuli offset. Therefore, here we expected to see better performance for brief delays compared to conditions with phase shift. **Figure S1A** shows recall error increased with both set size and delay. Both of these effects were statistically significant, as well as their interaction (set size: $F_{(2,18)} = 47.3$, $p < .001$, $\eta^2 = .31$; delay time: $F_{(5,45)} = 48.4$, $p < .001$, $\eta^2 = .26$; interaction: $F_{(10,90)} = 21.3$, $p < .001$, $\eta^2 = .14$), reminiscent of findings for data with phase shift.

Next, we focused on the comparison of conditions with and without phase shift of stimuli (**Fig S1B**). We illustrate the difference in performance by subtracting RMSE obtained in the condition without phase shift (**Fig 4B**) from RMSE shown in **Figure S1A**. Negative values indicate better performance in a condition without phase shift. As predicted, the overall pattern of data suggested performance was comparable for 1 item across all delays, and for all set sizes for extreme delays (simultaneous presentation and 1000 ms), indicated by the difference values around 0. We confirmed the difference in recall error for 1 item across all delays did not differ consistently with

and without phase shift, as neither phase shift ($F_{(1,9)} = 0.03$, $p = .86$, $\eta^2 < .001$, $BF_{incl} = 0.143$) nor the interaction of phase shift and delay ($F_{(5,45)} = 0.41$, $p = .89$, $\eta^2 = .00$, $BF_{incl} = 0.042$) reached significance. Based on this result, we conducted all remaining analyses using only the remaining two set sizes. We ran separate repeated measures ANOVAs for each delay using phase shift and set size as factors. The pattern of results we observed was clear: performance was comparable with and without phase shift with the simultaneous presentation and 1000 ms delay (phase shift, $F_{(1,9)} \leq 1.08$, $p \geq .33$, $\eta^2 \leq .002$, $BF_{excl} \geq 3.62$; interaction, $F_{(2,18)} \leq 0.8$, $p \geq .44$, $\eta^2 \leq .02$, $BF_{excl} \geq 3.39$), while for the remaining intermediate delays recall error was consistently lower when phase shift was omitted (phase shift, $F_{(1,9)} \geq 5.8$, $p \leq .039$, $\eta^2 \geq .06$; interaction, $F_{(1,9)} \leq 2.8$, $p \geq .13$, $\eta^2 \leq .001$).

Taken together, performance with and without phase shift of stimuli was comparable in perceptual condition (simultaneous presentation) and with the longest delay, suggesting phase shift did not change visibility or encoding of information into VWM. In contrast, we found strong evidence that observers had access to an additional source of information over intermediate delays when phase shift was not used, demonstrated by a better recall performance from 0 ms to 400 ms delay. Specifically, this source of information was available immediately after stimuli offset and was short-lived, consistent with the theoretical description of retinal afterimages [118].

Swap error estimates

Alternative models' fits

Additional dataset 1

To further investigate the role of diffusion in memory dynamics, we analysed an additional dataset collected in our lab [62]. In this experiment we varied the set size and delay duration similar to Experiment 1. In contrast to Experiment 1, we used longer memory delays, which allowed us to examine the diffusion mechanism on a more suitable time scale. Moreover, memory delays used in this study are out of reach of the decaying sensory information, enabling us to investigate the diffusion without changes in the neural signal strength post-cue.

Methods

Ten observers (6 females, 4 males, aged 18-34) took part in this experiment. The data for this experiment was collected using the same equipment and the testing setting as described for the main experiments. A typical trial sequence is illustrated in Fig. S8. Each trial began with the presentation of a central annulus which served as a fixation point. Once a stable fixation was achieved, the inner annulus radius changed indicating that stimuli would appear in 500 ms. The memory sample array was then presented for a duration of 500 ms. The array consisted of one or three randomly oriented black bars (length 2.8°). Each bar was positioned in one of six predetermined locations equally distributed around the circle with a radius of 5° around center of the screen. Each bar was presented along with a placeholder circle (radius 1.5°).

Memory array presentation was followed by a memory delay during which fixation circle and placeholders stayed visible. The retention interval was either 1 or 7 seconds long. After that, one stimulus was randomly cued for recall. The cue consisted of a second, larger circle drawn around one of the placeholders. Observers were instructed to start rotating a response dial (Griffin Technology PowerMate USB) once they were ready to respond. After the rotation of the response dial was detected, a randomly oriented black bar was displayed within the placeholder. Observers were instructed to rotate the dial until the displayed bar matched the remembered orientation of the cued item. Observers confirmed their response by pressing the dial. Trials with different set sizes and delay durations were randomly interleaved.

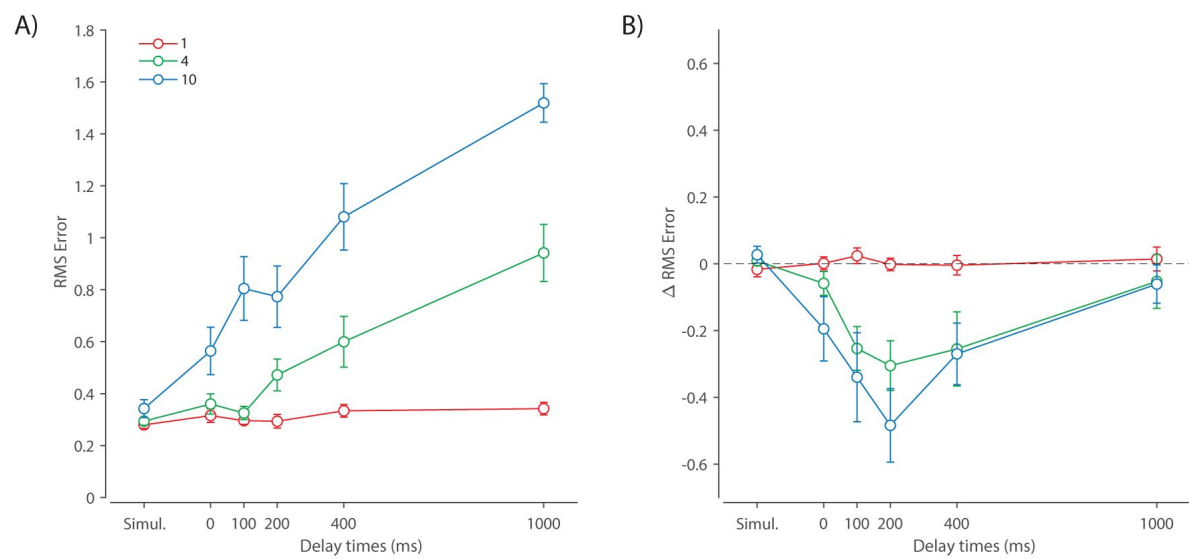


Figure S1

(A) Experiment 1 RMSE for trials without phase shift. (B) Differences in RMSE between trials with and without phase shift across set size and delay conditions. Negative values indicate better performance in the condition without phase shift.

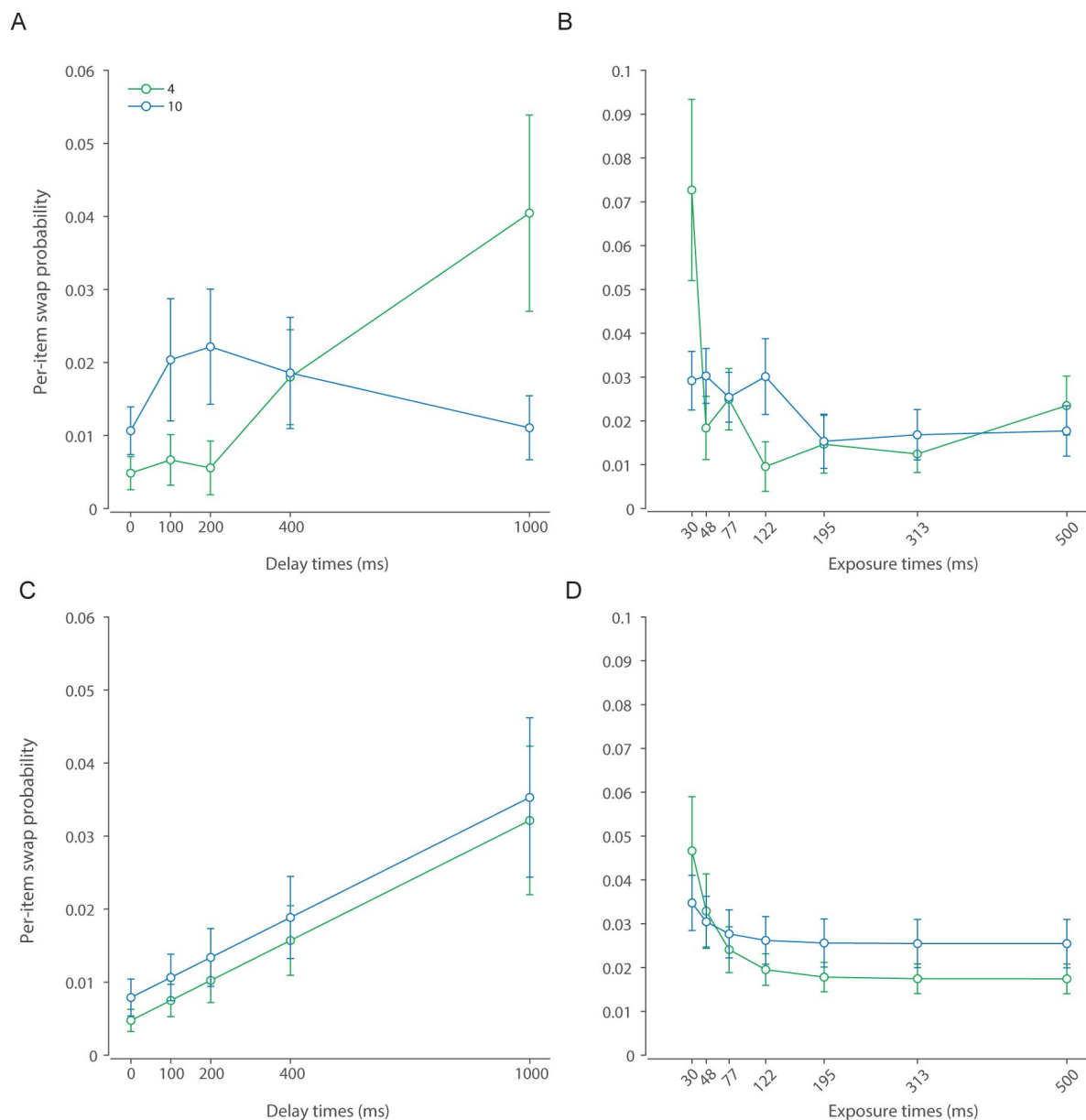


Figure S2

Swap error estimates. **(A&B)** Probability of swap errors estimated from empirical data using the three-component mixture model [108] in Experiment 1 (A) and Experiment 2 (B). **(C&D)** Probability of swap errors in best-fitting DyNR model in Experiment 1 (C) and Experiment 2 (D).

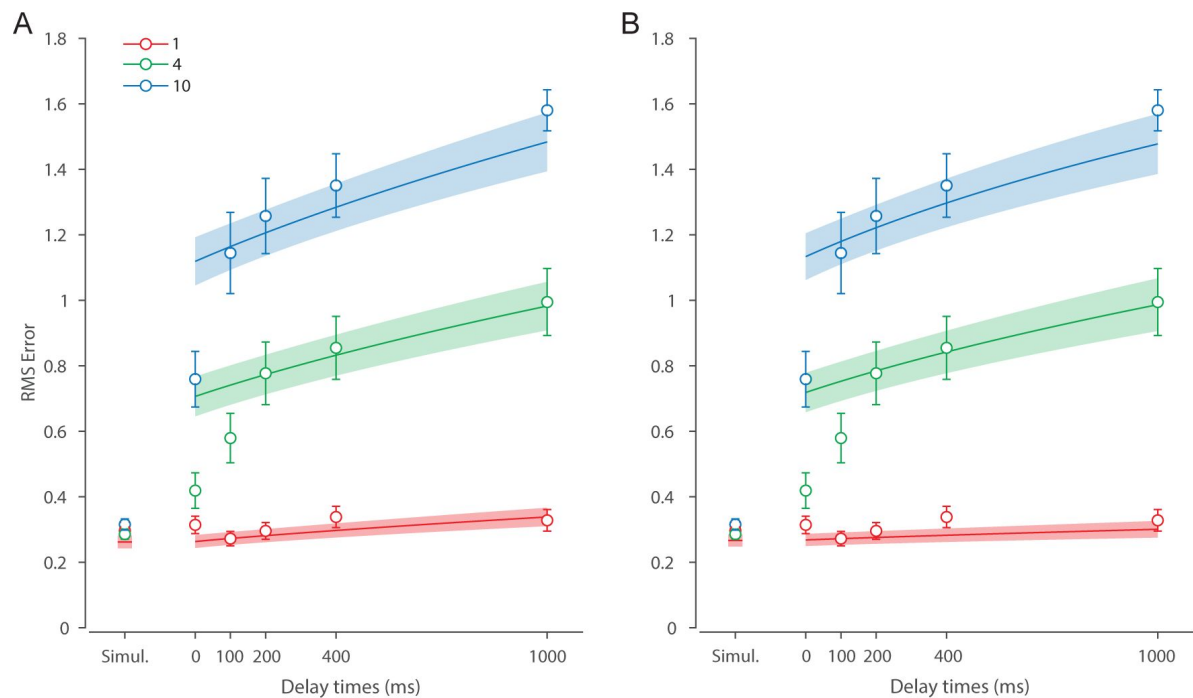


Figure S3

Experiment 1 behavioral data and model fit for the DyNR model without sensory persistence after stimulus offset. **(A)** A version of the DyNR model with equal diffusion across set sizes. **(B)** A version of the DyNR model with diffusion that scales with set size.

