J Neurophysiol 120: 1945–1961, 2018.
First published June 27, 2018; doi:10.1152/jn.00778.2017.

RESEARCH ARTICLE | *Working Memory: Neural Mechanisms*

Slot-like capacity and resource-like coding in a neural model of multiple-item working memory

Dominic Standage and Martin Paré
Centre for Neuroscience Studies, Queen's University, Kingston, Ontario, Canada

Submitted 30 October 2017; accepted in final form 22 June 2018

Standage D, Paré M. Slot-like capacity and resource-like coding in a neural model of multiple-item working memory. *J Neurophysiol* 120: 1945–1961, 2018. First published June 27, 2018; doi:10.1152/jn.00778.2017.—For the past decade, research on the storage limitations of working memory has been dominated by two fundamentally different hypotheses. On the one hand, the contents of working memory may be stored in a limited number of "slots," each with a fixed resolution. On the other hand, any number of items may be stored but with decreasing resolution. These two hypotheses have been invaluable in characterizing the computational structure of working memory, but neither provides a complete account of the available experimental data or speaks to the neural basis of the limitations it characterizes. To address these shortcomings, we simulated a multiple-item working memory task with a cortical network model, the cellular resolution of which allowed us to quantify the coding fidelity of memoranda as a function of memory load, as measured by the discriminability, regularity, and reliability of simulated neural spiking. Our simulations account for a wealth of neural and behavioral data from human and nonhuman primate studies, and they demonstrate that feedback inhibition lowers both capacity and coding fidelity. Because the strength of inhibition scales with the number of items stored by the network, increasing this number progressively lowers fidelity until capacity is reached. Crucially, the model makes specific, testable predictions for neural activity on multiple-item working memory tasks.

NEW & NOTEWORTHY Working memory is the ability to keep information in mind and is fundamental to cognition. It is actively debated whether the storage limitations of working memory reflect a small number of storage units (slots) or a decrease in coding resolution as a limited resource is allocated to more items. In a cortical model, we found that slot-like capacity and resource-like neural coding resulted from the same mechanism, offering an integrated explanation for storage limitations.

biophysically based model; working memory; working memory capacity; working memory precision; working memory storage

INTRODUCTION

Working memory refers to the retention of information for use in cognitive tasks over intervals on the order of seconds. Visual working memory (WM) is a particularly active research field, largely because the high precision of the visual system affords fine-grained measurements that address the storage limitations of WM. These limitations are highly correlated with measures of intelligence and are currently the subject of intense research interest (see Luck and Vogel 2013).

For several decades, research on storage limitations was dominated by the hypothesis that WM is supported by a small number of discrete "slots." According to this hypothesis, information is either stored with high precision in a slot or simply not encoded if the number of items n exceeds the number of slots (see Cowan 2001). More recently, evidence has emerged for an alternative hypothesis, according to which a limited "resource" R is allocated to n items, with no limit on n. Accordingly, the precision of WM representations decreases with increasing n, since less resource is available for the encoding of each item, i.e., precision tracks R/n. Thus the nature of WM storage limitations is fundamentally different under the slot and resource hypotheses, attributing constraints to capacity and resolution, respectively. It is increasingly clear, however, that neither is complete (see Luck and Vogel 2013; Ma et al. 2014). Generally, the slot hypothesis (Slot) is overconstrained with respect to resolution, since it does not account for a gradual decrease in precision with increasing n (Bays et al. 2009; Schneegans and Bays 2016). Equally, the resource hypothesis (Resource) is overconstrained with respect to capacity, since it does not account for a plateau in imprecision with a critical number of items, where this number appears to correspond to capacity (Zhang and Luck 2008). Consequently, several hybrid hypotheses have been presented, accounting for data that cannot be explained by Slot or Resource alone (van den Berg et al. 2012; Zhang and Luck 2008).

The above work has been invaluable in characterizing the storage limitations of WM but does not speak to its neural basis. WM is widely believed to be supported by "attractor states" in neocortex, emerging from recurrent excitation and feedback inhibition in local circuits. Under this framework, recurrent excitation sustains neural firing in the absence of driving stimuli (persistent activity), while feedback inhibition prevents this activity from running away (see Wang 2001). If R is instantiated by the cortical tissue mediating a task-relevant feature domain, e.g., spatial location, then feedback inhibition necessarily constrains capacity, since WM items will compete for representational space (see Franconeri et al. 2013). If so, R cannot be infinitely divisi-

Address for reprint requests and other correspondence: D. Standage, School of Psychology, University of Birmingham, Birmingham, B15 2TT UK (e-mail: d.standage@bham.ac.uk).

ble. Rather, it will be allocated to $n \leq K$ items, with capacity K determined by the properties of feedback inhibition, e.g., its strength and breadth. In other words, the simultaneous encoding of an arbitrary number of WM items is incompatible with feedback inhibition between stimulus-selective neural populations, a fundamental principle of neural information processing. The application of these principles to Resource leads to a strong hypothesis: a decrease in precision with increased memory load must be limited by capacity.

Here we test this hypothesis with a biophysically based model of a local circuit in posterior parietal cortex (PPC), a cortical area extensively correlated with WM (Christophel et al. 2012; Gnadt and Andersen 1988; Palva et al. 2010; Salazar et al. 2012; Todd and Marois 2004). Previous studies have used similar models to offer mechanistic explanations for capacity (Edin et al. 2009), precision (Almeida et al. 2015), and their relationship (Roggeman et al. 2014; Wei et al. 2012), but these studies did not explain precision under the principles of Resource. Rather, they equated imprecision with the "drift" of item-encoding neural populations in cortical tissue. As such, they make predictions different from those of our model (see DISCUSSION). According to Resource, imprecision reflects the signal-to-noise ratio (SNR) of neural representations. We extend this hypothesis from SNR to coding fidelity more generally, measuring the regularity and reliability of simulated spiking activity. In doing so, we demonstrate and explain the deterioration of coding fidelity with increasing n under established statistical measures, where this deterioration levels off at a critical n. Thus we offer a novel explanation for resource-like coding and its relationship with capacity, unifying a large body of neural and behavioral data and making specific predictions for experimental testing.

METHODS

Our local-circuit PPC model is a network of simulated pyramidal neurons and inhibitory interneurons, connected by AMPA (AMPAR), NMDA (NMDAR), and GABA (GABAR) receptor conductance synapses (Fig. 1*A*). Synaptic connectivity within and between classes of neuron was structured according to in vitro data, including structured and unstructured components of the connectivity to pyramidal neurons from interneurons (Fig. 1*B*; see *Parameter Values*). We refer to the former and latter components as local and broad inhibition, respectively.

We ran simulations of two common visual and WM tasks, a visually guided delayed saccade task (the visual task) and a memory-guided delayed-saccade task (the memory task) (e.g., Paré and Wurtz 1997). Each task consists of three intervals: a pretrial interval, a stimulus interval, and a delay interval. After the pretrial interval, items are presented during the stimulus interval on both tasks, remaining present during the delay interval on the visual task but not the memory task (Fig. 1*C*). We constrained the model by setting its parameter values according to anatomical and physiological data (see *Parameter Values*) and by stipulating that it must qualitatively reproduce signature neural data from PPC (see RESULTS; Fig. 2). We then measured its storage capacity and coding fidelity as a function of n. Capacity was defined as the mean number of accurately encoded items during the last 300 ms of the delay interval (the statistics window), where accurate encoding was determined by the rate, position (relative to stimulus position), and discriminability of item-encoding populations. We used three standard measures of coding fidelity: the SNR of stimulus-selective spiking, the coefficient of variation (CV) of interspike intervals (ISIs), and the Fano factor (FF) of between-trial spike counts. SNR quantifies the degree to which selective spiking is discriminable from baseline activity, while CV and FF quantify the within-trial regularity and between-trial reliability of spiking, respectively.

The Network Model

The local-circuit model is a fully connected network of leaky integrate-and-fire neurons (Tuckwell 1988), comprised of $N^{\mathrm{p}} = 400$ simulated pyramidal neurons and $N^{\mathrm{i}} = N^{\mathrm{p}}/4$ fast-spiking inhibitory interneurons (putative basket cells). Each model neuron is described by

$$C_{\mathrm{m}}^{\{\mathrm{p,i}\}}\frac{\mathrm{d}V}{\mathrm{d}t} = -g_{\mathrm{L}}^{\{\mathrm{p,i}\}}(V - E_{\mathrm{L}}^{\{\mathrm{p,i}\}}) - I \tag{1}$$

where C_{m} is the membrane capacitance of the neuron, g_{L} is the leakage conductance, V is the membrane potential, E_{L} is the equilibrium potential, and I is the total input current. When V reaches a threshold ϑ_v, it is reset to V_{res}, after which it is unresponsive to its input for an absolute refractory period of τ_{ref}. Here and below, superscripts p and i refer to pyramidal neurons and interneurons, respectively, indicating that parameter values are assigned separately to each class of neuron.

The total input current at each neuron is given by

$$I = I^{\mathrm{sel}} + I^{\mathrm{rec}} + I^{\mathrm{back}} \tag{2}$$

where I^{sel} is stimulus-selective synaptic current (set to 0 for interneurons), I^{rec} is recurrent (intrinsic) synaptic current, and I^{back} is background current. I^{sel} and I^{rec} are comprised of synaptic currents, and I^{back} is comprised of synaptic current and injected current. Synaptic currents driven by pyramidal neuron spiking are mediated by simulated AMPAR and/or NMDAR conductances, and synaptic currents driven by interneuron spiking are mediated by simulated GABAR conductances. For AMPAR and GABAR currents, synaptic activation (the proportion of open channels) is defined by

$$\begin{aligned}\frac{\mathrm{d}g_{\mathrm{AMPA}}^{\mathrm{a}}}{\mathrm{d}t} &= -\frac{g_{\mathrm{AMPA}}^{\mathrm{a}}}{\tau_{\mathrm{AMPA}}^{\{\mathrm{p,i}\}}} + \delta(t - t_{\mathrm{f}}) \\ \frac{\mathrm{d}g_{\mathrm{GABA}}^{\mathrm{a}}}{\mathrm{d}t} &= -\frac{g_{\mathrm{GABA}}^{\mathrm{a}}}{\tau_{\mathrm{GABA}}^{\{\mathrm{p,i}\}}} + \delta(t - t_{\mathrm{f}})\end{aligned} \tag{3}$$

where τ_{AMPA} and τ_{GABA} are the time constants of AMPAR and GABAR deactivation respectively, δ is the Dirac delta function, t_{f} is the time of firing of a presynaptic neuron and superscript a indicates that synapses are activated by different sources of spiking activity (selective, recurrent, and background). NMDAR activation has a slower rise and decay and is described by

$$\frac{\mathrm{d}g_{\mathrm{NMDA}}^{\mathrm{a}}}{\mathrm{d}t} = -\frac{g_{\mathrm{NMDA}}^{\mathrm{a}}}{\tau_{\mathrm{NMDA}}^{\{\mathrm{p,i}\}}} + \alpha_{\mathrm{NMDA}} \cdot \omega_{\mathrm{NMDA}}(1 - g_{\mathrm{NMDA}}^{\mathrm{a}}) \tag{4}$$

where τ_{NMDA} is the time constant of receptor deactivation and α_{NMDA} controls the saturation of NMDAR channels at high presynaptic spike frequencies. The slower opening of NMDAR channels is captured by

$$\frac{\mathrm{d}\omega_{\mathrm{NMDA}}}{\mathrm{d}t} = -\frac{\omega_{\mathrm{NMDA}}}{\tau_{\omega}} + \delta(t - t_{\mathrm{f}}) \tag{5}$$

where τ_{ω} determines the rate of channel opening.