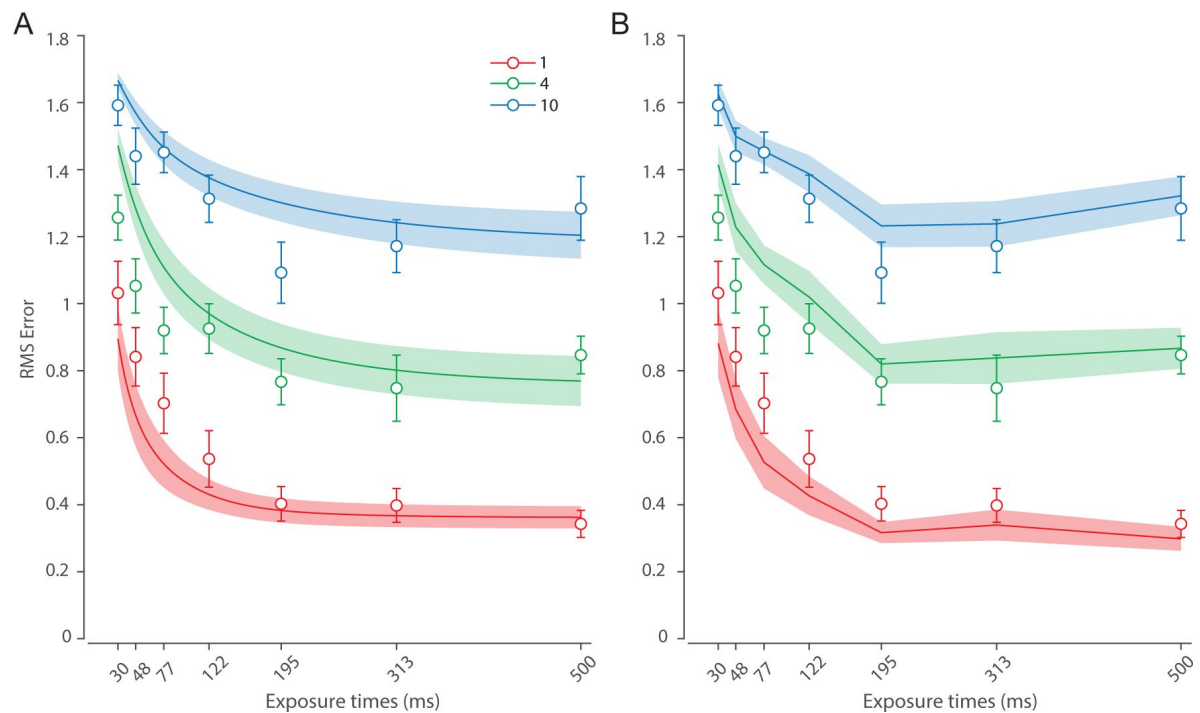


Figure S4

Experiment 2 behavioral data and model fit for the neural model without sensory persistence after stimulus offset. **(A)** A version of the DyNR model without sensory persistence. **(B)** Separate fits of the simplified neural model to each exposure time.

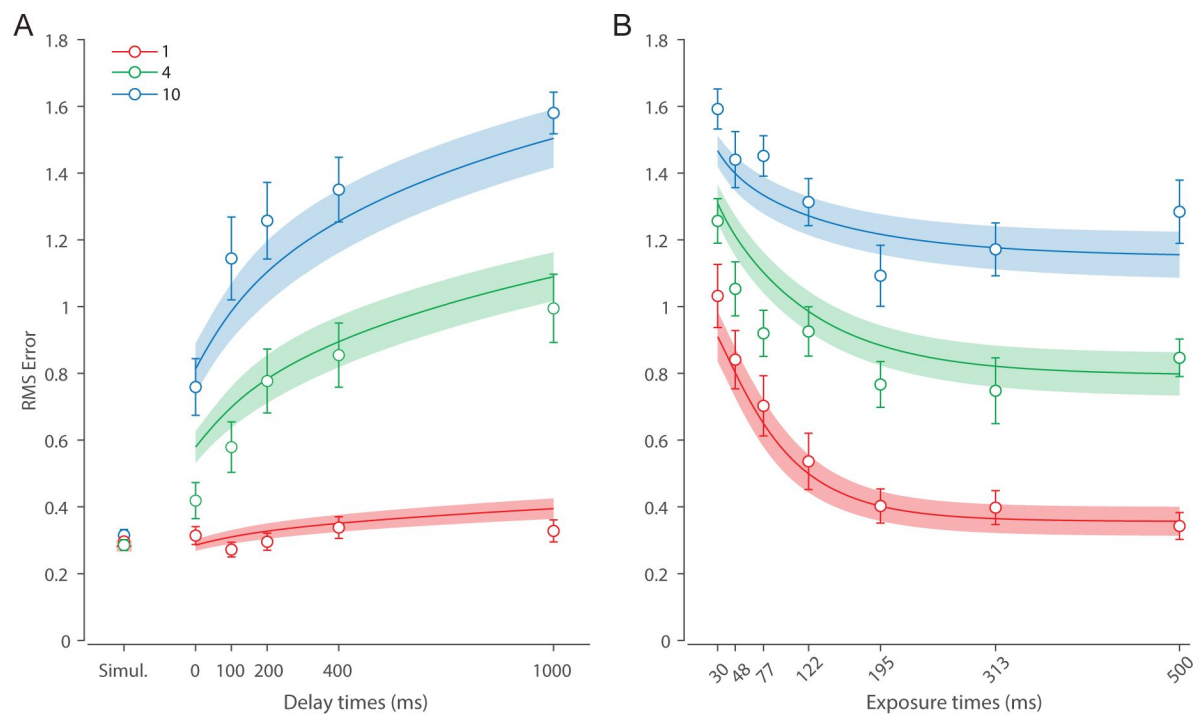


Figure S5

Behavioral data and model fit for a neural model with the direct read-out of information from sensory memory for (A) Experiment 1 and (B) Experiment 2.

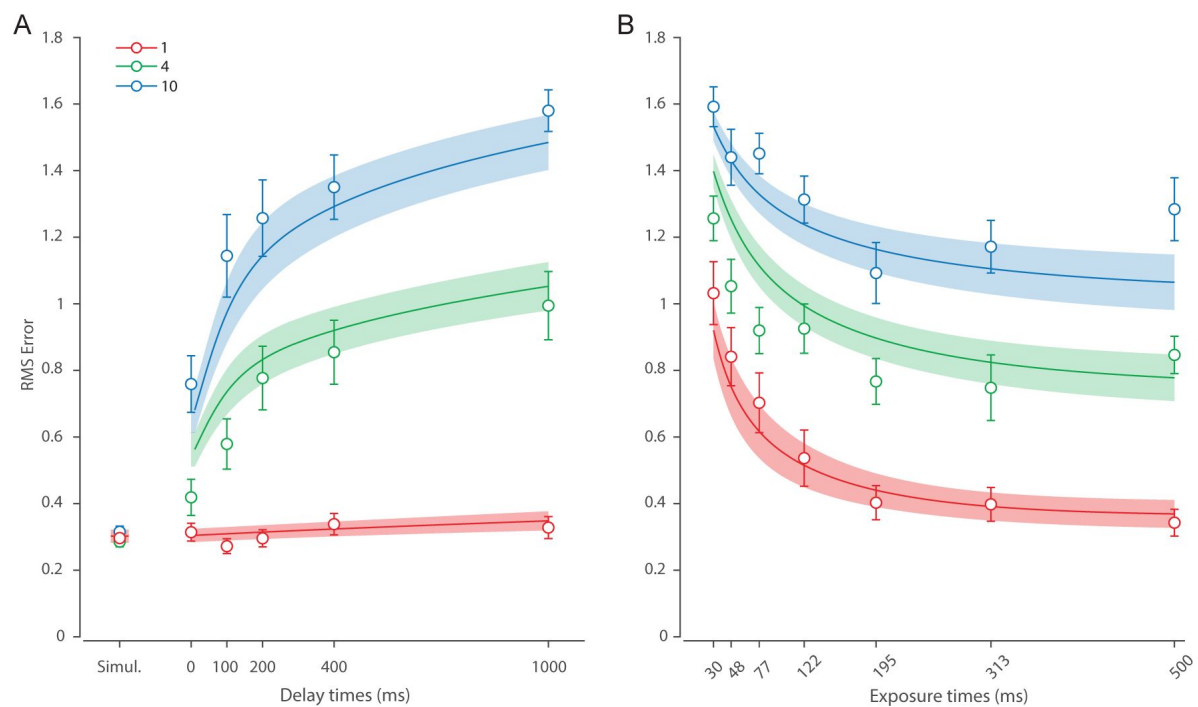


Figure S6

Behavioral data and model fit for the DyNR model without the cue processing time for (A) Experiment 1 and (B) Experiment 2.

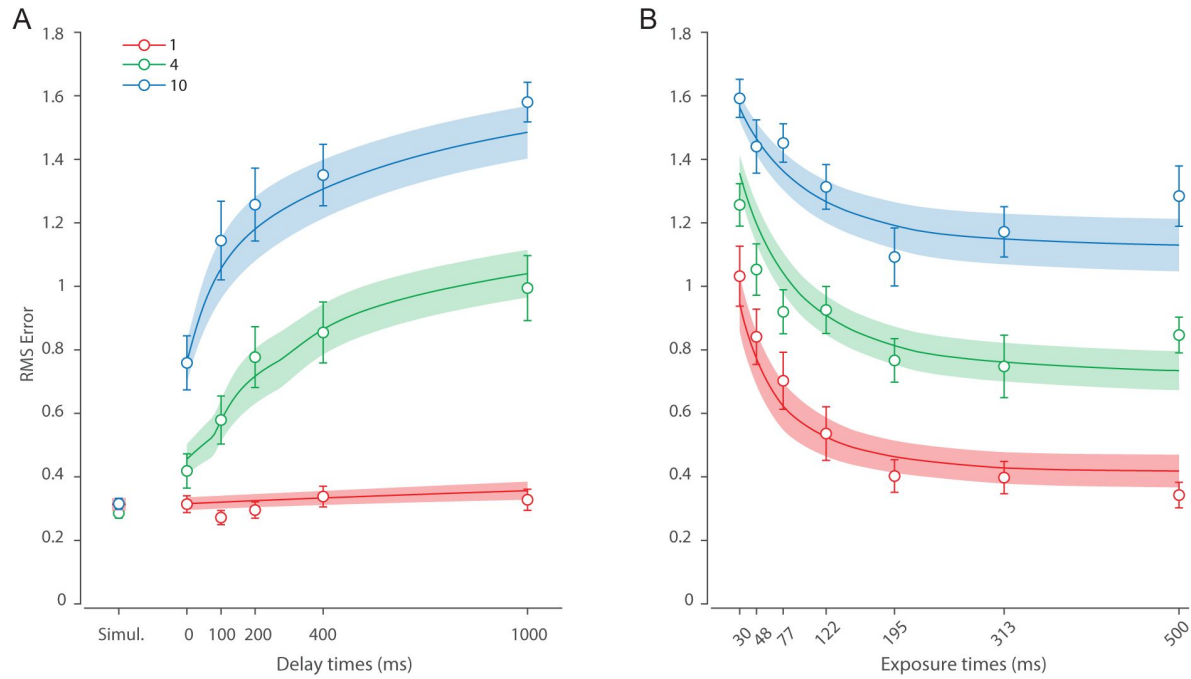


Figure S7

Behavioral data and model fit for a neural model with constant accumulation of information into WM for **(A)** Experiment 1 and **(B)** Experiment 2.

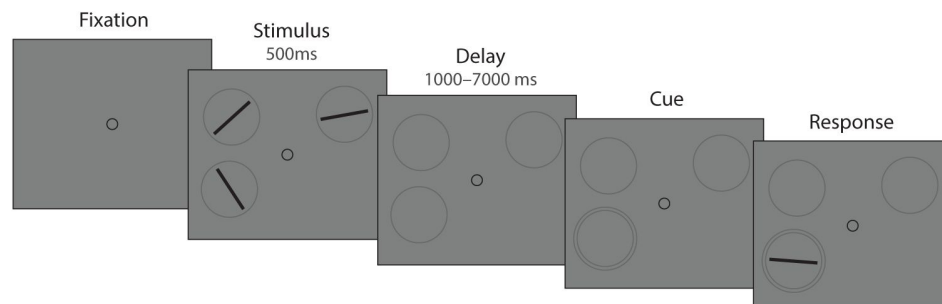


Figure S8

Experimental procedure. Stimuli are not drawn to scale.

Eye movements were monitored from the beginning of the trial until stimuli offset, and observers were required to hold steady fixation during that period. If the gaze position deviated by more than 2° a message appeared on the screen and the trial was aborted and restarted with new orientations. Each observer completed 700 trials, divided into two sessions and each consisting of 7 equal blocks.

Two sessions were separated by at least one day, and each lasted approximately 1 hour. At the beginning of each session observers familiarized themselves with the task and experimental setup by doing at most 50 practice trials.

Results

Behavioral data

Recall performance is shown in **Figure S9**. As predicted, response error increased with set size and memory delay. A repeated measures ANOVA revealed a significant effect of set size ($F_{(1,9)} = 111.17, p < .001, \eta^2 = .76$) and memory interval ($F_{(1,9)} = 58.14, p < .001, \eta^2 = .12$), and their interaction ($F_{(1,9)} = 10.66, p = .01, \eta^2 = .02$) on response error. Moreover, conducting paired t-tests within each set size revealed recall error increased with the delay with set size 1 ($t_{(9)} = 5.83, p < .001, d = 1.84$) and set size 3 ($t_{(9)} = 5.78, p < .001, d = 1.83$). The interaction effect was a consequence of a larger increase in error with delay for set size 3 compared to set size 1 ($\Delta\text{RMSE} = \text{RMSE}_{700\text{ms}} - \text{RMSE}_{100\text{ms}}; t_{(9)} = 3.27, p = .01, d = 1.03$). These results are consistent with Experiment 1, corroborating our finding that increasing the set size and delay time have a disadvantageous effect on memory fidelity.

Neural model

We fitted the DyNR model to the data to test whether noise-driven diffusion is sufficient to account for changes in recall fidelity with longer memory intervals. We applied a simplified version of the model without sensory decay and VWM accumulation components. This was justified given that estimate of sensory decay from Experiment 1 was shorter (mean life $\tau = 0.21$) than the shortest interval used in this experiment (1 s). Moreover, based on our findings in Experiment 2, we argue that a display duration of 500 ms is sufficient to fully encode objects into VWM.

Curves in **Figure S9** show fits of the model with maximum likelihood (ML) parameters (mean $\pm$ se: population gain $\gamma = 385.02 \pm 208.3$, tuning width $\kappa = 2.67 \pm 0.43$, cue processing constant $b = 0.68 \pm .67$, base diffusion $\sigma^2 = 0.009 \pm 0.001$, swap probability $p = 0.005 \pm 0.002$). The model provided an excellent quantitative fit to response distributions and summary statistics (**Fig. S9**), successfully explaining the adverse effects of set size and memory interval on recall fidelity. Critically, and consistent with results from Experiment 1, the proposed DyNR model provided a better fit to human response error compared to the matching model without diffusion ($\Delta\text{AIC} = 144.75$) or the model in which diffusion rate increases with set size ($\Delta\text{AIC} = 42.3$). In conclusion, this result shows that variability in representations over longer memory intervals can be fully accounted for by noise-driven accumulation without changes in the representational signal [9, 10, 28].

Additional dataset 2

To further validate predictions of the DyNR model we fitted it to an existing working memory study (Experiment 1 in 38). This study focused on the role of temporal dynamics during WM encoding, thereby addressing the same question as our Experiment 2. In contrast to our Experiment 2, Bays et al. [38] used a longer delay period (1100 ms), precluding the strengthening influence of decaying sensory information on recall. This dataset therefore isolates the initial information accumulation process during stimuli presentation.

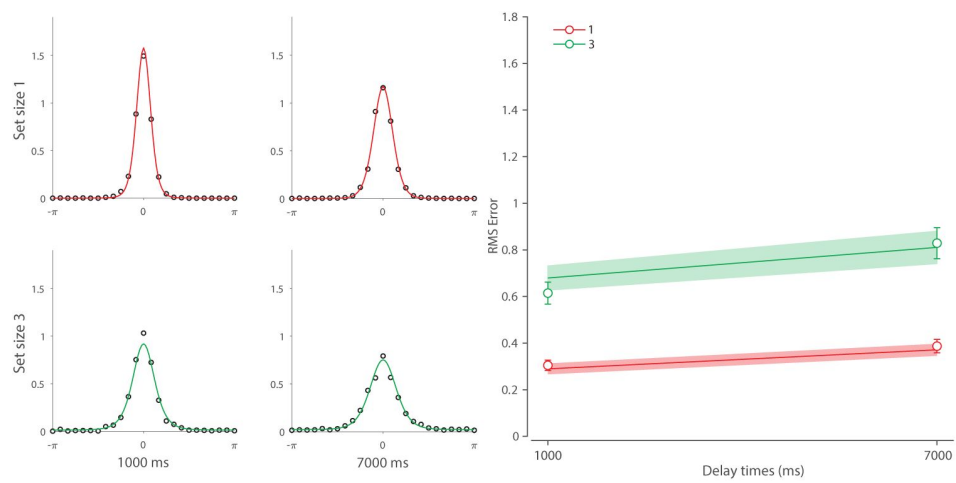


Figure S9

Behavioral data and model fit for Experiment 1a

Methods

The observers ($N = 32$) performed a continuous report task in which a variable number of oriented bars was presented for a variable duration, followed by a pattern mask (100 ms) and a 1-second delay period after which one of the items was probed for recall. Set size was manipulated between observers and exposure duration was manipulated within observers. Each observer performed 100 trials per exposure duration, for a total of 25600 trials in the study. A more detailed description of the experiment is provided in Bays et al. [38].

Analysis

Considering only exposure duration in this experiment was manipulated at the observer level, we decided to expand our modelling approach by employing a Bayesian hierarchical method as a compromise between fitting the data for each observer (i.e., set size) independently and pooling the data across all observers. Using a Bayesian hierarchical modelling, individual-observer parameters are considered samples from population distributions, whose means and variances are estimated based on all available data. In general, this approach has a desirable characteristic of constraining individual-level parameters with the population-level distribution and producing meaningful parameter estimates when a model is fitted across separate groups. The dynamic neural model fitted to the data is identical to the model fitted in Experiment 2, with the exception that here we assumed any existing post-stimulus sensory activity completely diminished by the time of the cue (1100 ms post-stimulus offset), and therefore we did not model sensory decay here. To obtain the hierarchical fit, we used the Differential Evolution Markov Chain algorithm [119]. All individual-level parameters were samples drawn from normal (i.e., Gaussian) distributions, with corresponding mean and standard deviation being constrained by uniform hyperprior distributions. We collected 240000 post-warmup samples across 12 chains and computed median and 95% equal-tailed intervals (ETI) of posterior distributions to obtain the group and individual-level parameter estimates. Prior specifications and empirical data for all analyses can be found along with the published code.

Results

Figure S10 and **Figure S11** show empirical distributions and summary statistics across all conditions. Similar to Experiment 2, increasing the exposure duration ($F_{(7,196)} = 110.9$, $p < .001$, $\eta^2 = .188$) and decreasing the set size ($F_{(3,28)} = 22.83$, $p < .001$, $\eta^2 = .53$) had beneficial effect on response error. Interaction of exposure duration and set size was significant ($F_{(21,196)} = 3.13$, $p < .001$, $\eta^2 = .02$). Critically, the pattern of memory fidelity dynamics largely matches the pattern observed in Experiment 2, with response errors decreasing rapidly as presentation duration was increased from the minimum duration, saturating at longer durations. This pattern was consistent across all set sizes, which only differed in the absolute error.

These dynamics were accurately predicted by the DyNR model, both at the level of response distributions (curves in **Fig. S10**) and summary statistics (curves in **Fig. S11**). The parameters used to generate model predictions were obtained by taking the individual observer's posterior medians. We observed the following hyperparameters (median and 95% ETI of hyperposterior): population gain $\gamma = 109.47$ (88.1 - 133.57), tuning width $\kappa = 3.23$ (2.6 - 4.03), sensory rise time constant $\tau_{\text{rise}} = 0.0049$ (0.0019 - 0.0091), VWM accumulation time constant $\tau_{\text{WM}} = 0.067 \pm (0.051 - 0.087)$, cue processing constant $b = 0.423$ (0.093 - 0.8436), base diffusion $\sigma^2 = 0.095$ (0.057 - 0.149), spatial uncertainty time constant $\tau_{\text{spatial}} = 0.031$ (0.022 - 0.041), swap probability $p = 0.02$ (0.011 - 0.034).

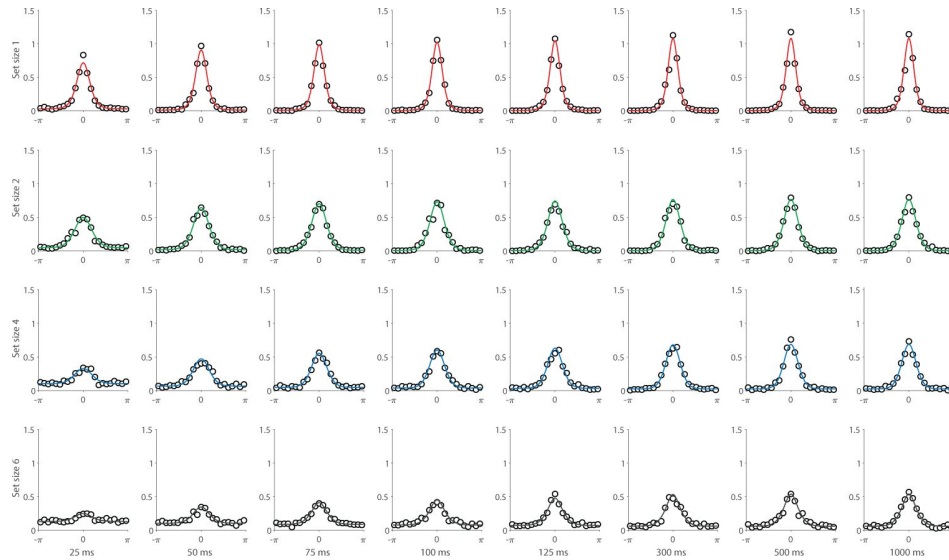


Figure S10

Empirical recall error distributions (black circles) from Experiment 1 in Bays et al. [38] and the DyNR model fits to the data (colored curves).

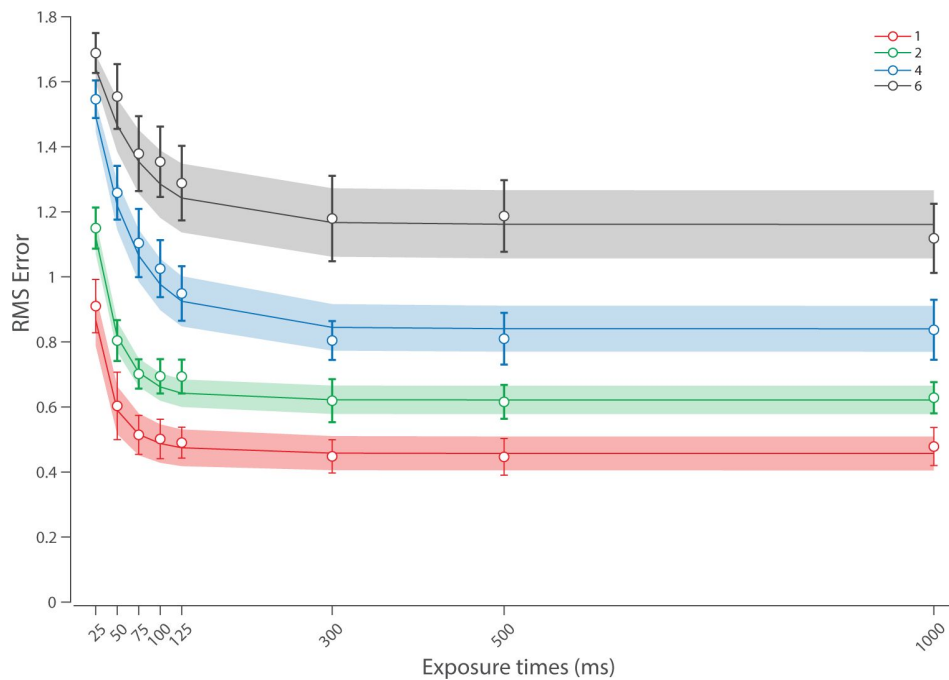


Figure S11

Summary statistics (black circles) from Experiment 1 in Bays et al. [38] and the DyNR model fits to the data (colored curves). The DyNR model was fit to the distributions of recall errors shown in Fig. S10.