Intrinsic (recurrent, local feedback) synaptic current to each neuron j is defined by

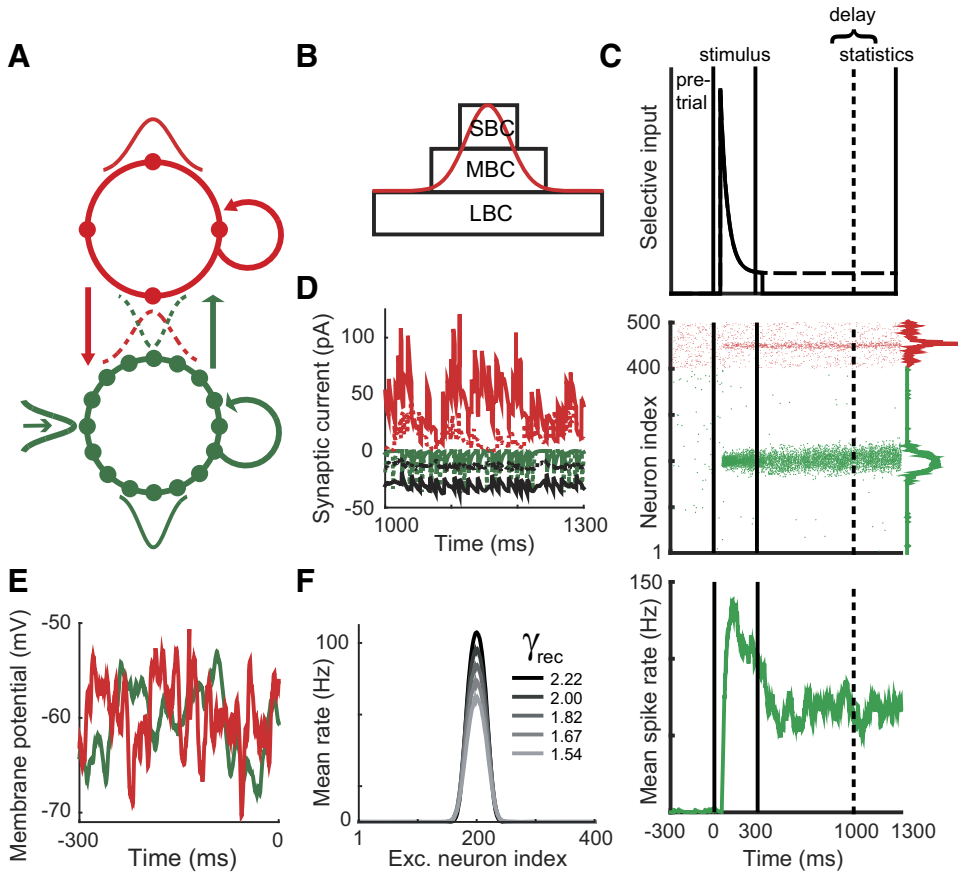

Fig. 1. Local-circuit posterior parietal cortex (PPC) model and simulated tasks. *A*: schematic of the model. Solid circles depict pyramidal neurons (green) and inhibitory interneurons (red), arranged periodically by their connectivity structures. The 4-to-1 ratio of pyramidal neurons to interneurons preserves their population sizes in the model. Arced and straight arrows depict synaptic connectivity within and between classes of neuron, respectively. Thin Gaussian curves depict the structure of this connectivity (within, solid; between, dotted). The thick Gaussian curve depicts the response field of a pyramidal neuron. Red arrows depict GABA receptor (GABAR) synapses, thin green arrow depicts AMPA receptor (AMPAR)-only synapses, and wide green arrow depicts synapses with AMPARs and NMDA receptors (NMDARs). *B*: approximation of synaptic connections onto pyramidal neurons and interneurons from small, medium, and large basket cells (SBC, MBC, and LBC respectively). Rectangles depict unstructured connectivity within each class of cell and onto pyramidal neurons from each class. Red curve approximates their combined structure. *C*, *top*: the visual and memory tasks are comprised of a pretrial interval, a stimulus interval, and a delay interval. Spiking statistics are taken during the last 300 ms of the delay interval, referred to as the statistics window. Stimulus onset follows a 50-ms visual response delay. On the visual (memory) task, stimuli persist (do not persist) throughout the delay interval, depicted by the dashed horizontal line. The decaying input signal simulates upstream response adaptation. *Middle* and *bottom*: example trial of the 1-item memory task. *Middle*: in the raster plot, pyramidal neurons and interneurons are indexed from 1 to 400 and from 401 to 500, respectively. Mean spike density function (SDF; see text) over all pyramidal neurons and interneurons during the statistics window is shown on *right*. *Bottom*: mean SDF over the item-encoding pyramidal population. *D*: synaptic currents onto a pyramidal neuron (solid) and an interneuron (dotted) during the delay interval of the 1-item memory task. Red, green, and black curves show GABAR, AMPAR, and NMDAR currents, respectively. *E*: membrane potential of a pyramidal neuron and an interneuron during the pretrial interval. *F*: mean rate over all pyramidal neurons during the statistics window of correct trials on the memory task for each value of control parameter $\gamma_{rec} = 1/\gamma_g$ (see text).

$$
\begin{aligned}
I_j^{\text{rec}} &= I_{\text{AMPA},j}^{\text{rec}} + I_{\text{NMDA},j}^{\text{rec}} + I_{\text{GABA},j}^{\text{rec}} \\
I_{\text{AMPA},j}^{\text{rec}} &= \sum_k \frac{G_{\text{AMPA}}^{\{\text{p,i}\}}}{\gamma_g} \cdot g_{\text{AMPA},k}^{\text{rec}}(V_j - V_{\text{E}}) \cdot W_{j,k}^{\text{rec}\,|\,\text{pp,ip}} \\
I_{\text{NMDA},j}^{\text{rec}} &= \sum_k \frac{G_{\text{NMDA}}^{\{\text{p,i}\}}}{\gamma_g} \cdot g_{\text{NMDA},k}^{\text{rec}}(V_j - V_{\text{E}}) \cdot \eta_j \cdot W_{j,k}^{\text{rec}\,|\,\text{pp,ip}} \\
I_{\text{GABA},j}^{\text{rec}} &= \sum_k \frac{G_{\text{GABA}}^{\{p,i\}}}{\gamma_g} \cdot g_{\text{GABA},k}^{\text{rec}}(V_j - V_{\text{I}}) \cdot W_{j,k}^{\text{rec}\,|\,\text{pi,ii}}
\end{aligned}
\tag{6}
$$

where γ_g is a scale factor controlling the relative strength of extrinsic and intrinsic synaptic conductance; G_{AMPA}, G_{NMDA}, and G_{GABA} are the respective strengths of AMPAR, NMDAR, and GABAR conductance; V_{E} is the reversal potential for AMPARs and NMDARs, and V_{I} is the reversal potential for GABARs; $g_{\text{AMPA},k}^{\text{rec}}$, $g_{\text{NMDA},k}^{\text{rec}}$, and $g_{\text{GABA},k}^{\text{rec}}$ are the activation of AMPAR, NMDAR, and GABAR receptors, respectively, by presynaptic neurons k; η governs the voltage dependence of NMDARs; and matrices $W^{\text{rec}\,|\,\text{pp,ip}}$ and $W^{\text{rec}\,|\,\text{pi,ii}}$ scale conductance strength or weight according to the connectivity structure of the network. This structure depends on the class of neuron receiving and projecting spiking activity, where superscripts pp, ip, pi, and ii denote connections to pyramidal neurons from pyramidal neurons, to interneurons from pyramidal neurons, to pyramidal neurons from interneurons, and to interneurons from interneurons, respectively. For each of these structures $s \in \{\text{pp, ip, pi, ii}\}$, $W^{\text{rec}\,|\,s}$ is a Gaussian function of the distance between periodically arranged neurons, where the weight $W_{j,k}^{\text{rec}\,|\,s}$ to neuron j from neuron k is given by

$$
W_{j,k}^{\text{rec}\,|\,s} = e^{-d^2/2\sigma_{\text{rec}\,|\,s}^2} \cdot (1 - \zeta_{\text{rec}\,|\,s}) + \zeta_{\text{rec}\,|\,s} \tag{7}
$$

The distance between neurons is defined by $d = min(|j - k|\Delta x^{\text{p}}, 2\pi - |j - k|\Delta x^{\text{p}})$ for $W^{\text{rec}\,|\,\text{pp}}$, $d = min(|j - k|\Delta x^{\text{i}}, 2\pi - |j - k|\Delta x^{\text{i}})$ for $W^{\text{rec}\,|\,\text{ii}}$, $d = min(|j - z^{\text{pi}}|\Delta x^{\text{p}}, 2\pi - |j - z^{\text{pi}}|\Delta x^{\text{p}})$ for $W^{\text{rec}\,|\,\text{pi}}$, and $d = min(|j - z^{\text{ip}}|\Delta x^{\text{i}}, 2\pi - |j - z^{\text{ip}}|\Delta x^{\text{i}})$ for $W^{\text{rec}\,|\,\text{ip}}$, with scale factors $\Delta x^{\text{p}} = 2\pi/N^{\text{p}}$ and $\Delta x^{\text{i}} = 2\pi/N^{\text{i}}$. For $W^{\text{rec}\,|\,\text{pi}}$ and $W^{\text{rec}\,|\,\text{ip}}$, $z^{\text{pi}} = N^{\text{p}}/$

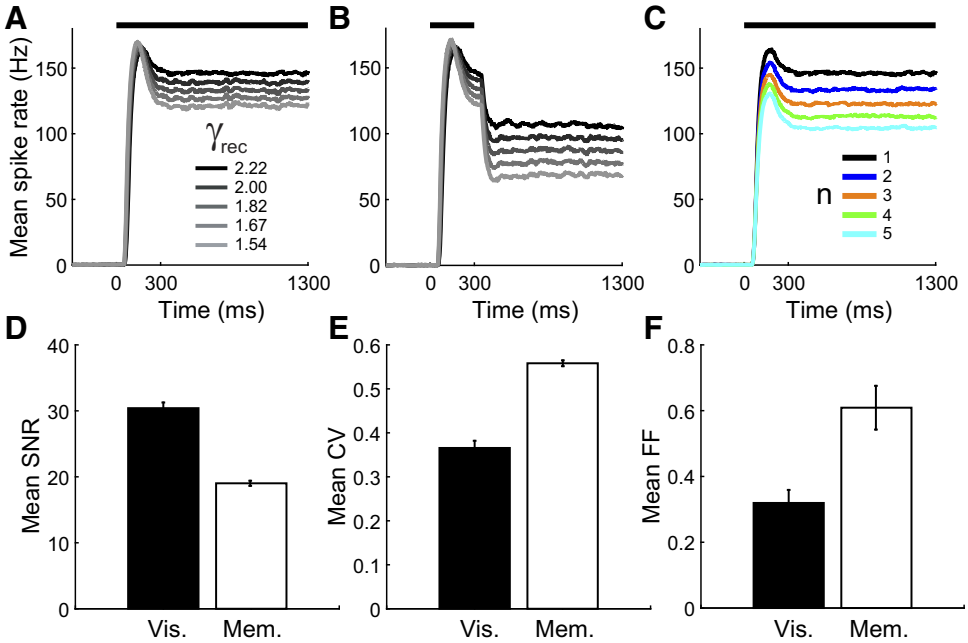

Fig. 2. The model qualitatively reproduces signature neural data recorded from posterior parietal cortex (PPC) during 1-item visual and memory tasks and multiple-item visual tasks. *A*: mean activity at the response field (RF) center of the item-encoding population on the 1-item visual task for each value of control parameter $\gamma_{rec} = 1/\gamma_g$ (each gain condition, see text). Darker shades correspond to higher γ_{rec}. *B*: mean activity at the RF center on the 1-item memory task. *C*: mean activity at the RF center of a single item-encoding population on the *n*-item visual task for all n ($1 \leq n \leq 5$). Results are shown for the highest-gain condition. Thick horizontal bars at *top* of *A–C* show the timing of the target stimuli. *D–F*: persistent activity in the model encodes a low-fidelity representation of an earlier stimulus, characterized by a lower signal-to-noise ratio (SNR; *D*), higher coefficient of variation (CV; *E*), and higher Fano factor (FF; *F*) during the memory delay than during the visual delay. Error bars show SE. Results are shown for the lowest-gain condition.

$N^i \times k$ and $z^{ip} = N^i/N^p \times k$, respectively. Parameter $\sigma_{rec\,|\,s}$ determines the spatial extent of connectivity, and parameter $\zeta_{rec\,|\,s}$ allows the inclusion of a baseline weight, with the function normalized to a maximum of 1 ($0 \leq \zeta_{rec\,|\,s} < 1$).

Background Activity

For each neuron, in vivo cortical background activity is simulated by current I^{back}, defined by

$$I^{back} = I^{back,syn} + I^{back,inj} \tag{8}$$

where $I^{back,syn}$ is driven by synaptic bombardment and $I^{back,inj}$ is noisy current injection. The former is generated by AMPAR synaptic activation, where independent, homogeneous Poisson spike trains are provided to all neurons at rate μ_{back}. $I^{back,syn}$ is therefore defined by

$$I^{back,syn} = \gamma_g \cdot \lambda \cdot G^{\{p,i\}}_{AMPA} \cdot g^{back}_{AMPA}(V - V_E) \tag{9}$$

where λ is a scale factor and g^{back}_{AMPA} is given in *Eq. 3*.

For $I^{back,inj}$, we used the point-conductance model by Destexhe et al. (2001):

$$I^{back,inj} = g_e(t)(V - V_E) + g_i(t)(V - V_I) \tag{10}$$

The time-dependent excitatory and inhibitory conductances $g_e(t)$ and $g_i(t)$ are updated at each timestep Δt according to

$$g_e(t + \Delta t) = g0_e + [g_e(t) - g0_e] \cdot e^{-\Delta t/\tau_e} + A_e \Upsilon \tag{11}$$

and

$$g_i(t + \Delta t) = g0_i + [g_i(t) - g0_i] \cdot e^{-\Delta t/\tau_i} + A_i \Upsilon \tag{12}$$

respectively, where $g0_e$ and $g0_i$ are average conductances, τ_e and τ_i are time constants, and Υ is normally distributed random noise with 0 mean and unit standard deviation. Amplitude coefficients A_e and A_i are defined by

$$A_e = \sqrt{\frac{D_e \tau_e}{2}\left[1 - \exp\left(\frac{-2\Delta t}{\tau_e}\right)\right]} \tag{13}$$

and

$$A_i = \sqrt{\frac{D_i \tau_i}{2}\left[1 - \exp\left(\frac{-2\Delta t}{\tau_i}\right)\right]} \tag{14}$$

respectively, where $D_e = 2\sigma_e^2/\tau_e$ and $D_i = 2\sigma_i^2/\tau_i$ are noise "diffusion" coefficients. See Destexhe et al. (2001) for the derivation of these equations.

Experimental Design and Statistical Analysis

Simulated experimental tasks. We simulated the target stimuli in both tasks by providing independent, homogeneous Poisson spike trains to all pyramidal neurons j in the network, where spike rates were drawn from a normal distribution with mean μ_{sel} corresponding to the center of a Gaussian response field (RF) defined by $W^{rf}_{j,k} = \exp(-d^2/2\sigma^2_{rf})$. Constant d is given above for recurrent synaptic structure $W^{rec\,|\,pp}$, σ_{rf} determines the width of the RF, and subscript k indexes the neuron at the RF center. Spike response adaptation by upstream visually responsive neurons was modeled by a step-and-decay function

$$\mu_{sel}(t) = \begin{cases} (\mu_{init} - \mu_{init}/\mu_{div})e^{-(t-t_{vrd})/\tau_\mu} + \mu_{init}/\mu_{div} & \text{for } t > t_{vrd} \\ 0 & \text{for } t \leq t_{vrd} \end{cases} \tag{15}$$

where μ_{init} determines the initial spike rate, μ_{div} determines the asymptotic rate, τ_μ determines the rate of upstream response adaptation, and t_{vrd} is a visual response delay. We simulated the visual task by providing these selective spike trains for 1,300 ms, following the 300-ms pretrial interval. We simulated the memory task by providing the selective spike trains for 300 ms, following the pretrial interval and followed by a 1,000-ms delay (Fig. 1*C*). The stimuli were mediated by AMPARs only, so for all pyramidal neurons j in the PPC network,

$$I^{sel}_j = \gamma_g \cdot \lambda \cdot G^{p}_{AMPA} \cdot g^{sel}_{AMPA,j}(V_j - V_E) \cdot W^{rf}_{j,k} \tag{16}$$

All simulations were run with the standard implementation of Euler's forward method, where the timestep was $\Delta t = 0.25$ ms.

Values of model parameters are provided in Table 1 and justified in *Parameter Values*.

Determining working memory performance. We ran 400 trials of the visual and memory tasks with 1–5 stimuli (henceforth the *n*-item

Table 1. *Model parameters*

Parameter	Pyramidal Neurons	Interneurons	Description
C_m	0.5 nF	0.2 nF	Membrane capacitance
g_L	25 nS	20 nS	Leakage conductance
E_L	–70 mV	–70 mV	Leakage equilibrium potential
ϑ_v	–50 mV	–50 mV	Spike threshold
V_{res}	–60 mV	–60 mV	Reset potential
τ_{ref}	2 ms	1 ms	Absolute refractory period
V_E	0 mV	0 mV	Reversal potential for AMPARs and NMDARs
V_I	–70 mV	–70 mV	Reversal potential for GABARs
τ_ω	2 ms	2 ms	Time constant of channel opening for NMDARs
α_{NMDA}	0.5 kHz	0.5 kHz	Saturation of NMDAR channels
Mg	1 mM	1 mM	Extracellular magnesium concentration
G_{AMPA}	0.2 nS	0.4 nS	Conductance strength of AMPARs
G_{NMDA}	4 nS	2 nS	Conductance strength of NMDARs
G_{GABA}	1.5 nS	0.75 nS	Conductance strength of GABARs
τ_{AMPA}	4 ms	2 ms	Time constant of deactivation of AMPARs
τ_{NMDA}	100 ms	50 ms	Time constant of deactivation of NMDARs
τ_{GABA}	10 ms	10 ms	Time constant of deactivation of GABARs
$\sigma_{rec\mid pp,ip}$	0.2	0.2	Width of connectivity from pyramidal neurons
$\sigma_{rec\mid pi,ii}$	0.4	0.4	Width of connectivity from interneurons
$\zeta_{rec\mid pp,ip}$	0	0	Unstructured connectivity from pyramidal neurons
$\zeta_{rec\mid pi,ii}$	1/3	1/3	Unstructured connectivity from interneurons
$g0_e$	2.5 nS	2.5 nS	Average exc. background conductance
$g0_i$	12.5 nS	12.5 nS	Average inh. background conductance
τ_e	2.5 ms	2.5 ms	Time constant of exc. background conductance
τ_i	10 ms	10 ms	Time constant of inh. background conductance
σ_e	5 nS	5 nS	Standard deviation of exc. diffusion coefficient
σ_i	12.5 nS	12.5 nS	Standard deviation of inh. diffusion coefficient
σ_{rf}	0.1		Width of response fields
μ_{init}	10,000/γ_g Hz		Initial (aggregate) spike rate to RF center
μ_{div}	10		Divisor for upstream response adaptation
τ_μ	50 ms		Time constant of upstream response adaptation
t_{vrd}	50 ms		Visual response delay
γ_g	0.45–0.65	0.45–0.65	Control parameter determining gain condition
γ_{rec}	1/γ_g	1/γ_g	Inverse of control parameter

Synaptic connectivity parameters (pp, ip, pi, ii, where p indicates pyramidal neuron and i indicates interneuron) are indexed to a receiving neuron from a transmitting neuron. AMPAR, AMPA receptor; NMDAR, NMDA receptor; GABAR, GABA receptor; exc, excitatory; inh, inhibitory; RF, response field. See *Parameter Values*.

visual and memory tasks; $1 \leq n \leq 5$). To determine WM performance on each trial of the memory task, spike density functions (SDFs) were calculated for all pyramidal neurons in the network by convolving their spike trains with a rise-and-decay function

$$\frac{(1 - e^{-t/\tau_r}) \cdot e^{-t/\tau_d}}{\frac{\tau_d^2}{\tau_r + \tau_d}} \quad (17)$$

where t is the time after stimulus onset and $\tau_r = 1$ ms and $\tau_d = 20$ ms are the time constants of rise and decay, respectively (Standage and Paré 2011; Thompson et al. 1996). On each n-item trial, we calculated the mean of the SDFs over the last 300 ms of the delay, obtaining the average activity over the network, and then partitioned the network into n equal regions. The location of each item was centered within each region. We then fit the mean activity in each region with a Gaussian function with four parameters: the height of the peak, the position of the peak, the standard deviation (SD) (controlling width), and the height that is approached asymptotically from the peak. An item was considered accurately stored if the fitted Gaussian satisfied three criteria: the height parameter exceeded 30 Hz, the difference between the height and the fitted asymptote on both sides of the peak exceeded 15 Hz, and the position parameter was within $\Delta c = 10°$ of the center of the RF for that item. For the first criterion, we chose 30 Hz because this spike rate implies ~10 spikes during the 300-ms statistics window, as required to faithfully calculate CV and FF (Nawrot 2010). The second criterion dictates that items are only considered accurately stored if the population response is discriminable. The third criterion ensures that the memory of the location of the item is close to the actual location, the precise value of which was not crucial to our results ($\Delta c >$ ~5).

Calculating spiking statistics. We selected $m = 20$ simulated pyramidal neurons from the network (the target neurons) and recorded their activity on m trials each. This population of neurons consisted of the neuron at the center of the RF for a given target and the $m - 1$ neurons closest to the RF center. For each of the two tasks, the SNR of each target neuron was calculated on each trial by subtracting the spike count during the 300-ms pretrial interval from the spike count during the statistics window and dividing the result by the latter [$SNR = (SC_{del} - SC_{pre})/SC_{pre}$, where SC is the spike count].

The CV of ISI was calculated for each target neuron on each trial by dividing the mean ISI by the SD of ISI during the statistics window ($CV = \sigma_{del}^{ISI}/\mu_{del}^{ISI}$). The FF was calculated for each target neuron by recording the spike count during the statistics window on each trial and dividing the variance by the mean over all trials for that neuron ($FF = \sigma_{del}^{SC^2}/\mu_{del}^{SC}$). These statistics were only calculated for accurately stored items and from neurons that emitted at least 9 spikes during the 300-ms statistics window (30 Hz; see *Determining working memory performance*). To increase statistical power on memory trials with $n >$ 1 items, if the network did not accurately store the "first" item we searched for a corresponding neuron in another item-encoding population, where correspondence was determined relative to the RF center, e.g., if the target neuron was located 3 indices below the RF center of *item 1*, we used a neuron located 3 indices below

the RF center of another item. If no items were accurately stored, the trial was discarded for statistical purposes.

Parameter Values

In setting parameter values in the model, our aim was to justify every value by anatomical and physiological data, thus constraining our choices as much as possible, and then to use a single control parameter to explore the model's performance and spiking statistics on the visual and memory tasks. Our control parameter was γ_g (*Eqs. 6*, *9*, and *16*), governing the relative strengths of extrinsic and intrinsic synaptic conductance and therefore the strength of recurrent processing.

For cellular parameters, we used standard values for integrate-and-fire neurons in cortical simulations (Compte et al. 2000), justified by electrophysiological data in earlier related work (Troyer and Miller 1997; Wang 1999). These values are $C_m^p = 0.5$ nF, $g_L^p = 25$ nS, $E_L^p = -70$ mV, $\vartheta_v^p = -50$ mV, $V_{res}^p = -60$ mV, and $\tau_{ref}^p = 2$ ms and $C_m^i = 0.2$ nF, $g_L^i = 20$ nS, $E_L^i = -70$ mV, $\vartheta_v^i = -50$ mV, $V_{res}^i = -60$ mV, and $\tau_{ref}^i = 1$ ms. Likewise, synaptic reversal potentials are $V_E = 0$ mV and $V_I = -70$ mV, and the parameters governing the opening and saturation of NMDARs are $\tau_\omega = 2$ ms and $\alpha_{NMDA} = 0.5$ kHz, respectively (Compte et al. 2000). The voltage dependence of NMDARs is given by $\eta = 1/[1 + Mg \cdot \exp(-0.062 \cdot V)/3.57]$, where $Mg = 1$ mM is the extracellular magnesium concentration and V is measured in millivolts (Jahr and Stevens 1990).

In setting parameters for the conductance strengths and time constants of decay of AMPARs and NMDARs, we followed Standage et al. (2013), emphasizing fast inhibitory recruitment in response to slower excitation (see Povysheva et al. 2006 for discussion). For AMPARs $G_{AMPA}^p = 0.2$ nS, $G_{AMPA}^i = 2G_{AMPA}^p$, $\tau_{AMPA}^p = 4$ ms, and $\tau_{AMPA}^i = \tau_{AMPA}^p/2$, and for NMDARs $G_{NMDA}^p = 4$ nS, $G_{NMDA}^i = G_{NMDA}^p/2$, $\tau_{NMDA}^p = 100$ ms, and $\tau_{NMDA}^i = \tau_{NMDA}^p/2$. These values produce fast-decaying AMPAR currents on the order of 10 pA (Angulo et al. 1999; Desai et al. 2002) that are stronger and shorter lived onto inhibitory interneurons than onto pyramidal neurons (Hestrin 1993; Hull et al. 2009; McBain and Fisahn 2001) and slow-decaying NMDAR currents on the order of 10 pA (Angulo et al. 1999; Berretta and Jones 1996) that are stronger and longer lived at synapses onto pyramidal neurons than onto inhibitory interneurons (Hull et al. 2009). For GABARs, $G_{GABA}^p = 1.5$ nS and $G_{GABA}^i = G_{GABA}^p/2$, producing GABAR currents several times stronger than the above excitatory currents, where the stronger conductance at synapses onto pyramidal neurons captures their greater prevalence of GABARs (Markram et al. 2004). GABAR time constants were set to $\tau_{GABA}^p = \tau_{GABA}^i = 10$ ms (Salin and Prince 1996; Xiang et al. 1998). Example synaptic currents are shown in Fig. 1*D*.