References

- [1] D'Esposito Mark, Postle Bradley R. (2015) **The Cognitive Neuroscience of Working Memory** *Annual Review of Psychology* **66**:115–142 <https://doi.org/10.1146/annurev-psych-010814-015031>
- [2] Pasternak Tatiana, Greenlee Mark W. (2005) **Working memory in primate sensory systems** *Nature Reviews Neuroscience* **6**:97–107 <https://doi.org/10.1038/nrn1603>
- [3] Ma Wei Ji, Husain Masud, Bays Paul M (2014) **Changing concepts of working memory** *Nature Neuroscience* **17**:347–356 <https://doi.org/10.1038/nn.3655>
- [4] Bays Paul M., Schneegans Sebastian, Ma Wei Ji, Brady Timothy (2022) **Representation and computation in working memory** *PsyArXiv* <https://doi.org/10.31234/osf.io/kubr9>
- [5] Barlow H. B., Lecture The Ferrier (1980) **Critical limiting factors in the design of the eye and visual cortex** *Proceedings of the Royal Society of London. Series B, Biological Sciences* **212**:1–34 <https://doi.org/10.1098/rspb.1981.0022>
- [6] Van Essen D. C., Anderson C. H., Felleman D. J. (1992) **Information processing in the primate visual system: an integrated systems perspective** *Science (New York, N.Y.)* **255**:419–423
- [7] Brunton Bingni W., Botvinick Matthew M., Brody Carlos D. (2013) **Rats and Humans Can Optimally Accumulate Evidence for Decision-Making** *Science* **340**:95–98 <https://doi.org/10.1126/science.1233912>
- [8] Gold Joshua I., Shadlen Michael N. (2007) **The Neural Basis of Decision Making** *Annual Review of Neuroscience* **30**:535–574 <https://doi.org/10.1146/an-nurev.neuro.29.051605.113038>
- [9] Schneegans Sebastian, Bays Paul M. (2018) **Drift in Neural Population Activity Causes Working Memory to Deteriorate Over Time** *The Journal of Neuroscience* **38**:4859–4869 <https://doi.org/10.1523/JNEUROSCI.3440-17.2018>
- [10] Panichello Matthew F., DePasquale Brian, Pillow Jonathan W., Buschman Timothy J. (2019) **Error-correcting dynamics in visual working memory** *Nature Communications* **10** <https://doi.org/10.1038/s41467-019-11298-3>
- [11] van Ede Freek, Chekroud Sammi R., Stokes Mark G., Nobre Anna C. (2019) **Concurrent visual and motor selection during visual working memory guided action** *Nature Neuroscience* <https://doi.org/10.1038/s41593-018-0335-6>
- [12] Bays P. M. (2014) **Noise in Neural Populations Accounts for Errors in Working Memory** *Journal of Neuroscience* **34**:3632–3645 <https://doi.org/10.1523/JNEUROSCI.3204-13.2014>
- [13] Schneegans Sebastian, Taylor Robert, Bays Paul M. (2020) **Stochastic sampling provides a unifying account of visual working memory limits** *Proceedings of the National Academy of Sciences, page 202004306* <https://doi.org/10.1073/pnas.2004306117>
- [14] Ma Wei Ji, Beck Jeffrey M, Latham Peter E, Pouget Alexandre (2006) **Bayesian inference with probabilistic population codes** *Nature Neuroscience* **9**:1432–1438 <https://doi.org/10.1038/nn1790>

- [15] Pouget A., Dayan P., Zemel R. (2000) **Information processing with population codes** *Nature Reviews. Neuroscience* **1**:125–132 <https://doi.org/10.1038/35039062>
- [16] Carandini M, Heeger D.J. (2011) **Normalization as a canonical neural computation** *Nature Reviews Neuroscience* **13** <https://doi.org/10.1038/nrn3136>
- [17] Bays Paul M. (2015) **Spikes not slots: noise in neural populations limits working memory** *Trends in Cognitive Sciences* **19**:431–438 <https://doi.org/10.1016/j.tics.2015.06.004>
- [18] Tomić Ivan, Bays Paul M. (2022) **Perceptual similarity judgments do not predict the distribution of errors in working memory.** *Journal of Experimental Psychology: Learning Memory, and Cognition* <https://doi.org/10.1037/xlm0001172>
- [19] Bays Paul M., Taylor Robert (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>
- [20] Schneegans Sebastian, Bays Paul M. (2017) **Neural Architecture for Feature Binding in Visual Working Memory** *The Journal of Neuroscience* **37**:3913–3925 <https://doi.org/10.1523/JNEUROSCI.3493-16.2017>
- [21] Bays Paul M. (2016) **A signature of neural coding at human perceptual limits** *Journal of Vision* **16** <https://doi.org/10.1167/16.11.4>
- [22] Tomić Ivan, Bays Paul M. (2018) **Internal but not external noise frees working memory resources** *PLOS Computational Biology* **14** <https://doi.org/10.1371/journal.pcbi.1006488>
- [23] Pertzov Yoni, Manohar Sanjay, Husain Masud (2017) **Rapid forgetting results from competition over time between items in visual working memory.** *Journal of Experimental Psychology: Learning Memory, and Cognition* **43**:528–536 <https://doi.org/10.1037/xlm0000328>
- [24] Rademaker Rosanne L., Park Young Eun, Sack Alexander T., Tong Frank (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* <https://doi.org/10.1037/xhp0000491>
- [25] Burak Y., Fiete I. R. (2012) **Fundamental limits on persistent activity in networks of noisy neurons** *Proceedings of the National Academy of Sciences* **109**:17645–17650 <https://doi.org/10.1073/pnas.1117386109>
- [26] Wimmer Klaus, Nykamp Duane Q, Constantinidis Christos, Compte Albert (2014) **Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory** *Nature Neuroscience* **17**:431–439 <https://doi.org/10.1038/nn.3645>
- [27] Lim Phui Cheng, Ward Emily J., Vickery Timothy J., Johnson Matthew 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, pages* :1–15
- [28] Wolff Michael J., Jochim Janina, Akyürek Elkan G., Buschman Timothy J., Stokes Mark G. (2020) **Drifting codes within a stable coding scheme for working memory** *PLOS Biology* **18** <https://doi.org/10.1371/journal.pbio.3000625>

- [29] Di Lollo Vincent, Dixon Peter (1988) **Two forms of persistence in visual information processing** *Journal of Experimental Psychology: Human Perception and Performance* **14**:671–681 <https://doi.org/10.1037/0096-1523.14.4.671>
- [30] Sperling George (1960) **The information available in brief visual presentations** *Psychological Monographs: General and Applied* **74**:1–29 <https://doi.org/10.1037/h0093759>
- [31] Bradley Claire, Pearson Joel (2012) **The Sensory Components of High-Capacity Iconic Memory and Visual Working Memory** *Frontiers in Psychology* **3** <https://doi.org/10.3389/fpsyg.2012.00355>
- [32] Pratte Michael S. (2018) **Iconic Memories Die a Sudden Death** *Psychological Science* **29**:877–887 <https://doi.org/10.1177/0956797617747118>
- [33] Neisser Ulric (1966) **in Century psychology series award. Appleton-Century-Crofts, New York NY 1967**
- [34] Priebe Nicholas J., Churchland Mark M., Lisberger Stephen G. (2002) **Constraints on the source of short-term motion adaptation in macaque area MT I. the role of input and intrinsic mechanisms.** *Journal of Neurophysiology* **88**:354–369 <https://doi.org/10.1152/jn.00852.2001>
- [35] Rolls Edmund T., Tovee M.J. (1994) **Processing speed in the cerebral cortex and the neuro-physiology of visual masking** *Proceedings of the Royal Society of London. Series B: Biological Sciences* **257**:9–15 <https://doi.org/10.1098/rspb.1994.0087>
- [36] Teeuwen Rob R.M., Wacongne Catherine, Schnabel Ulf H., Self Matthew W., Roelfsema Pieter R. (2021) **A neuronal basis of iconic memory in macaque primary visual cortex** *Current Biology* **31** <https://doi.org/10.1016/j.cub.2021.09.052>
- [37] van Kerkoerle Timo, Self Matthew W., Roelfsema Pieter R. (2017) **Layer-specificity in the effects of attention and working memory on activity in primary visual cortex** *Nature Communications* **8** <https://doi.org/10.1038/ncomms13804>
- [38] 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**:6–6 <https://doi.org/10.1167/11.10.6>
- [39] Shibuya Hitomi, Bundesen Claus (1988) **Visual selection from multielement displays: Measuring and modeling effects of exposure duration** *Journal of Experimental Psychology: Human Perception and Performance* **14**:591–600 <https://doi.org/10.1037/0096-1523.14.4.591>
- [40] Vogel Edward K., Woodman Geoffrey F., Luck Steven J. (2006) **The time course of consolidation in visual working memory** *Journal of Experimental Psychology: Human Perception and Performance* **32**:1436–1451 <https://doi.org/10.1037/0096-1523.32.6.1436>
- [41] Jazayeri Mehrdad, Anthony Movshon J (2006) **Optimal representation of sensory information by neural populations** *Nature Neuroscience* **9**:690–696 <https://doi.org/10.1038/nn1691>
- [42] Ohshiro Tomokazu, Angelaki Dora E, DeAngelis Gregory C (2011) **A normalization model of multisensory integration** *Nature Neuroscience* **14**:775–782 <https://doi.org/10.1038/nn.2815>
- [43] Reynolds John H., Heeger David J. (2009) **The Normalization Model of Attention** *Neuron* **61**:168–185 <https://doi.org/10.1016/j.neuron.2009.01.002>

- [44] Hess R. F., Snowden R. J. (1992) **Temporal properties of human visual filters: number, shapes and spatial covariation** *Vision Research* **32**:47–59 [https://doi.org/10.1016/0042-6989\(92\)90112-v](https://doi.org/10.1016/0042-6989(92)90112-v)
- [45] Derrington A M, Lennie P (1984) **Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque** *The Journal of Physiology* **357**:219–240 <https://doi.org/10.1113/jphysiol.1984.sp015498>
- [46] Lee B B, Martin P R, Valberg A (1989) **Sensitivity of macaque retinal ganglion cells to chromatic and luminance flicker** *The Journal of Physiology* **414**:223–243 <https://doi.org/10.1113/jphysiol.1989.sp017685>
- [47] Hawken M. J., Shapley R. M., Grosf D. H. (1996) **Temporal-frequency selectivity in monkey visual cortex** *Visual Neuroscience* **13**:477–492 <https://doi.org/10.1017/s0952523800008154>
- [48] Müller J. R., Metha A. B., Krauskopf J., Lennie P. (2001) **Information conveyed by onset transients in responses of striate cortical neurons** *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **21**:6978–6990
- [49] Oram M. W., Perrett D. I. (1992) **Time course of neural responses discriminating different views of the face and head** *Journal of Neurophysiology* **68**:70–84 <https://doi.org/10.1152/jn.1992.68.1.70>
- [50] Ringach Dario L., Hawken Michael J., Shapley Robert (2003) **Dynamics of Orientation Tuning in Macaque V1: The Role of Global and Tuned Suppression** *Journal of Neurophysiology* **90**:342–352 <https://doi.org/10.1152/jn.01018.2002>
- [51] Sit Yiu Fai, Chen Yuzhi, Geisler Wilson S., Miikkulainen Risto, Seidemann Eyal (2009) **Complex Dynamics of V1 Population Responses Explained by a Simple Gain-Control Model** *Neuron* **64**:943–956 <https://doi.org/10.1016/j.neuron.2009.08.041>
- [52] Kovacs G., Vogels R., Orban G. A. (1995) **Cortical correlate of pattern backward masking** *Proceedings of the National Academy of Sciences* **92**:5587–5591 <https://doi.org/10.1073/pnas.92.12.5587>
- [53] Turvey M. T. (1973) **On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli** *Psychological Review* **80**:1–52 <https://doi.org/10.1037/h0033872>
- [54] Aksay E., Gamkrelidze G., Seung H. S., Baker R., Tank D. W. (2001) **In vivo intracellular recording and perturbation of persistent activity in a neural integrator** *Nature Neuroscience* **4**:184–193 <https://doi.org/10.1038/84023>
- [55] 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 (New York, N.Y.: 1991)* **10**:910–923
- [56] Coltheart Max (1980) **Iconic memory and visible persistence** *Perception & Psychophysics* **27**:183–228 <https://doi.org/10.3758/BF03204258>
- [57] Oberauer Klaus (2018) **Removal of irrelevant information from working memory: sometimes fast, sometimes slow, and sometimes not at all: Removal of irrelevant information from working memory** *Annals of the New York Academy of Sciences* **1424**:239–255 <https://doi.org/10.1111/nyas.13603>