The connectivity structures $W^{rec\,|\,pp,ip,pi,ii}$ capture the probability of lateral synaptic contact within and between classes of neurons in local cortical circuitry (Somers et al. 1995; Wilson and Cowan 1973). A considerable volume of data indicates that the probability of lateral synaptic contact between cortical pyramidal neurons is normally distributed with mean 0 and half-width of ~0.25 mm (Berger et al. 2009; Hellwig 2000; Voges et al. 2010). Thus $\sigma_{rec\,|\,pp}$ corresponds to 0.25 mm, determining the size of the cortical region being modeled, and $\zeta_{rec\,|\,pp} = 0$. We are unaware of any data suggesting that the lateral projections of pyramidal neurons target basket cells differently than they target other pyramidal neurons, so we set $\sigma_{rec\,|\,ip} = \sigma_{rec\,|\,pp}$ and $\zeta_{rec\,|\,ip} = \zeta_{rec\,|\,pp}$. Arguably, $\sigma_{rec\,|\,ip}$ should be narrower than $\sigma_{rec\,|\,pp}$, since the dendritic trees of basket cells are less extensive than those of pyramidal neurons, but setting these parameters to equal values supported more stable network dynamics, i.e., it furnished sufficient local-circuit inhibition for the model to simulate the experimental tasks without modifications to other parameter values.

For connectivity structures $W^{rec\,|\,pi,ii}$, values for $\sigma_{rec\,|\,pi,ii}$ and $\zeta_{rec\,|\,pi,ii}$ are justified by four premises: first, we assume that basket cells are a major source of lateral inhibition (Krimer and Goldman-Rakic 2001), and we limit our focus to this class of inhibitory interneuron; second, basket cells synapse onto the somatic and perisomatic regions of their targets (see Markram et al. 2004); third, the axons of basket cells contact their targets indiscriminately throughout the range of their ramifications (Packer and Yuste 2011); and fourth, the basket cell population can be divided into small (local arbor), medium (medium arbor), and large (wide arbor) cells in equal proportion, i.e., one-third each (Krimer et al. 2005). Under the first and second premises, we do not need to consider the dendritic morphology of the targets of inhibitory interneurons. Under the second and third premises, we assume a uniform synaptic distribution for inhibitory targets, where the axonal ramifications of small, medium, and large basket cells cover progressively larger areas (Krimer et al. 2005; Krimer and Goldman-Rakic 2001), with large basket cells (LBCs) covering the entire local circuit (Kisvárday et al. 1993; Markram et al. 2004). We therefore approximate this connectivity structure by setting $\sigma_{rec\,|\,pi} = \sigma_{rec\,|\,ii} = 2\sigma_{rec\,|\,pp}$ and $\zeta_{rec\,|\,pi} = \zeta_{rec\,|\,ii} = 1/3$, where the former corresponds to a half-width of ~0.5 mm (cf. Kisvárday et al. 1993; Krimer et al. 2005; Krimer and Goldman-Rakic 2001) and the latter refers to the 1/3 proportion of LBCs. This approach to determining inhibitory connectivity parameters is depicted in Fig. 1*B*. We set $\sigma_{rec\,|\,pp} = 0.2$ because this value supported the simultaneous representation of 5 simulated visual stimuli, corresponding to the upper limit on human WM capacity, i.e., 4 ± 1 items (Cowan 2001; Luck and Vogel 1997). Finally, we set $\sigma_{rec\,|\,ii} = \sigma_{rec\,|\,pi}$ because LBCs make extensive contacts onto one another over the full range of their axonal ramifications (Kisvárday et al. 1993). Note that we do not attribute biological significance to the spatial periodicity of the network. Rather, this arrangement allows the implementation of $W^{rec\,|\,pp,ip,pi,ii}$ with all-to-all connectivity without biases due to asymmetric lateral interactions between neurons and further captures the topographic mapping of spatially periodic stimuli in many visual (e.g., Thomas and Paré 2007) and WM (e.g., Funahashi et al. 1989; Matsuyoshi et al. 2014) tasks. In *Broad Feedback Inhibition Underlies Slot-Like Capacity and Resource-Like Coding*, we describe simulations with alternative configurations of inhibitory connectivity by varying $\zeta_{rec\,|\,pi,ii}$. In all cases, we retained the total conductance strength of inhibitory synapses onto pyramidal neurons by multiplying G_{GABA}^p by $\sum W_{alt}^{rec\,|\,pi}/\sum W_{def}^{rec\,|\,pi}$, where alt refers to a given alternative configuration and def refers to the "default" configuration described above. Similarly, we retained the total conductance strength of inhibitory synapses onto interneurons by multiplying G_{GABA}^i by $\sum W_{alt}^{rec\,|\,ii}/\sum W_{def}^{rec\,|\,ii}$. In other words, we normalized feedback inhibition according to the area under $W^{rec\,|\,pi,ii}$.

In setting parameter values for background activity in each network, we initially omitted background synaptic input $I^{back,syn}$ and followed the data by Fellous et al. (2003) to produce $I^{back,inj}$, where $g0_e = 5$ nS and $g0_i = 25$ nS, $\tau_e = 2.5$ ms, $\tau_i = 10$ ms, $\sigma_e = 5$ nS, and $\sigma_i = 12.5$ nS. Because the average inhibitory background conductance $g0_i$ is five times the average excitatory background conductance $g0_e$ (see Destexhe 2010), our simulated pyramidal neurons did not respond adequately to selective stimuli under these parameter values. We therefore reduced the average conductances by a factor of 2, setting $g0_e = 2.5$ nS, retaining the ratio of inhibitory to excitatory conductance strength $g0_i = 5 \cdot g0_e = 12.5$ nS, and simulating the "other half" of upstream cortical background activity by providing independent, homogeneous Poisson spike trains to all neurons in the network. As such, we assumed that each neuron forms ~10,000 synapses with upstream cortical neurons (Douglas et al. 2004) and that by dividing $g0_e$ and $g0_i$ by 2 we were effectively omitting ~5,000 background inputs. We therefore approximated 5,000 upstream cortical neurons firing at 1 Hz each by setting the rate of background Poisson spike trains to $\mu_{back} = 500$ Hz and setting the extrinsic synaptic scale factor to $\lambda = 10$, trading temporal summation for spatial summation (Prescott and De Koninck 2003; Standage et al.

2013). As noted above, background spike trains were provided to all pyramidal neurons and interneurons in each network, mediated by AMPARs on the assumption that spike trains converging on PPC (an association cortical area) are predominantly ascending. Evidence for AMPAR-mediated ascending activity is provided by Self et al. (2012). This approach simultaneously released the network model from the overly strong background inhibitory currents and implemented an established, biologically plausible form of gain modulation [balanced background inputs (Chance et al. 2002)], rendering the PPC network responsive to simulated visual stimuli. It should be noted that our parameter values for background current injection ($g0_e$, $g0_i$, τ_e, τ_i, σ_e, and σ_i) were based on recordings from pyramidal neurons (Fellous et al. 2003), but since we are unaware of any data to guide these parameters for inhibitory interneurons, we assigned them the same values for all neurons. The effect of this background activity on the membrane potential of a pyramidal neuron and an interneuron is shown in Fig. 1*E*.

For the target stimuli, the width of RFs was determined by $\sigma_{rf} = \sigma_{pp}/2$. This narrow width captures the less extensive dendritic branching in cortical (input) layer 4 compared with layers 2/3 and 5 (see above for justification of lateral connectivity in the model). The initial spike rate at the RF center (*Eq. 15*) was $\mu_{init} = 10{,}000/\gamma_g$ Hz, which (for $\gamma_g = 1$) can be equated with, e.g., 100 upstream, visually responsive neurons firing at 100 Hz each, given our use of homogeneous, independent Poisson spike trains. It should be noted, however, that the synaptic scale factor $\lambda = 10$ probably renders this spike rate unrealistically high, since it implies, e.g., 1,000 upstream neurons firing at 100 Hz. Nonetheless, the high initial spike rate ensured a rapid-onset, high-rate visual response in the network for all processing regimes furnished by control parameter γ_g, as observed experimentally (e.g., Churchland et al. 2008; Paré and Wurtz 1997; Thomas and Paré 2007). Upstream, visual response adaptation was simulated by $\mu_{div} = 10$ and $\tau_\mu = 50$ ms. The former is somewhat extreme but allowed the rate of the initial population response in PPC to exceed the steady-state response on the visual task for all values of γ_g (e.g., Churchland et al. 2008; Paré and Wurtz 1997). Our use of γ_g as a denominator in determining μ_{init} supported stronger selective inputs when the network had stronger recurrent processing (smaller γ_g), allowing the rapid-onset, high-rate visual response described above. For larger γ_g, the network more readily gives way to its inputs, so a weaker input is sufficient to elicit a similar response. The visual response delay was $t_{vrd} = 50$ ms (Thomas and Paré 2007).

RESULTS

To systematically investigate network performance on the visual and memory tasks, we varied a single parameter γ_g, scaling the relative strength of intrinsic (*Eq. 6*) and extrinsic (*Eqs. 9* and *16*) synaptic conductance. We ran a block of trials for a range of values of this parameter (increments of 0.05), searching for values supporting a mean capacity of at least 0.95 items on the *n*-item memory task for $n \leq 5$ and for which all item-encoding populations on the 5-item task coexisted at the end of the stimulus interval (with excessively strong intrinsic synapses, feedback inhibition produced strong competition between populations, so that not all populations were extant at the onset of the memory delay). Thus we interpolated between upper and lower bounds on the strength of recurrent drive that support performance of the task, finding that our criteria were satisfied by $\gamma_g \in \{0.45, 0.5, 0.55, 0.6, 0.65\}$. We confirmed that these values support a stable background state (no structured activity before stimulus onset) by running a single trial with no stimuli for 10 s and that they support performance of the visual task (>99% of items were accurately encoded during the statistics window for all *n* and gain conditions). Because lower values of γ_g produce stronger recurrent drive and higher neuronal gain (Fig. 1*F*), it is convenient to define $\gamma_{rec} = 1/\gamma_g$. We refer to the values of γ_{rec} (equivalently γ_g) as the gain conditions of the network. A single trial of the 1-item memory task is shown in Fig. 1*C*.

The Model Complies with Signature Neural Data from PPC

Electrophysiological recordings from PPC show that on 1-item visual and memory tasks the rate of stimulus-selective activity is higher during the visual delay than the memory delay and on the memory task, the rate is higher during the stimulus interval than the memory interval (Paré and Wurtz 1997). More generally, PPC activity consistently shows several characteristics across visual tasks, including a rapid-onset, high-rate response that drops to a steady state before movement-related activity (e.g., Churchland et al. 2008; Louie et al. 2011; Paré and Wurtz 1997) and a decrease in rate with an increase in the number of stimuli (e.g., Churchland et al. 2008; Louie et al. 2011; Thomas and Paré 2007). Consistent with these data, the mean rate of stimulus-selective spiking in the model was higher during the visual delay than the memory delay on the 1-item tasks (Fig. 2, *A* and *B*) and was higher during the stimulus interval than the delay interval on the memory task (Fig. 2*B*). On the multiple-item visual tasks ($2 \leq n \leq 5$), selective spike rates were higher during the stimulus interval than the delay interval and the rate of stimulus-selective activity decreased with increasing *n* (Fig. 2*C*). These results were the case for all gain conditions, indicating that the model captured the relevant aspects of PPC processing over its full dynamic range.

For all gain conditions on the 1-item tasks, delay activity in the model had a lower SNR, a higher CV, and higher FF during the memory delay than the visual delay (Fig. 2, *D–F*). Thus persistent activity encoded a low-fidelity representation of the stimulus, as reported in monkey PPC (Johnston et al. 2009). Higher-gain conditions supported higher-fidelity encoding of the stimulus (see results for $n = 1$ in Fig. 5, *C–E*). Quantitative consideration of these results is provided in DISCUSSION.

Working Memory Performance in the Model Is Consistent with That of Monkeys and Humans

To measure WM performance, we calculated the mean number of accurately stored items on each *n*-item memory task, referring to this quantity as capacity $K(n)$. Example trials are shown in Fig. 3. Mean spike rates at the RF centers of all accurate item-encoding populations are shown in Fig. 4. Moderate- to high-gain conditions supported a maximum capacity ($\max[K(n)]$) of around 2 and 3 items, respectively (Fig. 5*A*), consistent with WM capacity in monkeys (Heyselaar et al. 2011) and humans (Luck and Vogel 2013; Vogel and Awh 2008). In keeping with earlier models of this class (Edin et al. 2009; Wei et al. 2012), $K(n)$ decreased beyond a critical *n* for all gain conditions, consistent with WM "overload" (Matsuyoshi et al. 2014). Indeed, for the range of *n* used here, overload was more pronounced in lower-capacity gain conditions, consistent with experimental data showing more pronounced overload among lower-capacity subjects (Fukuda et al. 2015; Linke et al. 2011; Matsuyoshi et al. 2014), on lower-capacity tasks (Xu 2007), and in lower-capacity conditions of the same task (Chee and Chuah 2007). Finally, capacity was roughly tracked

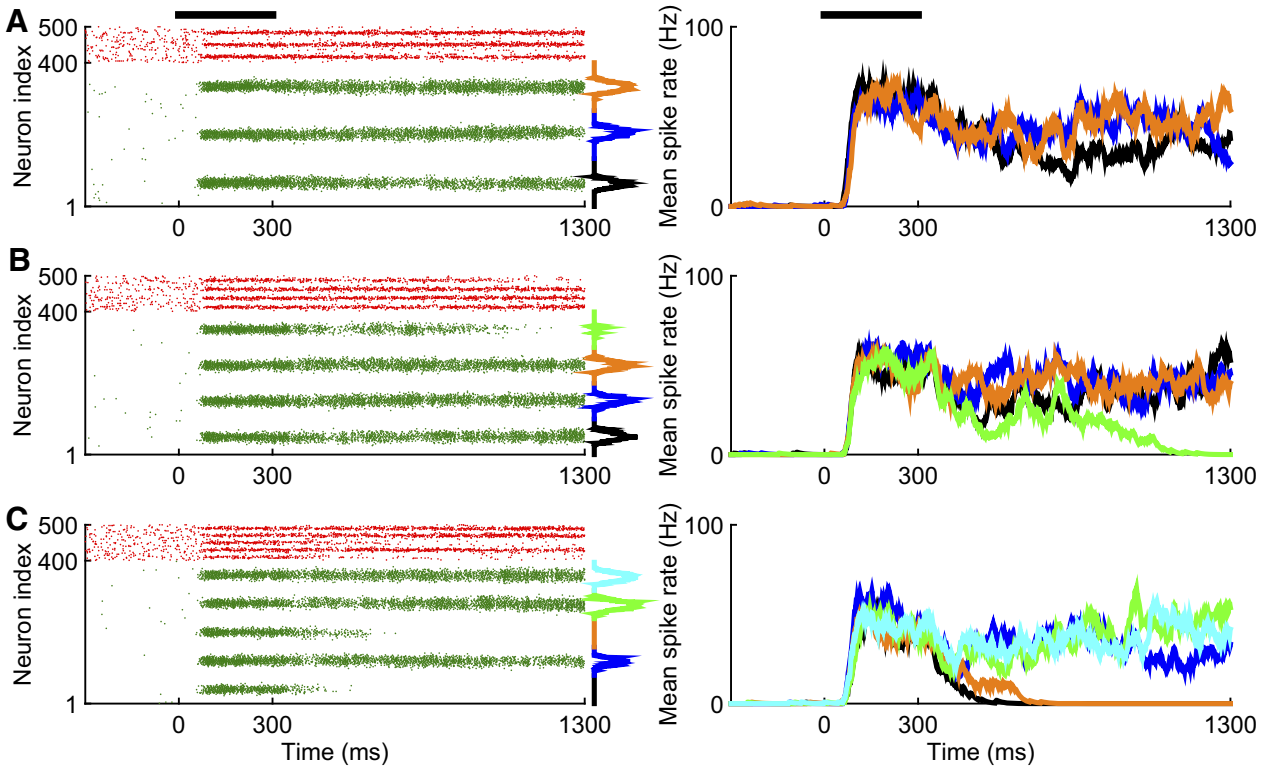

Fig. 3. Three trials of the multiple-item memory task for the highest-gain condition, with 3 (*A*), 4 (*B*), and 5 (*C*) items. The model accurately stored 3 items on each of these trials. Mean rates over all pyramidal neurons during the statistics window are inset on *right* of raster plots (see Fig. 1*C*), where shades match the mean spike density functions in *right* panels. Thick horizontal bars at *top* show the timing of the stimuli.

by the total pyramidal neuron activity in the network (Fig. 5*B*), similar to electroencephalogram (EEG) recordings from PPC (Vogel and Machizawa 2004).

Predictions for Experimental Testing and Their Implications for Slot and Resource

Having demonstrated that our simulations are qualitatively consistent with a range of electrophysiological and behavioral data from visual and WM tasks, we investigated the model's predictions for coding fidelity on multiple-item WM tasks and the implications of these predictions for Slot and Resource. For all gain conditions, the coding fidelity of persistent activity on the memory task deteriorated as the number of items increased from $n = 1$ to $n = 2$, characterized by a decrease in SNR and an increase in CV and FF (Fig. 5, *C–E*). With higher gain ($0.45 \leq \gamma_g \leq 0.55$), for which $K(3) > K(2)$, this reduction in coding fidelity continued as the number of items increased from $n = 2$ to $n = 3$, as measured by all three statistics. Within this range of γ_g, coding fidelity leveled off as the number of items increased beyond $n = 3$, roughly tracking $K(n)$ (Fig. 5*A*). This finding is strikingly consistent with behavioral data showing that WM precision decreases with increasing n until it reaches a plateau at around 3 or 4 items (Zhang and Luck 2008), although our model does not speak to the strategies by which subjects may guess the values of forgotten items on WM tasks, the assumptions of which can influence the interpretation of these data (see DISCUSSION). It also suggests that the same mechanism is responsible for constraints on capacity and resolution: the competition between item-encoding populations, mediated by broad inhibition.

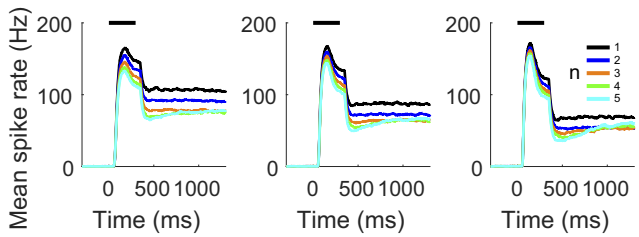

Fig. 4. Mean rates at the response field centers of all accurate item-encoding populations on the memory tasks for each n. Results are shown for the highest (*left*)-, median (*center*)-, and lowest (*right*)-gain conditions. Horizontal bars show the timing of the target stimuli.

Broad Feedback Inhibition Underlies Slot-Like Capacity and Resource-Like Coding

If broad inhibition is responsible for slot-like capacity and resource-like deterioration of coding fidelity, then removing it will increase $K(n)$ and eliminate the dependence of coding fidelity on n. We therefore removed the unstructured component of feedback inhibition ($\zeta_{\text{rec}\,|\,\text{pi,ii}} = 0$), preserving the total inhibitory conductance in the model by increasing the strength of local (structured) feedback inhibition (see *Parameter Values*). We then determined the corresponding upper and lower values of γ_g according to the criteria above and repeated our simulations of the memory task under these modified parameter values. As expected, these changes led to a dramatic increase in capacity ($\max[K(n)] > 4.2$ for all γ_g) and rendered coding fidelity roughly independent of n (Fig. 6*D*).

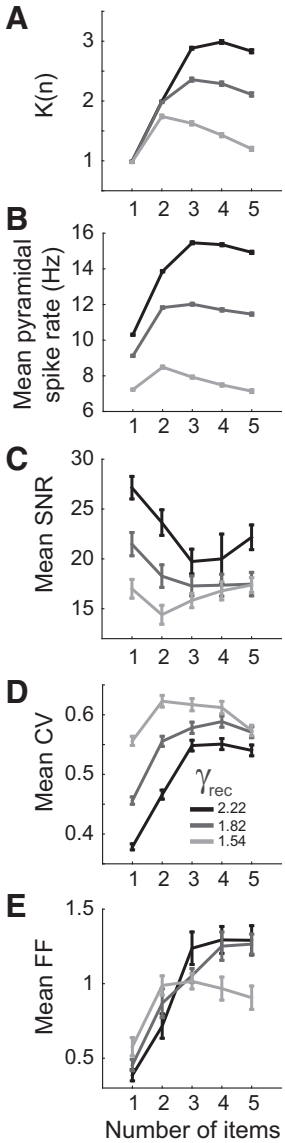

Fig. 5. *A*: mean number of items accurately stored on each *n*-item memory task, referred to as capacity $K(n)$. Error bars show SE. Results are shown for the highest-, median-, and lowest-gain conditions, where darker shades correspond to higher gain. Other gain conditions are omitted for clarity. *B*: mean spike rate of all pyramidal neurons in the model as a function of *n*. *C*: signal-to-noise ratio (SNR) over *n*. *D*: coefficient of variation (CV) over *n*. *E*: Fano factor (FF) over *n*. Statistics were taken from accurate item-encoding neurons only (see METHODS).

To further characterize the influence of broad inhibition on the relationship between capacity and coding fidelity, and to determine the robustness of our findings to variation in the relative strengths of local and broad inhibition, we ran simulations in which the strength of broad inhibition was half that of the "default" configuration ($\zeta_{rec\,|\,pi,ii} = 1/6$), again preserving the total strength of inhibitory conductance in the model. Capacity was between that of the default configuration and the configuration with local inhibition only, and coding fidelity roughly tracked $K(n)$ by all three measures (more so in higher-gain conditions; Fig. 6*C*). For completeness, we ran simulations without the structured component of feedback inhibition ($\zeta_{rec\,|\,pi,ii} = 1$, broad inhibition only). Unsurprisingly, this approach led to a drastic reduction in capacity, but coding fidelity again tracked $K(n)$ by all three measures (Fig. 6*A*). Indeed, in the higher-gain condition, $K(n)$ and coding fidelity were perfectly in step. Thus the relationship between capacity and coding fidelity was highly robust to the magnitude of broad inhibition and its relative contribution to total feedback inhibition. Also of note, the total pyramidal activity in the network was load dependent in all configurations that included local inhibition. This result contrasts with that of Wei et al. (2012), in which total pyramidal spiking was roughly fixed in a model with broad inhibition only, similar to the model with broad inhibition only here (Fig. 6*A*, second row). Thus our model (with local and broad inhibition) does not imply that total pyramidal spiking is a resource to be shared by memoranda during WM tasks. Rather, our results are consistent with the more general hypothesis that neural tissue is a resource to be shared. The respective roles played by local and broad inhibition in the model are apparent in Fig. 6: broad inhibition imposes capacity and the load dependence of coding fidelity, whereas local feedback inhibition stabilizes item-encoding populations, while limiting interactions between them.

Finally, because the target stimuli were equidistant in our simulations (per Heyselaar et al. 2011; Oemisch et al. 2016) but were nonequidistant in some of the experiments providing the data to which we compare model performance (e.g., Bays et al. 2009; Luck and Vogel 1997; Zhang and Luck 2008), we ran additional simulations to test the robustness of our results to nonequidistant target stimuli. In these simulations (with the default configuration) we placed the stimuli adjacent to one another on the 2-item, 3-item, and 4-item memory tasks, using the spacing of the 5-item task (72° apart; Fig. 7). This change had no qualitative effect on our results, i.e., the relationship between capacity and coding fidelity was retained, as was that between capacity and total pyramidal neuron activity (Fig. 6*B*).