- [58] Hick W. E. (1952) **On the Rate of Gain of Information** *Quarterly Journal of Experimental Psychology* **4**:11–26 <https://doi.org/10.1080/17470215208416600>
- [59] Shih Shui-I, Sperling George (2002) **Measuring and modeling the trajectory of visual spatial attention** *Psychological Review* **109**:260–305 <https://doi.org/10.1037/0033-295X.109.2.260>
- [60] Bouchacourt Flora, Buschman Timothy J. (2019) **A Flexible Model of Working Memory** *Neuron* **103**:147–160 <https://doi.org/10.1016/j.neuron.2019.04.020>
- [61] McMaster Jessica M.V., Tomić Ivan, Schneegans Sebastian, Bays Paul M. (2022) **Swap errors in visual working memory are fully explained by cue-feature variability** *Cognitive Psychology* **137** <https://doi.org/10.1016/j.cogpsych.2022.101493>
- [62] Tomić Ivan, Girones Zahara, Lengyel Máté, Bays Paul M. **Effects of retention interval on stimulus-specific biases in visual working memory**
- [63] Doost Rainer, Turvey M. T. (1971) **Iconic memory and central processing capacity** *Perception & Psychophysics* **9**:269–274 <https://doi.org/10.3758/BF03212646>
- [64] Koyluoglu Onur Ozan, Pertzov Yoni, Manohar Sanjay, Husain Masud, Fiete Ila R (2017) **Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity** *eLife* **6** <https://doi.org/10.7554/eLife.22225>
- [65] Sperling George, Lachmann Thomas, Weis Tina (2018) **A Brief Overview of Computational Models of Spatial, Temporal, and Feature Visual Attention Invariances in Human Information Processing** :143–182
- [66] Boerlin Martin, Denève Sophie (2011) **Spike-Based Population Coding and Working Memory** *PLoS Computational Biology* **7** <https://doi.org/10.1371/journal.pcbi.1001080>
- [67] Beck Jeffrey M., Ma Wei Ji, Kiani Roozbeh, Hanks Tim, Churchland Anne K., Roitman Jamie, Shadlen Michael N., Latham Peter E., Pouget Alexandre (2008) **Probabilistic Population Codes for Bayesian Decision Making** *Neuron* **60**:1142–1152 <https://doi.org/10.1016/j.neuron.2008.09.021>
- [68] Loftus Geoffrey R., Duncan Janine, Gehrig Paul (1992) **On the time course of perceptual information that results from a brief visual presentation** *Journal of Experimental Psychology: Human Perception and Performance* **18**:530–549 <https://doi.org/10.1037/0096-1523.18.2.530>
- [69] Lu Z.-L., Neuse J., Madigan S., Doshier B. A. (2005) **Fast decay of iconic memory in observers with mild cognitive impairments** *Proceedings of the National Academy of Sciences* **102**:1797–1802 <https://doi.org/10.1073/pnas.0408402102>
- [70] Ricker Timothy J., Spiegel Lauren R., Cowan Nelson (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**:1510–1523 <https://doi.org/10.1037/xlm0000018>
- [71] Shin Hongsup, Zou Qijia, Ma Wei Ji (2017) **The effects of delay duration on visual working memory for orientation** *Journal of Vision* **17** <https://doi.org/10.1167/17.14.10>
- [72] Zhang Weiwei, Luck Steven J. (2009) **Sudden Death and Gradual Decay in Visual Working Memory** *Psychological Science* **20**:423–428 <https://doi.org/10.1111/j.1467-9280.2009.02322.x>

- [73] Brody Carlos D, Romo Ranulfo, Kepecs Adam (2003) **Basic mechanisms for graded persistent activity: discrete attractors, continuous attractors, and dynamic representations** *Current Opinion in Neurobiology* **13**:204–211 [https://doi.org/10.1016/S0959-4388\(03\)00050-3](https://doi.org/10.1016/S0959-4388(03)00050-3)
- [74] Khona Mikail, Fiete Ila R. (2022) **Attractor and integrator networks in the brain** *Nature Reviews Neuroscience* **23**:744–766 <https://doi.org/10.1038/s41583-022-00642-0>
- [75] Khanna Sanjeev B., Snyder Adam C., Smith Matthew A. (2019) **Distinct Sources of Variability Affect Eye Movement Preparation** *The Journal of Neuroscience* **39**:4511–4526 <https://doi.org/10.1523/JNEUROSCI.2329-18.2019>
- [76] Chang M. H., Armstrong K. M., Moore T. (2012) **Dissociation of Response Variability from Firing Rate Effects in Frontal Eye Field Neurons during Visual Stimulation, Working Memory, and Attention** *Journal of Neuroscience* **32**:2204–2216 <https://doi.org/10.1523/JNEUROSCI.2967-11.2012>
- [77] Hussar C., Pasternak T. (2010) **Trial-to-trial variability of the prefrontal neurons reveals the nature of their engagement in a motion discrimination task** *Proceedings of the National Academy of Sciences* **107**:21842–21847 <https://doi.org/10.1073/pnas.1009956107>
- [78] Aagten-Murphy David, Bays Paul M. (2018) **Functions of Memory Across Saccadic Eye Movements.** Springer Berlin Heidelberg, Berlin Heidelberg
- [79] Rolfs Martin, Schweitzer Richard (2022) **Coupling perception to action through incidental sensory consequences of motor behaviour** *Nature Reviews Psychology* **1**:112–123 <https://doi.org/10.1038/s44159-021-00015-x>
- [80] Bundesen Claus (1990) **A theory of visual attention** *Psychological Review* **97**:523–547 <https://doi.org/10.1037/0033-295X.97.4.523>
- [81] Sperling George, Weichselgartner Erich (1995) **Episodic theory of the dynamics of spatial attention** *Psychological Review* **102**:503–532 <https://doi.org/10.1037/0033-295X.102.3.503>
- [82] Smith Philip L., Ratcliff Roger (2009) **An integrated theory of attention and decision making in visual signal detection** *Psychological Review* **116**:283–317 <https://doi.org/10.1037/a0015156>
- [83] Oberauer Klaus, Kliegl Reinhold (2006) **A formal model of capacity limits in working memory** *Journal of Memory and Language* **55**:601–626 <https://doi.org/10.1016/j.jml.2006.08.009>
- [84] Usher M., McClelland J. L. (2001) **The time course of perceptual choice: the leaky, competing accumulator model** *Psychological Review* **108**:550–592 <https://doi.org/10.1037/0033-295X.108.3.550>
- [85] McClelland James L. (1979) **On the time relations of mental processes: An examination of systems of processes in cascade** *Psychological Review* **86**:287–330 <https://doi.org/10.1037/0033-295X.86.4.287>
- [86] Zhou Jingyang, Benson Noah C., Kay Kendrick, Winawer Jonathan (2019) **Predicting neuronal dynamics with a delayed gain control model** *PLOS Computational Biology* **15** <https://doi.org/10.1371/journal.pcbi.1007484>
- [87] Sewell David K., Lilburn Simon D., Smith Philip L. (2014) **An information capacity limitation of visual short-term memory** *Journal of Experimental Psychology: Human Perception and Performance* **40**:2214–2242 <https://doi.org/10.1037/a0037744>

- [88] Buschman T. J., Siegel M., Roy J. E., Miller E. K. (2011) **Neural substrates of cognitive capacity limitations** *Proceedings of the National Academy of Sciences* **108**:11252–11255 <https://doi.org/10.1073/pnas.1104666108>
- [89] Sprague Thomas C., Ester Edward F., Serences John T. (2014) **Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load** *Current Biology* **24**:2174–2180 <https://doi.org/10.1016/j.cub.2014.07.066>
- [90] Taylor Robert, Bays Paul M. (2020) **Theory of neural coding predicts an upper bound on estimates of memory variability** *Psychological Review* **127**:700–718 <https://doi.org/10.1037/rev0000189>
- [91] Kriegeskorte Nikolaus, Wei Xue-Xin (2021) **Neural tuning and representational geometry** *Nature Reviews Neuroscience*, pages :1–16 <https://doi.org/10.1038/s41583-021-00502-3>
- [92] Zylberberg Ariel, Dehaene Stanislas, Mindlin Gabriel B., Sigman Mariano (2009) **Neurophysiological bases of exponential sensory decay and top-down memory retrieval: a model** *Frontiers in Computational Neuroscience* **3** <https://doi.org/10.3389/neuro.10.004.2009>
- [93] Yu Qing, Panichello Matthew F., Cai Ying, Postle Bradley R., Buschman Timothy J. (2020) **Delay-period activity in frontal, parietal, and occipital cortex tracks noise and biases in visual working memory** *PLOS Biology* **18** <https://doi.org/10.1371/journal.pbio.3000854>
- [94] Bonin Vincent, Mante Valerio, Carandini Matteo (2005) **The suppressive field of neurons in lateral geniculate nucleus** *Journal of Neuroscience* **25**:10844–10856 <https://doi.org/10.1523/JNEUROSCI.3562-05.2005>
- [95] Busse Laura, Wade Alex R., Carandini Matteo (2009) **Representation of concurrent stimuli by population activity in visual cortex** *Neuron* **64**:931–942 <https://doi.org/10.1016/j.neuron.2009.11.004>
- [96] Ni Amy M., Maunsell John H. R. (2017) **Spatially tuned normalization explains attention modulation variance within neurons** *Journal of Neurophysiology* **118**:1903–1913 <https://doi.org/10.1152/jn.00218.2017>
- [97] Bloem Ilona M., Watanabe Yurika L., Kibbe Melissa M., Ling Sam (2018) **Visual memories bypass normalization** *Psychological Science* **29**:845–856 <https://doi.org/10.1177/0956797617747091>
- [98] Lamme Victor Af, Supèr Hans, Spekreijse Henk (1998) **Feedforward, horizontal, and feedback processing in the visual cortex** *Current Opinion in Neurobiology* **8**:529–535 [https://doi.org/10.1016/S0959-4388\(98\)80042-1](https://doi.org/10.1016/S0959-4388(98)80042-1)
- [99] Lamme Victor A.F., Roelfsema Pieter R. (2000) **The distinct modes of vision offered by feedforward and recurrent processing** *Trends in Neuroscience* **23**:571–579 [https://doi.org/10.1016/S0166-2236\(00\)01657-X](https://doi.org/10.1016/S0166-2236(00)01657-X)
- [100] Semedo João D., Jasper Anna I., Zandvakili Amin, Krishna Aravind, Aschner Amir, Machens Christian K., Kohn Adam, Yu Byron M. (2022) **Feedforward and feedback interactions between visual cortical areas use different population activity patterns** *Nature Communications* **13** <https://doi.org/10.1038/s41467-022-28552-w>

- [101] Rademaker Rosanne L., Chunharas Chaipat, Serences John T. (2019) **Coexisting representations of sensory and mnemonic information in human visual cortex** *Nature Neuroscience* **22**:1336–1344 <https://doi.org/10.1038/s41593-019-0428-x>
- [102] Serences John T. (2016) **Neural mechanisms of information storage in visual short-term memory** *Vision Research* **128**:53–67 <https://doi.org/10.1016/j.visres.2016.09.010>
- [103] Xu Yaoda (2017) **Reevaluating the Sensory Account of Visual Working Memory Storage** *Trends in Cognitive Sciences* <https://doi.org/10.1016/j.tics.2017.06.013>
- [104] Goldman-Rakic P.S (1995) **Cellular basis of working memory** *Neuron* **14**:477–485 [https://doi.org/10.1016/0896-6273\(95\)90304-6](https://doi.org/10.1016/0896-6273(95)90304-6)
- [105] Bettencourt Katherine C, Xu Yaoda (2016) **Decoding the content of visual short-term memory under distraction in occipital and parietal areas** *Nature Neuroscience* **19**:150–157 <https://doi.org/10.1038/nn.4174>
- [106] Harrison Stephenie A., Tong Frank (2009) **Decoding reveals the contents of visual working memory in early visual areas** *Nature* **458**:632–635 <https://doi.org/10.1038/nature07832>
- [107] Christophel Thomas B., Iamshchinina Polina, Yan Chang, Allefeld Carsten, Haynes John-Dylan (2018) **Cortical specialization for attended versus unattended working memory** *Nature Neuroscience* <https://doi.org/10.1038/s41593-018-0094-4>
- [108] 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**:7–7 <https://doi.org/10.1167/9.10.7>
- [109] Emrich S. M., Ferber S. (2012) **Competition increases binding errors in visual working memory** *Journal of Vision* **12**:12–12 <https://doi.org/10.1167/12.4.12>
- [110] Rerko Laura, Oberauer Klaus, Lin Hsuan-Yu (2014) **Spatial transposition gradients in visual working memory** *The Quarterly Journal of Experimental Psychology* **67**:3–15 <https://doi.org/10.1080/17470218.2013.789543>
- [111] Bays Paul M (2016) **Evaluating and excluding swap errors in analogue tests of working memory** *Scientific Reports* **6** <https://doi.org/10.1038/srep19203>
- [112] Zimmermann Eckart, Concetta Morrone M., Burr David C. (2013) **Spatial position information accumulates steadily over time** *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **33**:18396–18401 <https://doi.org/10.1523/JNEUROSCI.1864-13.2013>
- [113] Williams Melonie, Hong Sang W., Kang Min-Suk, Carlisle Nancy B., Woodman Geoffrey F. (2013) **The benefit of forgetting** *Psychonomic Bulletin & Review* **20**:348–355 <https://doi.org/10.3758/s13423-012-0354-3>
- [114] LaRocque Joshua J., Lewis-Peacock Jarrod A., Drysdale Andrew T., Oberauer Klaus, Postle Bradley R. (2013) **Decoding Attended Information in Short-term Memory: An EEG Study** *Journal of Cognitive Neuroscience* **25**:127–142
- [115] Taylor Robert, Tomić Ivan, Aagten-Murphy David, Bays Paul M. (2022) **Working memory is updated by reallocation of resources from obsolete to new items. Attention, Perception & Psychophysics** <https://doi.org/10.3758/s13414-022-02584-2>