The Mechanism by Which Broad Inhibition Underlies Resource-Like Coding

The mechanism underlying resource-like coding is that a larger number of active item-encoding populations drives more broad inhibition, which reduces the (absolute) mean net current onto item-encoding neurons. This reduction in current decreases SNR for the simple reason that it decreases stimulus-selective spike rates (Fig. 4) but pretrial rates are fixed (constant of average). Indeed, this finding would have been the case with earlier biophysical models of WM capacity (Edin et al. 2009; Wei et al. 2012) and precision (Almeida et al. 2015; Roggeman et al. 2014; Wei et al. 2012) if they had measured it, since these models included broad inhibition. However, the reduction in current increases CV because the SD of net current increases relative to the mean (Fig. 8*A*). In other words, CV (the coefficient of variation of ISIs for accurate item-encoding neurons) reflects the coefficient of variation of net current onto these neurons (Pearson correlation coefficient $r > 0.93$ for all gain conditions). This increase in the relative variability of net current within each trial entails an increase in the relative variability of total net current across trials ($r > 0.88$ for all gain conditions), manifest as an increase in the relative variability of spike counts and therefore FF. In other words, FF (the Fano factor of spike counts) reflects the Fano factor of across-trial total net current (Fig. 8*B*). Of course, this explanation of FF assumes a tight correspondence between the total net current

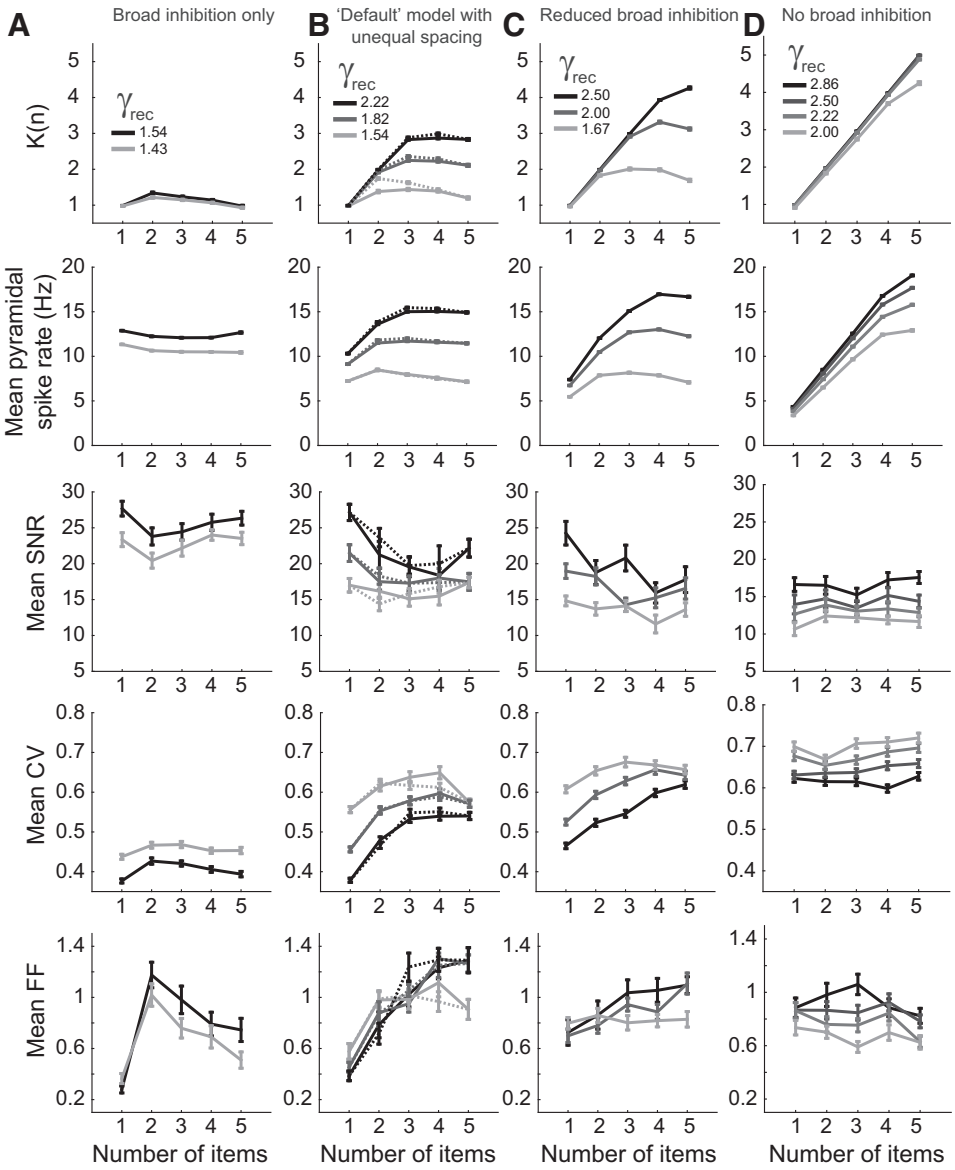

Fig. 6. Capacity $K(n)$ (*top* row), total pyramidal neuron activity (2nd row), and coding fidelity (*bottom* 3 rows) over n under different configurations of local and broad inhibition (*A*, *C*, and *D*) and with unequal spacing of the target stimuli (*B*). Error bars show SE. *A*: results with broad inhibition only ($\zeta_{rec\,|\,pi,ii} = 1$). *B*: results with unequal spacing of the target stimuli for the "default" model, where targets were placed adjacent to one another on the 2-, 3-, and 4-item memory tasks under the spacing of the 5-item task (72° apart). Dotted curves show results with equal spacing (reproduced from Fig. 5), as do results for $n = 1$ and $n = 5$. Results are shown for the highest-, median-, and lowest-gain conditions, where darker shades correspond to higher gain. Other gain conditions are omitted for clarity. *C*: results with reduced broad inhibition ($\zeta_{rec\,|\,pi,ii} = 1/6$), shown for the highest-, median-, and lowest-gain conditions, where darker shades correspond to higher gain. *D*: results with local inhibition only ($\zeta_{rec\,|\,pi,ii} = 0$).

during the statistics window and the spike count, which is indeed the case ($r > 0.97$ for all gain conditions).

DISCUSSION

Our local-circuit PPC model provides an integrated, mechanistic explanation for slot-like capacity and resource-like coding. Both are consequences of broad inhibition, which limits capacity by imposing competition (Edin et al. 2009) and reduces coding fidelity by lowering spike rates and rendering neurons more sensitive to current fluctuations. The model makes testable predictions for electrophysiological studies of WM. Most prominently, it predicts that on multiple-item tasks the SNR (CV and FF) of persistent activity will decrease (increase) with increasing n until capacity is reached, leveling off thereafter. These predictions (Fig. 5, *C–E*) are consistent with behavioral data showing that WM precision decreases with increasing n until capacity is reached, plateauing thereafter (Zhang and Luck 2008) (bilinear precision; but see below for discussion). They also explicitly demonstrate the incompatibility between Resource and mutual inhibition. The latter dictates that coding fidelity cannot decrease indefinitely with increasing memory load. Rather, it must be limited by competition. In this regard, we do not suggest that competitive dynamics in WM circuitry are immutable, dictating rigid capacity limitations. Far from it, we consider context-dependent control of neural dynamics to be fundamental to cognition (see Standage et al. 2014), a view supported by recent studies in relation to WM storage (Almeida et al. 2015; Edin et al. 2009; Roggeman et al. 2014). Thus we expect that capacity and precision will fluctuate with task demands but that their inherent relationship will hold: imprecision must be limited by capacity. Our findings offer a neural mechanism for this relationship.

Quantitative Considerations of Coding Fidelity

We have focused on the qualitative effect of memory load on coding fidelity, i.e., the direction of change in SNR, CV, and FF as a function of n, but quantitative considerations warrant

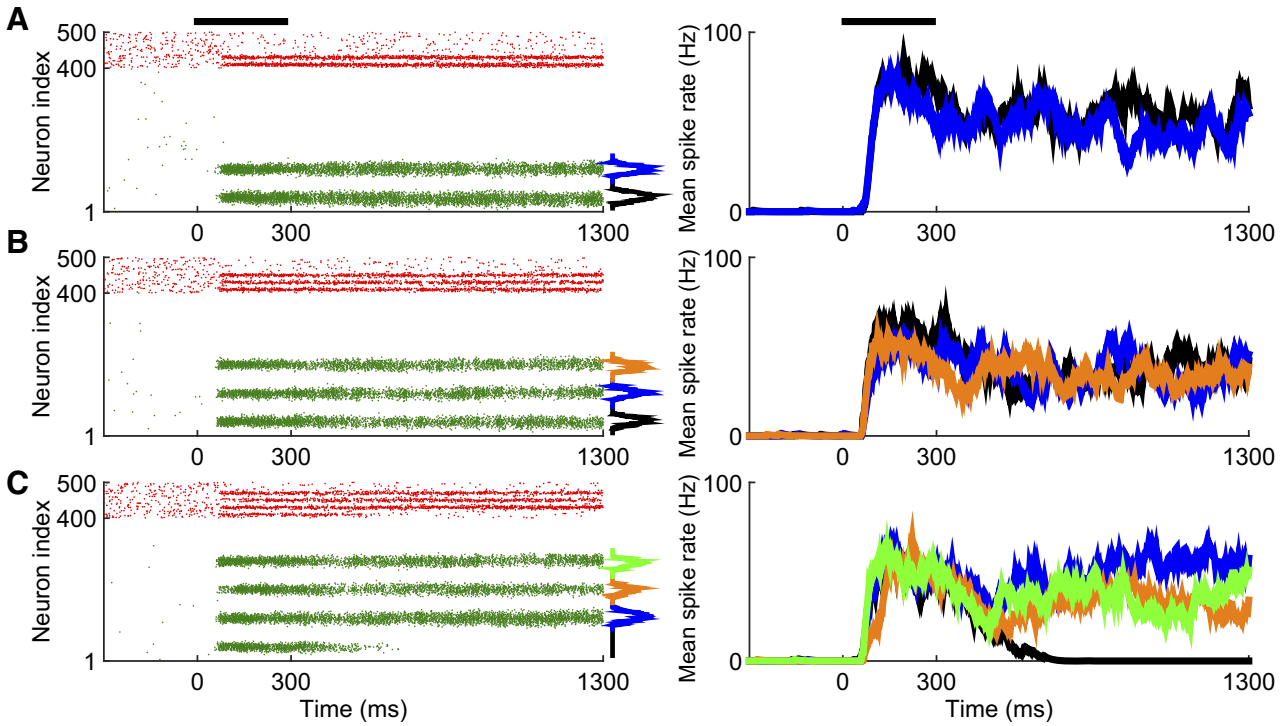

Fig. 7. Three trials of the multiple-item memory task for the highest-gain condition, with 2 (*A*), 3 (*B*), and 4 (*C*) items, placed adjacent to one another under the spacing of the 5-item task. Mean rates over all pyramidal neurons during the statistics window are inset on *right* of raster plots (see Fig. 1*C*), where shades match the mean spike density functions in *right* panels. Thick horizontal bars at *top* show the timing of the stimuli.

further comment. In particular, coding fidelity in our model was somewhat high with low n according to all three measures (Fig. 5, *C* and *D*). SNR is explained by low pretrial rates (mean < 1 Hz for all γ_g), because of the high ratio of inhibitory to excitatory conductance in our method of background current injection. Our parameter values were determined by in vivo cortical data (see *Parameter Values*) and the low pretrial rates they engender are consistent with neurons in the output layers of monkey primary visual cortex, in which spontaneous activity has been thoroughly investigated (Gur et al. 2005; Gur and Snodderly 2008; Snodderly and Gur 1995; see also Maier et al. 2010). In rodents, spontaneous activity is known to depend on intrinsic neuronal properties and connectivity and differs between cortical layers (see Harris and Mrsic-Flogel 2013). Unfortunately, there is a lack of such data from extrastriate and association areas, an issue that should be addressed by future neurophysiological studies. Our focus on coding fidelity concerns task-related spiking, and we do not further pursue background activity here. Suffice it to say, higher background rates would lower SNR in the model. As for CV and FF, it has long been maintained that their values should be around 1 in vivo, per the assumption that cortical spiking is Poissonian (see, e.g., Shadlen and Newsome 1998). More recent data and analyses cast doubt on this assumption, as these measures are sensitive to the finite time windows from which they are calculated, nonstationary spike rates, and serial correlations in spike timing (see Farkhooi et al. 2009; Nawrot 2010; Nawrot et al. 2008; Rajdl and Lansky 2014). Of particular note, failure to account for within-trial fluctuations in spike rate (e.g., due to sensory stimuli) can lead to overestimates of CV (Maimon and Assad 2009; Nawrot 2010), and between-trial fluctuations in rate (e.g., due to attentional state) or an insufficient number of spikes (less than ~5–10) can lead to overestimates of FF (Nawrot et al. 2008). It is worth noting that CV has been reported to be as low as ~0.5 in PPC (Maimon and Assad 2009), whereas FF has been reported in the range of ~0.3–0.4 in visual cortex (Gur et al. 1997; Kara et al. 2000). These are important issues for the understanding of non-task-related activity and neural coding, but none of them impacts our explanation of WM capacity and precision or its qualitative predictions.

WM capacity in our model also warrants further comment. As noted above, max[$K(n)$] was 2 or 3 items in moderate- to high-gain conditions (Fig. 5*A*), consistent with data from nonhuman primate (NHP) and human subjects (Heyselaar et al. 2011; Luck and Vogel 2013). Human studies have reported WM capacity higher than 3 items, though (see Vogel and Awh 2008). Increasing the strength and decreasing the width of recurrent excitation readily increases capacity in our model, but these modifications do not change the finding that coding fidelity tracks capacity by all three measures used (not shown). Thus the specific value of $K(n)$ in each gain condition is parameter dependent, but our chosen parameter values are consistent with the majority of experimental data on capacity (see Luck and Vogel 2013). These values are justified in *Parameter Values*.

Behavioral Data Accounted For by Slot and Resource

To a great degree, the conclusions drawn about WM storage limitations from experimental data depend on the nature of WM tasks and the ways in which performance is measured. For our purposes, WM tasks can be divided into two classes, referred to below as categorical and continuous report tasks. In both classes, information provided in a stimulus array must be

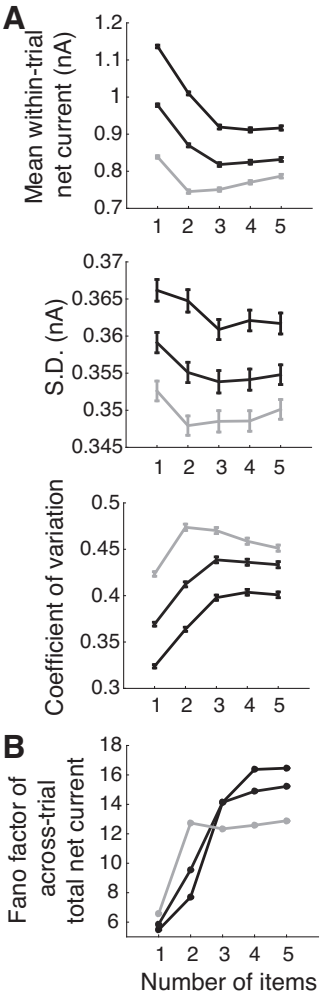

Fig. 8. Inhibition raises the coefficient of variation (CV) of interspike intervals and Fano factor (FF) of spike counts by increasing the relative variability of net current. Within each trial, the standard deviation of net current increases relative to the suprathreshold mean (*A*, rheobase = 0.5 nA for all *n* and gain conditions), so the higher coefficient of variation of net current deregularizes spike timing to a greater degree. This increase in the variability of within-trial net current entails an increase in the variability of total net current on each trial (Pearson's $r > 0.88$ for all gain conditions), resulting in an increase in the Fano factor (variance divided by the mean) of across-trial total net current (*B*). Given the monotonic relationship between the total net current and the spike count ($r > 0.97$ for all gain conditions), the FF (of spike counts) reflects the Fano factor of total net current. *A*: mean (*top*), standard deviation (*middle*), and coefficient of variation (*bottom*) of within trial net current at the RF center of an item-encoding population during the statistics window (highest, median, and lowest gain conditions; see legend). *B*: Fano factor (variance divided by the mean) of the across-trial total net current for each *n*.

retained over a delay interval. On categorical report tasks, performance is measured according to an all-or-none report on that information, such as whether the value of a particular feature is unchanged in a subsequent, postdelay stimulus array (e.g., Luck and Vogel 1997). On continuous report tasks, subjects report the memory of an analog feature value (e.g., Wilken and Ma 2004).

Slot accounts for behavioral data showing bilinear capacity on categorical report tasks (Luck and Vogel 1997), i.e., for capacity *K*, subjects retain *n* items for $n \leq K$ and they retain K/n items for $n > K$. Slot also accounts for EEG (Vogel and Machizawa 2004) and functional magnetic resonance imaging (fMRI) (Linden et al. 2003; Todd and Marois 2004) data showing bilinear signal amplitude, where these signals correlate with $K(n)$ (see below). On continuous report tasks, Slot cannot account for data showing decreasing precision with increasing *n* without recourse to the resource framework. Recent work has therefore referred to "discrete resource" and "continuous resource" hypotheses (see Fukuda et al. 2010). The underlying premise of the former is that slots are a kind of resource that are used in a quantized manner, i.e., discrete subunits of slots can be allocated flexibly to memoranda (e.g., Zhang and Luck 2008). This hybrid approach accounts for bilinear precision, since, according to this hypothesis, the subunits would be spread more thinly with increasing *n*, but all would be occupied for $n > K$.

Conversely, Resource cannot account for bilinear capacity on categorical report tasks but accounts for monotonically decreasing precision with increasing *n* (monotonic precision) on continuous report tasks (Bays et al. 2009; Schneegans and Bays 2016). In its original form (described above) Resource cannot account for bilinear precision, but it can do so with the addition of trial-to-trial noise in the amount of resource allocated to each item (van den Berg et al. 2012). Such trial-to-trial variability has long been employed by hypotheses on cognition [e.g., perceptual decision making (Brown and Heathcote 2005; Carpenter and Williams 1995)] and does not deviate in principle from the original Resource formulation. Thus we do not consider this approach to be a hybrid one. However, the bilinear signal amplitude shown by EEG (Vogel and Machizawa 2004) and fMRI (Todd and Marois 2004) studies on categorical report tasks poses a challenge to the resource framework, which can account for these data if the relevant resource can be continuously and partially allocated to memoranda (see Fukuda et al. 2010).

Neural Models of WM Storage Limitations

Abstract models instantiating Slot and Resource have been invaluable in characterizing WM storage limitations (van den Berg et al. 2012; Zhang and Luck 2008), but they do not speak to the neural mechanisms that may implement their principles. In particular, these models do not account for persistent activity, widely believed to instantiate WM storage (for discussion, see Christophel et al. 2017; Curtis and Lee 2010; D'Esposito and Postle 2015; Riley and Constantinidis 2016). A number of studies have used implementation-level models to address the neural basis of capacity (Edin et al. 2009; Lisman and Idiart 1995; Macoveanu et al. 2006; Raffone and Wolters 2001; Rolls et al. 2013; Tanaka 2002; Wei et al. 2012). These models can be divided into two classes, attributing capacity to fundamentally different mechanisms. In one class, WM items are stored in oscillatory subcycles (e.g., beta-gamma oscillations nested inside alpha-theta oscillations), where different phases effectively isolate memoranda from one another (Lisman and Idiart 1995; Raffone and Wolters 2001). As such, capacity is limited by the ratio of high-frequency to low-frequency oscillations. This compelling possibility relates feature binding more broadly to WM (Raffone and Wolters 2001), accounting for the finding that capacity does not depend on the complexity of WM items (Awh et al. 2007; Luck and Vogel 1997; although see Alvarez and Cavanagh 2004; Brady et al. 2011). Notably, phase separation is fully consistent with the notion of discrete

slots and further accounts for data showing that different phases of gamma oscillations contain information about different memoranda (Axmacher et al. 2010; Leszczyński et al. 2015; Siegel et al. 2009). It is unclear, however, whether simultaneously presented items could be allocated different phases. If not, these models imply that different WM mechanisms may encode simultaneously presented and sequentially presented memoranda.

In the other class of implementation-level model, multiple WM items are stored by attractor states (Almeida et al. 2015; Edin et al. 2009; Macoveanu et al. 2006; Roggeman et al. 2014; Rolls et al. 2013; Tanaka 2002; Wei et al. 2012), i.e., the balance between recurrent excitation and feedback inhibition allows a limited number of memoranda to coexist over a delay interval. A major difference between models of this class is the structure of local-circuit connectivity, where different connectivity structures embody different assumptions about the circuitry being simulated. In relation to capacity, the upshot of these studies is that feedback inhibition necessarily limits capacity (see Edin et al. 2009 for analysis) but the degree to which it does so can be ameliorated by mechanisms that localize and strengthen recurrent excitation.

Several of these studies also considered the neural basis of WM precision, equating imprecision with the SD of trial-to-trial differences between the locations of item-encoding populations and their target locations (Almeida et al. 2015; Roggeman et al. 2014; Wei et al. 2012). In our model, SD was negligible for all n and gain conditions ($1.96° < SD < 2.93°$), calculated over all "accurately" encoded items, where accuracy was recalculated with $C = 360/n/2$. This recalculation provided maximum tolerance for angular deviations, allowing up to 180° with 1 item, 90° with 2 items, and so on. Thus our model predicts that drift has little bearing on WM precision under the task conditions simulated here. At first glance, this finding appears to differ from those of earlier studies, but closer inspection mitigates this difference. In the study by Almeida et al. (2015) a guessing strategy was simulated (drawing from a uniform distribution of feature values) for item-encoding populations that faded out before 500 ms, so we do not know how much drift occurred in their model. In the study by Wei et al. (2012), SD was ~2–3° with equidistant stimuli as n increased from 1 to maximum capacity ($\max[K(n)]$ here), regardless of the model's connectivity structure or the length of their delay interval (see their Figs. 2, B and D, 3C, and 4B). For example, with their "wide" connectivity, SD was ~2° as n increased from 1 to 4 items over a 1-s delay (their Fig. 2, B and D). Thus their model predicts that if drift underlies WM precision then precision will not decrease with increasing subcapacity memory load and equidistant targets. Our model (like Resource) predicts that if coding fidelity underlies WM precision then precision will decrease with increasing subcapacity memory load. To the best of our knowledge, there are no extant data to confirm or refute these predictions, since earlier studies showing load-dependent WM precision used nonequidistant targets (Bays et al. 2009; Schneegans and Bays 2016; Zhang and Luck 2008). This discrepancy makes for good science: different models offer different mechanistic explanations for the same data but make a different, testable prediction for a future experiment. Running this experiment would provide important evidence for one hypothesis over the other, but a definitive test requires neural data. Drift can be tested by constructing tuning curves from electrophysiological recordings of persistent activity on multiple-item WM tasks. To the best of our knowledge, no studies have done so, but we are aware of one study to quantify drift in this way from single-item WM data (Wimmer et al. 2014). These authors showed no appreciable drift in the average tuning bias of prefrontal cortical neurons before ~2 s (their Fig. 3c). Thus these data are in agreement with our findings over the timescale considered here. It should be noted that our results were qualitatively unchanged with a 2-s memory delay (not shown) and drift remained severely limited ($2.75° < SD < 4.18°$). Ultimately, WM storage limitations may reflect constraints on the encoding, maintenance, and/or decoding of memoranda (Ma et al. 2014). Our results emphasize the discriminability, regularity, and reliability of persistent spiking as sources of WM imprecision, thereby implicating maintenance and decoding, but we do not suggest that these factors are the only sources of imprecision. Crucially, our study uses established measures of coding fidelity for single-neuron data (SNR, CV, and FF), so our predictions for coding fidelity on multiple-item tasks are testable with single-neuron recordings.

Limitations of the Model

Of course, our model has limitations. To begin with, it only considers the spatial location of memoranda, ignoring other features and their conjunctions. In effect, our simulations assume that everything encoded by PPC satisfies a set of rules for selection. This approach is common among attractor models (e.g., Almeida et al. 2015; Edin et al. 2009; Macoveanu et al. 2006; Tanaka 2002; Wei et al. 2012) and is justified in studies focused on storage limitations, i.e., it focuses on the mutual influence of persistently active neural populations, regardless of the features or rules that lead to their initial activation. As noted above, models in which memoranda are stored in oscillatory subcycles can account for feature binding with sequentially presented stimuli (Lisman and Idiart 1995; Raffone and Wolters 2001). A more general understanding of feature-bound memoranda will likely require hierarchical models with converging feature maps. Such models have the further potential to explain the neural basis of "swap errors," where subjects report the value of an item from the memory array other than the one probed (see Bays 2016). Hierarchical models also have the potential to explain the flexibility of WM precision, that is, the finding that one item can be maintained with higher precision than the others, but at a cost to those other items (see Ma et al. 2014). This finding points to the relationship between WM and attention and to the modulation of item-encoding populations in distributed circuitry. These and related phenomena are beyond the scope of the present study, the focus of which is the mechanistic relationship between capacity and precision.