- [116] Souza Alessandra S., Rerko Laura, Oberauer Klaus (2014) **Unloading and reloading working memory: Attending to one item frees capacity** *Journal of Experimental Psychology: Human Perception and Performance* **40**:1237–1256 <https://doi.org/10.1037/a0036331>
- [117] Kelly D. H., Martinez-Uriegas E. (1993) **Measurements of chromatic and achromatic afterimages** *Journal of the Optical Society of America. A, Optics and Image Science* **10**:29–37
- [118] Tsuchiya Naotsugu, Koch Christof (2005) **Continuous flash suppression reduces negative afterimages** *Nature Neuroscience* **8**:1096–1101 <https://doi.org/10.1038/nn1500>
- [119] Ter Braak Cajo J. F. (2006) **A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces** *Statistics and Computing* **16**:239–249 <https://doi.org/10.1007/s11222-006-8769-1>

Article and author information

Ivan Tomić

University of Cambridge, Department of Psychology, Cambridge, UK, University of Zagreb, Department of Psychology, Zagreb, CRO

For correspondence: ivn.tomic@gmail.com

Paul M. Bays

University of Cambridge, Department of Psychology, Cambridge, UK

Copyright

© 2023, Ivan Tomić & Paul M. Bays

This article is distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use and redistribution provided that the original author and source are credited.

Editors

Reviewing Editor

Emilio Salinas

Wake Forest School of Medicine, Winston-Salem, United States of America

Senior Editor

Joshua Gold

University of Pennsylvania, Philadelphia, United States of America

Reviewer #2 (Public Review):

Summary:

Previous work has shown subjects can use a form of short-term sensory memory, known as 'iconic memory', to accurately remember stimuli over short periods of time (several hundred milliseconds). Working memory maintains representations for longer periods of time but is strictly limited in its capacity. While it has long been assumed that sensory information acts as the input to working memory, a process model of this transfer has been missing. To address this, Tomic and Bays present the Dynamic Neural Resource (DyNR) model. The DyNR

model captures the dynamics of processing sensory stimuli, transferring that representation into working memory, the diffusion of representations within working memory, and the selection of a memory for report.

The DyNR model captures many of the effects observed in behavior. Most importantly, psychophysical experiments found the greater the delay between stimulus presentation and the selection of an item from working memory, the greater the error. This effect also depended on working memory load. In the model, these effects are captured by the exponential decay of sensory representations (i.e., iconic memory) following the offset of the stimulus. Once the selection cue is presented, residual information in iconic memory can be integrated into working memory, improving the strength of the representation and reducing error. This selection process takes time, and is slower for larger memory loads, explaining the observation that memory seems to decay instantly. The authors compare the DyNR model to several variants, demonstrating the importance of persistence of sensory information in iconic memory, normalization of representations with increasing memory load, and diffusion of memories over time.

Strengths:

The manuscript provides a convincing argument for the interaction of iconic memory and working memory, as captured by the DyNR model. The analyses are cutting-edge and the results are well captured by the DyNR model. In particular, a strength of the manuscript is the comparison of the DyNR model to several alternative variants.

The results provide a process model for interactions between iconic memory and working memory. This will be of interest to neuroscientists and psychologists studying working memory. I see this work as providing a foundation for understanding behavior in continuous working memory tasks, from either a mechanistic, neuroscience, perspective or as a high-water mark for comparison to other psychological process models.

Finally, the manuscript is well written. The DyNR model is nicely described and an intuition for the dynamics of the model are clearly shown in Figures 2 and 4.

Weaknesses:

The manuscript appropriately acknowledges and addresses several minor weaknesses that are due to the limited ability of the approach to disambiguate underlying neural mechanisms. Nevertheless, the manuscript adds significant value to the literature on working memory.

<https://doi.org/10.7554/eLife.91034.2.sa1>

Reviewer #3 (Public Review):

Summary

The authors set out to formally contrast several theoretical models of working memory, being particularly interested in comparing the models regarding their ability to explain cueing effects at short cue durations. These benefits are traditionally attributed to the existence of a high capacity, rapidly decaying sensory storage which can be directly read out following short latency retro-cues. Based on the model fits, the authors alternatively suggest that cue-benefits arise from a freeing of working memory resources, which at short cue latencies can be utilized to encode additional sensory information into VWM.

A dynamic neural population model consisting of separate sensory and VWM populations was used to explain temporal VWM fidelity of human behavioral data collected during several working memory tasks. VWM fidelity was probed at several timepoints during

encoding, while sensory information was available and maintenance, when sensory information was no longer available. Furthermore, set size and exposure durations were manipulated to disentangle contributions of sensory and visual working memory.

Overall, the model explained human memory fidelity well, accounting for set size, exposure time, retention time, error distributions and swap errors. Crucially the model suggests that recall at short delays is due to post-cue integration of sensory information into VWM as opposed to direct readout from sensory memory. The authors formally address several alternative theories, demonstrating that models with reduced sensory persistence, direct readout from sensory memory, no set-size dependent delays in cue processing and constant accumulation rate provide significantly worse fits to the data.

I congratulate the authors for this rigorous scientific work. All my remarks were thoroughly addressed.

<https://doi.org/10.7554/eLife.91034.2.sa0>

Author Response

The following is the authors' response to the original reviews.

Reviewer #1 (Recommendations For The Authors):

I only have a few minor suggestions:

Abstract: I really liked the conclusion (that IM and VWM are two temporal extremes of the same process) as articulated in lines 557--563. (It is always satisfying when the distinction between two things that seem fundamentally different vanishes). If something like this but shorter could be included in the Abstract, it would highlight the novel aspects of the results a little more, I think.

Thank you for this comment. We have added the following to the abstract:

“A key conclusion is that differences in capacity classically thought to distinguish IM and VWM are in fact contingent upon a single resource-limited WM store.”

L 216: There's an orphan parenthesis in "(justifying the use".

Fixed.

L 273: "One surprising result was the observed set size effect in the 0 ms delay condition". In this paragraph, it might be a good idea to remind the reader of the difference between the simultaneous and zero-delay conditions. If I got it right, the results differ between these conditions because it takes some amount of processing time to interpret the cue and free the resources associated with the irrelevant stimuli. Recalling that fact would make this paragraph easier to digest.

That is correct. However, at this point in the text, we have not yet fitted the DyNR model to the data. Therefore, we believe that introducing cue processing and resource reallocation as concepts that differentiate between those two conditions would disrupt the flow of this paragraph. We address these points soon after, in a paragraph starting on line 341.

Figures 3, 5: The labels at the bottom of each column in A would be more clear if placed at the top of each column instead. That way, the x-axis for the plots in A could be labeled appropriately, as "Error in orientation estimate" or something to that effect.

We edited both figures, now Figure 4 and Figure 6, as suggested.

L 379: It should be "(see Eq 6)", I believe.

That is correct, line 379 (currently line 391) should read 'Eq 6'. Fixed.

L 379--385: I was a bit mystified as to why the scaled diffusion rate produced a worse fit than a constant rate. I imagine the scaled version was set to something like

$$\sigma^2_{diff_scaled} = \sigma^2_{base} + K*(N-1)$$

where N is the set size and σ^2_{base} and K are parameters. If this model produced a similar fit as with a constant diffusion rate, the AIC would penalize it because of the extra parameter. But why would the fit be worse (i.e., not match the pattern of variability)? Shouldn't the fitter just find that the K=0 solution is the best? Not a big deal; the Nelder-Mead solutions can wobble when that many parameters are involved, but if there's a simple explanation it might be worth commenting on.

The scaled diffusion was implemented by extending Eq 6 in the following way:

$$\sigma(t)^2 = (t - t_{offset}) * \sigma^2_{diff} * N$$

where N is set size. Therefore, the scaling was not associated with a free parameter that could become 0 if set size did not affect diffusion rate, but variability rather mandatory increased with set size. We now clarify this in the text:

"The second variant was identical to the proposed model, except that we replaced the constant diffusion rate with a set size scaled diffusion rate by multiplying the right side of Eq 6 by N."

Figure 4 is not mentioned in the main text. Maybe the end of L 398 would be a good place to point to it. The paragraph at L 443-455 would also benefit from a couple of references to it.

Thank you for this suggestion. Figure 4 (now Figure 5) was previously mentioned on line 449 (previously line 437), but now we have included it on line 410 (previously line 398), within the paragraph spanning lines 455-467 (previously 443-455), and also on line 136 where we first discuss masking effects.

L 500: Figure S7 is mentioned before Figures S5 and S6. Quite trivial, I know....

Thank you for this comment. There was no specific reason for Figure S7 to appear after S5 & S6, so we simply swapped their order to be consistent with how they are referred to in the manuscript (i.e., S7 became S5, S5 became S6, and S6 became S7).

Reviewer #2 (Recommendations For The Authors):

(1) One potential weakness is that the model assumes sensory information is veridical. However, this isn't likely the case. Acknowledging noise in sensory representations could affect the model interpretation in a couple of different ways. First, neurophysiological recordings have shown normalization affects sensory representations, even when a stimulus is still present on the screen. The DyNR model partially addresses this concern because reports are drawn from working memory, which is normalized. However, if sensory representations were also normalized, then it may improve the model variant

where subjects draw directly from sensory representations (an alternative model that is currently described but discarded).

Thank you for this suggestion. We can consider two potential mechanisms through which divisive normalization might be incorporated into sensory processing within the DyNR model.

The first possibility involves assuming that normalization is pre-attentive. In this scenario, the sensory activity of each object would be rescaled at the lowest level of sensory processing, occurring before the allocation of attentional or VWM resources. One strong prediction of such an implementation is that recall error in the simultaneous cue condition (Experiment 1) should vary with set size. However, this prediction is inconsistent with the observed data, which failed to show a significant difference between set sizes, and is more closely aligned with the hypothesis of no-difference ($F(2,18) = 1.26$, $p = .3$, $\eta^2 = .04$, $BF_{10} = 0.47$). On that basis, we anticipate that introducing normalization as a pre-attentive mechanism would impair the model fit.

An alternative scenario is to consider normalization as post-attentive. In the simultaneous cueing condition, only one item is attended (i.e., the cued one), regardless of the displayed set size. Here, we would expect normalized activity for a single item, regardless of the number of presented objects, which would then be integrated into VWM. This expanded DyNR model with post-attentive normalization would make exactly the same predictions as the proposed DyNR for recall fidelity, so distinguishing between these models would not be possible based on working memory experiments.

To acknowledge the possibility that sensory signals could undergo divisive normalization and to motivate future research, we have added the following to our manuscript:

“As well as being implicated in higher cognitive processes including VWM (Buschman et al, 2011; Sprague et al., 2014), divisive normalization has been shown to be widespread in basic sensory processing (Bonin et al., 2005; Busse et al., 2009; Ni et al., 2017). The DyNR model presently incorporates the former but not the latter type of normalization. While the data observed in our experiments do not provide evidence for normalization of sensory signals (note comparable recall errors across set size in the simultaneous cue condition of Experiment 1), this may be because sensory suppressive effects are localized and our stimuli were relatively widely separated in the visual field: future research could explore the consequences of sensory normalization for recall from VWM using, e.g., centre-surround stimuli (Bloem et al., 2018).”

Bloem, I. M., Watanabe, Y. L., Kibbe, M. M., & Ling, S. (2018). Visual Memories Bypass Normalization. *Psychological Science*, 29(5), 845–856. <https://doi.org/10.1177/0956797617747091>

Bonin, V., Mante, V., & Carandini, M. (2005). The Suppressive Field of Neurons in Lateral Geniculate Nucleus. *The Journal of Neuroscience*, 25(47), 10844–10856. <https://doi.org/10.1523/JNEUROSCI.3562-05.2005>

Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, 108(27), 11252–11255. <https://doi.org/10.1073/pnas.1104666108>

Busse, L., Wade, A. R., & Carandini, M. (2009). Representation of Concurrent Stimuli by Population Activity in Visual Cortex. *Neuron*, 64(6), 931–942. <https://doi.org/10.1016/j.neuron.2009.11.004>

Ni, A. M., & Maunsell, J. H. R. (2017). Spatially tuned normalization explains attention modulation variance within neurons. *Journal of Neurophysiology*, 118(3), 1903–1913. <https://doi.org/10.1152/jn.00804.2017>

doi.org/10.1152/jn.00218.2017

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. *Current Biology*, 24(18), 2174–2180. <https://doi.org/10.1016/j.cub.2014.07.066>

Second, visual adaptation predicts sensory information should decrease over time. This would predict that for long stimulus presentation times, the error would increase. Indeed, this seems to be reflected in Figure 5B. This effect is not captured by the DyNR model.

Indeed, neural responses in the visual cortex have been observed to quickly adapt during stimulus presentation, showing reduced responses to prolonged stimuli after an initial transient (Groen et al., 2022; Sawamura et al., 2006; Zhou et al., 2019). This adaptation typically manifests as 1) reduced activity towards the end of stimulus presentation and 2) a faster decay towards baseline activity after stimulus offset.

In the DyNR model, we use an idealized solution in which we convolve the presented visual signal with a response function (i.e., temporal filter). At the longest presentation durations, in DyNR, the sensory signal plateaus and remains stable until stimulus offset. Because our psychophysical data does not allow us to identify the exact neural coding scheme that underlies the sensory signal, we tend to favour this simple implementation, which is broadly consistent with some previous attempts to model temporal dynamics in sensory responses (e.g., Carandini and Heeger, 1994). However, we agree with the reviewer that some adaptation of the sensory signal with prolonged presentation would also be consistent with our data.

We have added the following to the manuscript:

“In Experiment 2, the longest presentation duration shows an upward trend in error at set sizes 4 and 10. While this falls within the range of measurement error, it is also possible that this is a meaningful pattern arising from visual adaptation of the sensory signal, whereby neural populations reduce their activity after prolonged stimulation. This would mean less residual sensory signal would be available after the cue to supplement VWM activity, predicting a decline in fidelity at higher set sizes. Visual adaptation has previously been successfully accounted for by a type of delayed normalization model in which the sensory signal undergoes a series of linear and nonlinear transformations (Zhou et al., 2019). Such a model could in future be incorporated into DyNR and validated against psychophysical and neural data.”

Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, 264(5163), 1333–1336. <https://doi.org/10.1126/science.8191289>

Groen, I. I. A., Piantoni, G., Montenegro, S., Flinker, A., Devore, S., Devinsky, O., Doyle, W., Dugan, P., Friedman, D., Ramsey, N. F., Petridou, N., & Winawer, J. (2022). Temporal Dynamics of Neural Responses in Human Visual Cortex. *The Journal of Neuroscience*, 42(40), 7562–7580. <https://doi.org/10.1523/JNEUROSCI.1812-21.2022>

Sawamura, H., Orban, G. A., & Vogels, R. (2006). Selectivity of Neuronal Adaptation Does Not Match Response Selectivity: A Single-Cell Study of the fMRI Adaptation Paradigm. *Neuron*, 49(2), 307–318. <https://doi.org/10.1016/j.neuron.2005.11.028>

Zhou, J., Benson, N. C., Kay, K., & Winawer, J. (2019). Predicting neuronal dynamics with a delayed gain control model. *PLOS Computational Biology*, 15(11), e1007484. <https://doi.org/10.1371/journal.pcbi.1007484>

(2) A second potential weakness is that, in Experiment 1, the authors briefly change the sensory stimulus at the end of the delay (a 'phase shift', Fig. 6A). I believe this is intended to act as a mask. However, I would expect that, in the DyNR model, this should be modeled as a new sensory input (in Experiment 2, 50 ms is plenty of time for the subjects to process the stimuli). One might expect this change to disrupt sensory and memory representations in a very characteristic manner. This seems to make a strong testable hypothesis. Did the authors find evidence for interference from the phase shift?

The phase shift was implemented with the intention of reducing retinal after-effects, essentially acting as a mask for retinal information only; crucially the orientation of the stimulus is unchanged by the phase shift, so from the perspective of the DyNR model, it transmits the same orientation information to working memory as the original stimulus.

If our objective were to model sensory input at the level of individual neurons and their receptive fields, we would indeed need to treat this phase shift as a novel input. Nevertheless, for DyNR, conceived as an idealization of a biological system for encoding orientation information, we can safely assume that visual areas in biological organisms have a sufficient number of phase-sensitive simple cells and phase-indifferent complex cells to maintain the continuity of input to VWM.

When comparing conditions with and without the phase shift of stimuli (Fig S1B), we found performance to be comparable in the perceptual condition (simultaneous presentation) and with the longest delay (1 second), suggesting that the phase shift did not change the visibility or encoding of information into VWM. In contrast, we found strong evidence that observers had access to an additional source of information over intermediate delays when the phase shift was not used. This was evident through enhanced recall performance from 0 ms to 400 ms delay. Based on this, we concluded that the additional source of information available in the absence of a phase shift was accessible immediately following stimulus offset and had a brief duration, aligning with the theoretical concept of retinal afterimages.

(3) It seems odd that the mask does not interrupt sensory processing in Experiment 2. Isn't this the intended purpose of the mask? Should readers interpret this as all masks not being effective in disrupting sensory processing/iconic memory? Or is this specific to the mask used in the experiment?

Visual masks are often described as instantly and completely halting the visual processing of information that preceded the mask. We also anticipated the mask would entirely terminate sensory processing, but our data indicate the effect was not complete (as indicated by model variants in Experiment 2). Nevertheless, we believe we achieved our intended goal with this experiment – we observed a clear modulation of response errors with changing stimulus duration, indicating that the post-stimulus information that survived masking did not compromise the manipulation of stimulus duration. Moreover, the DyNR model successfully accounted for the portion of signal that survived the mask.

We can identify two possible reasons why masking was incomplete. First, it is possible that the continuous report measure used in our experiments is more sensitive than the discrete measures (e.g., forced-choice methods) commonly employed in experiments that found masks to be 100% effective. Second, despite using a flickering white noise mask at full contrast, it is possible that it may not have been the most effective mask; for instance, a mask consisting of many randomly oriented Gabor patches matched in spatial frequency to the stimuli could prove more effective. We decided against such a mask because we were concerned that it could potentially act as a new input to orientation-sensitive neurons, rather than just wiping out any residual sensory activity.

(4) I apologize if I missed it, but the authors did not compare the DyNR model to a model without decaying sensory information for Experiment 1.

We tested two DyNR variants in which the diffusion process was solely responsible for memory fidelity dynamics. These models assumed that the sensory signal terminates abruptly with stimuli offset, and the VWM signal encoding the stimuli was equal to the limit imposed by normalization, independent of the delay duration.

As variants of this model failed to account for the observed response errors both quantitatively (see 'Fixed neural signal' under Model variants) and qualitatively (Figure S3), we decided not to test any more restrictive variants, such as the one without sensory decay and diffusion.

(5) In the current model, selection is considered to be absolute (all or none). However, this need not be the case (previous work argues for graded selection). Could a model where memories are only partially selected, in a manner that is mediated by load, explain the load effects seen in behavior?

Thank you for this point. If attentional selection was partial, it would affect the observers' efficiency in discarding uncued objects to release allocated resources and encode additional information about the cued item. We and others have previously examined whether humans can efficiently update their VWM when previous items become obsolete. For example, Taylor et al. (2023) showed that observers could efficiently remove uncued items from VWM and reallocate the released resources to new visual information. These findings align with results from other studies (e.g., Ecker, Oberauer, & Lewandowsky, 2014; Kessler & Meiran, 2006; Williams et al., 2013).

Based on these findings, we feel justified in assuming that observers in our current task were capable of fully removing all uncued objects, allowing them to continue the encoding process for the cued orientation that was already partially stored in VWM, such that the attainable limit on representational precision for the cued item equals the maximum precision of VWM.

Partial removal could in principle be modelled in the DyNR model by introducing an additional plateau parameter specifying a maximum attainable precision after the cue. Our concern would be that such a plateau parameter would trade off with the parameter associated with Hick's law (i.e., cue interpretation time). The former would control the amount of information that can be encoded into VWM, while the latter regulates the amount of sensory information available for encoding. We are wary of adding additional parameters, and hence flexibility, to the model where we do not have the data to sufficiently constrain them.

Ecker, U. K. H., Oberauer, K., & Lewandowsky, S. (2014b). Working memory updating involves item-specific removal. *Journal of Memory and Language*, 74, 1–15. <https://doi.org/10.1016/j.jml.2014.03.006>

Kessler, Y., & Meiran, N. (2006). All updateable objects in working memory are updated whenever any of them are modified: Evidence from the memory updating paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 570–585. <https://doi.org/10.1037/0278-7393.32.3.570>

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>

Williams, M., & Woodman, G. F. (2012). Directed forgetting and directed remembering in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5), 1206–1220. <https://doi.org/10.1037/a0027389>

(6) Previous work, both from the authors and others, has shown that memories are biased as if they are acted on by attractive/repulsive forces. For example, the memory of an oriented bar is biased away from horizontal and vertical and biased towards diagonals. This is not accounted for in the current model. In particular, this could be one mechanism to generate a non-uniform drift rate over time. As noted in the paper, a non-uniform drift rate could capture many of the behavioral effects reported.

The reviewer is correct that the model does not currently include stimulus-specific effects, although our work on that topic provides a clear template for incorporating them in future (e.g. Taylor & Bays, 2018). Specifically on the question of generating a non-uniform drift, we have another project that currently looks at this exact question (cited in our manuscript as Tomic, Girones, Lengyel, and Bays; in prep.). By examining various datasets with varying memory delays, including the Additional Dataset 1 reported in the Supplementary Information, we found that stimulus-specific effects on orientation recall remain constant with retention time. Specifically, although there is a clear increase in overall error over time, estimation biases remain constant in direction and amplitude, indicating that the bias does not manifest in drift rates (see also Rademaker et al., 2018; Figure S1).

Taylor, R., & Bays, P. M. (2018). Efficient coding in visual working memory accounts for stimulus-specific variations in recall. *The Journal of Neuroscience*, 1018–18. <https://doi.org/10.1523/JNEUROSCI.1018-18.2018>

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*. <https://doi.org/10.1037/xhp0000491>

(7) Finally, the authors use AIC to compare many different model variants to the DyNR model. The delta-AICs are high (>10), indicating a strong preference for the DyNR model over the variants. However, the overall quality of fit to the data is not clear. What proportion of the variance in data was the model able to explain? In particular, I think it would be helpful for the reader if the authors reported the variance explained on withheld data (trials, conditions, or subjects).

Thank you for this comment.

Below we report the estimates of r^2 , representing the goodness of fit between observed data (i.e., RMSE) and the DyNR model predictions.

In Experiment 1, the r^2 values between observations and predictions were computed across delays for each set size, yielding the following estimates: $r^2_{ss1} = 0.60$; $r^2_{ss4} = 0.87$; $r^2_{ss10} = 0.95$. Note that lower explained variance for set size 1 arises from both data and model predictions having near-constant precision.

In Experiment 2, we calculated r^2 between observations and predictions across presentation durations, separately for each set size, resulting in the following estimates: $r^2_{ss1} = 0.88$; $r^2_{ss4} = 0.71$; $r^2_{ss10} = 0.70$. Note that in this case the decreasing percentage of explained variance with set size is a consequence of having less variability in both data and model predictions with larger set sizes.

While these estimates suggest that the DyNR model effectively fits the psychophysical data, a more rigorous validation approach would involve cross-validation checks across all conditions with a withheld portion of trials. Regrettably, due to the large number of conditions in each experiment, we could only collect 50 trials per condition. We are sceptical that fitting the model to even fewer trials, as necessary for cross-validation, would provide a reliable assessment of model performance.

Minor: It isn't clear to me why the behavioral tasks are shown in Figure 6. They are important for understanding the results and are discussed earlier in the manuscript (before Figure 3). This just required flipping back and forth to understand the task before I could interpret the results.

Thank you for this comment. We have now moved the behavioural task figure to appear early in the manuscript (as Figure 3).

Reviewer #3 (Recommendations For The Authors):

(1) Dynamics of sensory signals during perception

I believe that the modeled sensory signal is a reasonable simplification and different ways to model the decay function are discussed. I would like to ask the authors to discuss the implications of slightly more complex initial sensory transients such as the ones shown in Teeuwen (2021). Specifically for short exposure times, this might be particularly relevant for the model fits as some of the alternative models diverge from the data for short exposures. In addition, the role of feedforward (initial transient?) and feedback signaling (subsequent "plateau" activity) could be discussed. The first one might relate more strongly to sensory signals whereas the latter relates more to top-down attention/recurrent processing/VWM.

Particularly, this latter response might also be sensitive to the number of items present on the screen which leads to a related question pertaining to the limitations of attention during perception. Some work suggests that perception is similarly limited in the amount of information that can be represented concurrently (Tsubomi, 2013). Could the authors discuss the implications of this hypothesis? What happens if maximum sensory amplitude is set as a free parameter in the model?

*Tsubomi, H., Fukuda, K., Watanabe, K., & Vogel, E. K. (2013). Neural limits to representing objects still within view. *Journal of Neuroscience*, 33(19), 8257-8263.*

Thank you for this question. Below, we unpack it and answer it point by point.

While we agree our model of the sensory response is justified as an idealization of the biological reality, we also recognise that recent electrophysiological recordings have illuminated intricacies of neuronal responses within the striate cortex, a critical neural region associated with sensory memory (Teeuwen et al, 2021). Notably, these recordings reveal a more nuanced pattern where neurons exhibit an initial burst of activity succeeded by a lower plateau in firing rate, and stimulus offset elicits a second small burst in the response of some neurons, followed by a gradual decrease in activity after the stimulus disappears (Teeuwen et al, 2021).

In general, asynchronous bursts of activity in individual neurons will tend to average out in the population making little difference to predictions of the DyNR model. Synchronized bursts at stimulus onset could affect predictions for the shortest presentations in Exp 2, however the model appears to capture the data very well without including them. We would be wary of incorporating these phenomena into the model without more clarity on their

universality (e.g., how stimulus-dependent they are), their significance at the population level (as opposed to individual neurons), and most importantly, their prominence in visual areas outside striate cortex. Specifically, while Teeuwen et al. (2021) described activity in V1, our model does not make strong assumptions about which visual areas are the source of the sensory input to WM. Based on these uncertainties we believe the idealized sensory response is justified for use in our model.

Next, thank you for the comment on feedforward and feedback signals. We have added the following to our manuscript:

“Following onset of a stimulus, the visual signal ascends through visual areas via a cascade of feedforward connections. This feedforward sweep conveys sensory information that persists during stimulus presentation and briefly after it disappears (Lamme et al., 1998). Simultaneously, reciprocal feedback connections carry higher-order information back towards antecedent cortical areas (Lamme and Roelfsema, 2000). In our psychophysical task, feedback connections likely play a critical role in orienting attention towards the cued item, facilitating the extraction of persisting sensory signals, and potentially signalling continuous information on the available resources for VWM encoding. While our computational study does not address the nature of these feedforward and feedback signals, a challenge for future research is to describe the relative contributions of these signals in mediating transmission of information between sensory and working memory (Semedo et al., 2022).”

Lamme, V. A., Supér, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, 8(4), 529–535. [https://doi.org/10.1016/S0959-4388\(98\)80042-1](https://doi.org/10.1016/S0959-4388(98)80042-1)

Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11), 571–579. [https://doi.org/10.1016/S0166-2236\(00\)01657-X](https://doi.org/10.1016/S0166-2236(00)01657-X)

Semedo, J. D., Jasper, A. I., Zandvakili, A., Krishna, A., Aschner, A., Machens, C. K., Kohn, A., & Yu, B. M. (2022). Feedforward and feedback interactions between visual cortical areas use different population activity patterns. *Nature Communications*, 13(1), 1099. <https://doi.org/10.1038/s41467-022-28552-w>

Finally, both you and Reviewer 2 raised a similar interesting question regarding capacity limitations of attention during perception. Such a limitation could be modelled by freely estimating sensory amplitude and implementing divisive normalization to that signal, similar to how VWM is constrained. We can consider two potential mechanisms through which divisive normalization might be incorporated into sensory processing within the DyNR model.

The first possibility involves assuming that normalization is pre-attentive. In this scenario, the sensory activity of each object would be rescaled at the lowest level of sensory processing, occurring before the allocation of attentional or VWM resources. One strong prediction of such an implementation is that recall error in the simultaneous cue condition (Experiment 1) should vary with set size. However, this prediction is inconsistent with the observed data, which failed to show a significant difference between set sizes, and is more closely aligned with the hypothesis of no-difference ($F(2,18) = 1.26$, $p = .3$, $\eta^2 = .04$, $BF_{10} = 0.47$). On that basis, we anticipate that introducing normalization as a pre-attentive mechanism would impair the model fit.

An alternative scenario is to consider normalization as post-attentive. In the simultaneous cueing condition, only one item is attended (i.e., the cued one), regardless of the displayed set size. Here, we would expect normalized activity for a single item, regardless of the number of presented objects, which would then be integrated into VWM. This expanded DyNR model with post-attentive normalization would make exactly the same predictions as the proposed

DyNR for recall fidelity, so distinguishing between these models would not be possible based on working memory experiments.

To acknowledge the possibility that sensory signals could undergo divisive normalization and to motivate future research, we have added the following to our manuscript:

“As well as being implicated in higher cognitive processes including VWM (Buschman et al, 2011; Sprague et al., 2014), divisive normalization has been shown to be widespread in basic sensory processing (Bonin et al., 2005; Busse et al., 2009; Ni et al., 2017). The DyNR model presently incorporates the former but not the latter type of normalization. While the data observed in our experiments do not provide evidence for normalization of sensory signals (note comparable recall errors across set size in the simultaneous cue condition of Experiment 1), this may be because sensory suppressive effects are localized and our stimuli were relatively widely separated in the visual field: future research could explore the consequences of sensory normalization for recall from VWM using, e.g., centre-surround stimuli (Bloem et al., 2018).”

Bloem, I. M., Watanabe, Y. L., Kibbe, M. M., & Ling, S. (2018). Visual Memories Bypass Normalization. *Psychological Science*, 29(5), 845–856. <https://doi.org/10.1177/0956797617747091>

Bonin, V., Mante, V., & Carandini, M. (2005). The Suppressive Field of Neurons in Lateral Geniculate Nucleus. *The Journal of Neuroscience*, 25(47), 10844–10856. <https://doi.org/10.1523/JNEUROSCI.3562-05.2005>

Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, 108(27), 11252–11255. <https://doi.org/10.1073/pnas.1104666108>

Busse, L., Wade, A. R., & Carandini, M. (2009). Representation of Concurrent Stimuli by Population Activity in Visual Cortex. *Neuron*, 64(6), 931–942. <https://doi.org/10.1016/j.neuron.2009.11.004>

Ni, A. M., & Maunsell, J. H. R. (2017). Spatially tuned normalization explains attention modulation variance within neurons. *Journal of Neurophysiology*, 118(3), 1903–1913. <https://doi.org/10.1152/jn.00218.2017>

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. *Current Biology*, 24(18), 2174–2180. <https://doi.org/10.1016/j.cub.2014.07.066>

(2) Effectivity of retro-cues at long delays

Can the authors discuss how cues presented at long delays (>1000 ms) can still lead to increased memory fidelity when sensory signals are likely to have decayed? A list of experimental work demonstrating this can be found in Souza & Oberauer (2016).

Souza, A. S., & Oberauer, K. (2016). In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Attention, Perception, & Psychophysics*, 78, 1839–1860.

The increased memory fidelity observed with longer delays between memory array offset and cue does not result from integrating available sensory signals into VWM because the sensory signal would have completely decayed by that time. Instead, research so far has indicated several alternative mechanisms that could lead to higher recall precision for cued items, and we can briefly summarize some of them, which are also reviewed in more detail in Souza and Oberauer (2016).

One possibility is that, after a highly predictive retro-cue indicates the to-be-tested item, uncued items can simply be removed from VWM. This could result in decreased interference for the cued item, and consequently higher recall precision. Secondly, the retro-cue could also indicate which item can be selectively attended to, and thereby differentially strengthening it in memory. Furthermore, the retro-cue could allow evidence to accumulate for the target item ahead of decision-making, and this could increase the probability that the correct information will be selected for response. Finally, the retro-cued stimulus could be insulated from interference by subsequent visual input, while the uncued stimuli may remain prone to such interference.

A neural account of this retro-cue effect based on the original neural resource model has been proposed in Bays & Taylor, *Cog Psych*, 2018. However, as we did not use a retro-cue design in the present experiments, we have decided not to elaborate on this in the manuscript.

(3) Swap errors

I am somewhat surprised by the empirically observed and predicted pattern of swap errors displayed in Figure S2. For set size 10, swap probability does not consistently increase with the duration of the retention interval, although this was predicted by the author's model. At long intervals, swap probability is significantly higher for large compared to small set sizes, which also seems to contrast with the idea of shared, limited VWM resources. Can the authors provide some insight into why the model fails to reproduce part of the behavioral pattern for swap errors? The sentence in line 602 might also need some reconsideration in this regard.

Determining the ground truth for swap errors poses a challenge. The prevailing approach has been to employ a simpler model that estimates swap errors, such as a three-component mixture model, and use those estimates as a proxy for ground truth. However, this method is not without its shortcomings. For example, the variability of swap frequency estimates tends to increase with variability in the report feature dimension (here, orientation). This is due to the increasing overlap of response probability distributions for swap and non-swap responses. Consequently, the discrepancy between any two methods of swap estimation is most noticeable when there is substantial variability in orientation reports (e.g., 10 items and long delay or short exposure).

When modelling swap frequency in the DyNR model, our aim was to provide a parsimonious account of swap errors while implementing similar dynamics in the spatial (cue) feature as in the orientation (report) feature. This parametric description captured the overall pattern of swap frequency with set size and retention and encoding time, but is still only an approximation of the predictions if we fully modelled memory for the conjunction of cue and report features (as in e.g. Schneegans & Bays, 2017; McMaster et al, 2020).

We expanded the existing text in the section ‘Representational dynamics of cue-dimension features’ of our manuscript:

“... Although we did not explicitly model the neural signals representing location, the modelled dynamics in the probability of swap errors were consistent with those of the primary memory feature. We provided a more detailed neural account of swap errors in our earlier works that is theoretically compatible with the DyNR model (McMaster et al., 2020; Schneegans & Bays, 2017).

The DyNR model successfully captured the observed pattern of swap frequencies (intrusion errors). The only notable discrepancy between DyNR and the three-component mixture model (Fig. S2) arises with the largest set size and longest delay, although with considerable

interindividual variability. As the variability in report-dimension increases, the estimates of swap frequency become more variable due to the growing overlap between the probability distributions of swap and non-swap responses. This may explain apparent deviations from the modelled swap frequencies with the highest set size and longest delay where orientation response variability was greatest. “

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, 101493. <https://doi.org/10.1016/j.cogpsych.2022.101493>

Schneegans, S., & Bays, P. M. (2017). Neural Architecture for Feature Binding in Visual Working Memory. *The Journal of Neuroscience*, 37(14), 3913–3925. <https://doi.org/10.1523/JNEUROSCI.3493-16.2017>

(4) Direct sensory readout

The model assumes that readout from sensory memory and from VWM happens with identical efficiency. Currently, we don't know if these two systems are highly overlapping or are fundamentally different in terms of architecture and computation. In the case of the latter, it might be less reasonable to assume that information readout would happen at similar efficiencies, as it is currently assumed in the manuscript. Perhaps the authors could briefly discuss this possibility.

In the direct sensory read-out model, we did not explicitly model the efficiency of readout from either sensory or VWM store. However, the distinctive prediction of this model is that the precision of recall changes exponentially with delay at every set size, including one item. This prediction does not depend on the relative efficiency of readout from sensory and working memory, but only on the principle that direct readout from sensory memory bypasses the capacity limit on working memory. This prediction is inconsistent with the pattern of results observed in Experiment 1, where early cues did not show a beneficial effect on recall error for set size 1. While the proposal raised by the reviewer is intriguing, even if we were to model the process of readout from both the sensory and VWM stores with different efficiencies, the direct read-out model could not account for the near-constant recall error with delay for set size one.

(5) Encoding of distractors

One of the model assumptions is that, for simultaneous presentations of memory array and cue only the cued feature will be encoded. Previous work has suggested that participants often accidentally encode distractors even when they are cued before memory array onset (Vogel 2005). Given these findings, how reasonable is this assumption in the authors' model?

*Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, 438(7067), 500-503.*

Although previous research suggested that observers can misinterpret the pre-cue and encode one of the uncued items, our results argue against this being the case in the current experiment. Such encoding failures would manifest in overall recall error, resulting in a gradient of error with set size, owing to the presence of more adjacent distractors in larger set sizes. However, when we compared recall errors between set sizes in the simultaneous cue condition, we did not find a significant difference between set sizes, and moreover, our results were more likely under the hypothesis of no-difference ($F(2,18) = 1.26$, $p = .3$, $\eta^2 = .04$, $BF_{10} = 0.47$). If observers occasionally encoded and reported one of the uncued items in the

simultaneous cue condition, those errors were extremely infrequent and did not affect the overall error distributions.