Like earlier neural models of multiple-item WM (e.g., Wei et al. 2012), our model does not account for monotonic precision (Bays et al. 2009; Schneegans and Bays 2016) in and of itself. It is important to note, however, that the model simulates WM storage in PPC and does not speak to the strategies that subjects may use when reporting forgotten items. Spiking activity in the model predicts that coding fidelity tracks capacity, which is roughly bilinear in higher-gain conditions (Fig. 5). As such, the model explains decreasing precision with increas-

ing memory load for subcapacity loads (the "first line" of the bilinearity) and it explains the plateau in precision (the second line) when retained items are reported. For forgotten items, it accounts for the plateau under the assumption that guesses are drawn from a uniform distribution of feature values, an assumption made by Slot (Zhang and Luck 2008) and by previous neural modeling studies (e.g., Almeida et al. 2015). Along a similar vein, our model (and other models with capacity limitations) may account for monotonic precision under the assumption that information about retained items shapes the distribution from which subjects draw when guessing. For example, if a subject has a capacity of 3 or 4 items, then the proportion of known feature values relative to memory load is smaller on an 8-item task than on a 6-item task. As such, our model (and other models with capacity limitations) can account for monotonic precision if the greater proportion of unknown values causes a broader distribution of guesses. This possibility overlaps with "chunking" strategies, or the enhancement of WM storage by the grouping of items along a feature dimension (see Brady et al. 2011; Cowan 2001; Miller 1956). Experiments in which feature values are assigned nonequidistantly [e.g., randomly (Bays et al. 2009; Zhang and Luck 2008)] to target stimuli are more susceptible to chunking because the values of stimuli are more readily grouped in a given feature space, but the effect of this grouping on guessing strategies is unknown. There is strong evidence that subjects improve their WM performance by leveraging the statistical structure of stimulus arrays (Brady et al. 2009; Brady and Alvarez 2015; Lew and Vul 2015; Orhan and Jacobs 2013), and it seems unlikely that they would disregard this information when guessing. For example, on a continuous report task for color (Bays et al. 2009; Zhang and Luck 2008), if three of six target stimuli were red (or reddish) on a given trial, then this information not only would be useful when reporting the value of the three red (chunked) items but would also be useful when reporting the value of the other three items (they were not red). The systematic investigation of guessing is a recent addition to the literature on WM storage (Adam et al. 2017), and we look forward to further studies characterizing this important factor in the interpretation of WM performance (e.g., Nassar et al. 2018).

Finally, it is worth noting that our study is part of an ongoing research program investigating the neural basis of WM storage limitations with NHP subjects, in which in vivo electrophysiological recordings can be made during WM tasks. Thus our simulations are purposefully constrained by our experiments, e.g., our use of equidistant targets (see Heyselaar et al. 2011 for justification). This approach facilitates the testing of our predictions for single-neuron activity. In recent years, several studies have used categorical report tasks with NHP subjects (e.g., Elmore et al. 2011; Heyselaar et al. 2011; Lara and Wallis 2012), but we are unaware of any studies using continuous report tasks with a nonhuman species. Future work must address this challenging gap.

Conclusions

Although recent studies have investigated neural mechanisms for WM capacity and precision (Okimura et al. 2015; Roggeman et al. 2014; Wei et al. 2012), to the best of our knowledge no previous study has proposed a neural mechanism for precision under the principles of Resource that also accounts for persistent activity. Studies have proposed that Resource is implemented by the gain of item-encoding populations (van den Berg and Ma 2014) and by divisively normalized probabilistic spiking (Bays 2014), but these studies have taken persistent activity for granted, i.e., they did not address the mechanisms by which an unlimited number of low-gain or divisively normalized neural populations would be sustained over a memory delay. As described above, nested oscillations and attractor dynamics constrain capacity, so other mechanisms would be required. Our simulations without broad inhibition (Fig. 6*D*) are instructive in this regard, since limiting competition rendered coding fidelity independent of memory load. In other words, the very mechanism that might alleviate Resource from strong capacity constraints rendered coding fidelity less resource-like. This conundrum points to the need for continuous report tasks with set sizes that significantly exceed estimates of capacity on categorical report tasks. Such large set sizes can have significant effects on capacity [e.g., 12 items (Matsuyoshi et al. 2014)] and may provide insight into its mechanistic relationship with precision. This relationship has received much less attention than each of its elements. In this regard, tremendous insight into WM capacity and precision has been gleaned from studies focusing on their differences, guided by the principles of Slot and Resource, respectively. Our study suggests that their commonalities may be just as informative.

GRANTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/06686-2015).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

D.S. and M.P. conceived and designed research; D.S. performed experiments; D.S. analyzed data; D.S. and M.P. interpreted results of experiments; D.S. prepared figures; D.S. drafted manuscript; D.S. and M.P. edited and revised manuscript; D.S. and M.P. approved final version of manuscript.

REFERENCES

Adam KC, Vogel EK, Awh E. Clear evidence for item limits in visual working memory. *Cognit Psychol* 97: 79–97, 2017. doi:10.1016/j.cogpsych.2017.07.001.

Almeida R, Barbosa J, Compte A. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *J Neurophysiol* 114: 1806–1818, 2015. doi:10.1152/jn.00362.2015.

Alvarez GA, Cavanagh P. The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychol Sci* 15: 106–111, 2004. doi:10.1111/j.0963-7214.2004.01502006.x.

Angulo MC, Rossier J, Audinat E. Postsynaptic glutamate receptors and integrative properties of fast-spiking interneurons in the rat neocortex. *J Neurophysiol* 82: 1295–1302, 1999. doi:10.1152/jn.1999.82.3.1295.

Awh E, Barton B, Vogel EK. Visual working memory represents a fixed number of items regardless of complexity. *Psychol Sci* 18: 622–628, 2007. doi:10.1111/j.1467-9280.2007.01949.x.

Axmacher N, Henseler MM, Jensen O, Weinreich I, Elger CE, Fell J. Cross-frequency coupling supports multi-item working memory in the human hippocampus. *Proc Natl Acad Sci USA* 107: 3228–3233, 2010. doi:10.1073/pnas.0911531107.

Bays PM. Noise in neural populations accounts for errors in working memory. *J Neurosci* 34: 3632–3645, 2014. doi:10.1523/JNEUROSCI.3204-13.2014.

Bays PM. Evaluating and excluding swap errors in analogue tests of working memory. *Sci Rep* 6: 19203, 2016. doi:10.1038/srep19203.

Bays PM, Catalao RF, Husain M. The precision of visual working memory is set by allocation of a shared resource. *J Vis* 9: 7.1–7.11, 2009. doi:10.1167/9.10.7.

Berger TK, Perin R, Silberberg G, Markram H. Frequency-dependent disynaptic inhibition in the pyramidal network: a ubiquitous pathway in the developing rat neocortex. *J Physiol* 587: 5411–5425, 2009. doi:10.1113/jphysiol.2009.176552.

Berretta N, Jones RS. A comparison of spontaneous EPSCs in layer II and layer IV-V neurons of the rat entorhinal cortex in vitro. *J Neurophysiol* 76: 1089–1100, 1996. doi:10.1152/jn.1996.76.2.1089.

Brady TF, Alvarez GA. Contextual effects in visual working memory reveal hierarchically structured memory representations. *J Vis* 15: 6, 2015. doi:10.1167/15.15.6.

Brady TF, Konkle T, Alvarez GA. Compression in visual working memory: using statistical regularities to form more efficient memory representations. *J Exp Psychol Gen* 138: 487–502, 2009. doi:10.1037/a0016797.

Brady TF, Konkle T, Alvarez GA. A review of visual memory capacity: beyond individual items and toward structured representations. *J Vis* 11: 4, 2011. doi:10.1167/11.5.4.

Brown S, Heathcote A. A ballistic model of choice response time. *Psychol Rev* 112: 117–128, 2005. doi:10.1037/0033-295X.112.1.117.

Carpenter RH, Williams ML. Neural computation of log likelihood in control of saccadic eye movements. *Nature* 377: 59–62, 1995. doi:10.1038/377059a0.

Chance FS, Abbott LF, Reyes AD. Gain modulation from background synaptic input. *Neuron* 35: 773–782, 2002. doi:10.1016/S0896-6273(02)00820-6.

Chee MW, Chuah YM. Functional neuroimaging and behavioral correlates of capacity decline in visual short-term memory after sleep deprivation. *Proc Natl Acad Sci USA* 104: 9487–9492, 2007. doi:10.1073/pnas.0610712104.

Christophel TB, Hebart MN, Haynes JD. Decoding the contents of visual short-term memory from human visual and parietal cortex. *J Neurosci* 32: 12983–12989, 2012. doi:10.1523/JNEUROSCI.0184-12.2012.

Christophel TB, Klink PC, Spitzer B, Roelfsema PR, Haynes JD. The distributed nature of working memory. *Trends Cogn Sci* 21: 111–124, 2017. doi:10.1016/j.tics.2016.12.007.

Churchland AK, Kiani R, Shadlen MN. Decision-making with multiple alternatives. *Nat Neurosci* 11: 693–702, 2008. [Erratum in *Nat Neurosci* 11: 851, 2008.] doi:10.1038/nn.2123.

Compte A, Brunel N, Goldman-Rakic PS, Wang XJ. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb Cortex* 10: 910–923, 2000. doi:10.1093/cercor/10.9.910.

Cowan N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav Brain Sci* 24: 87–114, 2001. doi:10.1017/S0140525X01003922.

Curtis CE, Lee D. Beyond working memory: the role of persistent activity in decision making. *Trends Cogn Sci* 14: 216–222, 2010. doi:10.1016/j.tics.2010.03.006.

D'Esposito M, Postle BR. The cognitive neuroscience of working memory. *Annu Rev Psychol* 66: 115–142, 2015. doi:10.1146/annurev-psych-010814-015031.

Desai NS, Cudmore RH, Nelson SB, Turrigiano GG. Critical periods for experience-dependent synaptic scaling in visual cortex. *Nat Neurosci* 5: 783–789, 2002. doi:10.1038/nn878.

Destexhe A. Inhibitory "noise". *Front Cell Neurosci* 4: 9, 2010. doi:10.3389/fncel.2010.00009.

Destexhe A, Rudolph M, Fellous JM, Sejnowski TJ. Fluctuating synaptic conductances recreate in vivo-like activity in neocortical neurons. *Neuroscience* 107: 13–24, 2001. doi:10.1016/S0306-4522(01)00344-X.

Douglas R, Markram H, Martin K. Neocortex. In: *The Synaptic Organization of the Brain*, edited by Shepherd GM. Oxford, UK: Oxford Univ. Press, 2004, p. 499–558. doi:10.1093/acprof:oso/9780195159561.003.0012.

Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, Compte A. Mechanism for top-down control of working memory capacity. *Proc Natl Acad Sci USA* 106: 6802–6807, 2009. doi:10.1073/pnas.0901894106.

Elmore LC, Ma WJ, Magnotti JF, Leising KJ, Passaro AD, Katz JS, Wright AA. Visual short-term memory compared in rhesus monkeys and humans. *Curr Biol* 21: 975–979, 2011. doi:10.1016/j.cub.2011.04.031.

Farkhooi F, Strube-Bloss MF, Nawrot MP. Serial correlation in neural spike trains: experimental evidence, stochastic modeling, and single neuron variability. *Phys Rev E Stat Nonlin Soft Matter Phys* 79: 021905, 2009. doi:10.1103/PhysRevE.79.021905.

Fellous JM, Rudolph M, Destexhe A, Sejnowski TJ. Synaptic background noise controls the input/output characteristics of single cells in an in vitro model of in vivo activity. *Neuroscience* 122: 811–829, 2003. doi:10.1016/j.neuroscience.2003.08.027.

Franconeri SL, Alvarez GA, Cavanagh P. Flexible cognitive resources: competitive content maps for attention and memory. *Trends Cogn Sci* 17: 134–141, 2013. doi:10.1016/j.tics.2013.01.010.

Fukuda K, Awh E, Vogel EK. Discrete capacity limits in visual working memory. *Curr Opin Neurobiol* 20: 177–182, 2010. doi:10.1016/j.conb.2010.03.005.

Fukuda K, Woodman GF, Vogel EK. Individual differences in visual working memory capacity: contributions of attentional control to storage. In: *Mechanisms of Sensory Working Memory: Attention and Performance XXV*, edited by Jolicoeur P, Lefebvre C, Martinez-Trujillo J. London, UK: Elsevier, 2015, p. 105–119. doi:10.1016/B978-0-12-801371-7.00009-0.

Funahashi S, Bruce CJ, Goldman-Rakic PS. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J Neurophysiol* 61: 331–349, 1989. doi:10.1152/jn.1989.61.2.331.

Gnadt JW, Andersen RA. Memory related motor planning activity in posterior parietal cortex of macaque. *Exp Brain Res* 70: 216–220, 1988.

Gur M, Beylin A, Snodderly DM. Response variability of neurons in primary visual cortex (V1) of alert monkeys. *J Neurosci* 17: 2914–2920, 1997. doi:10.1523/JNEUROSCI.17-08-02914.1997.

Gur M, Kagan I, Snodderly DM. Orientation and direction selectivity of neurons in V1 of alert monkeys: functional relationships and laminar distributions. *Cereb Cortex* 15: 1207–1221, 2005. doi:10.1093/cercor/bhi003.

Gur M, Snodderly DM. Physiological differences between neurons in layer 2 and layer 3 of primary visual cortex (V1) of alert macaque monkeys. *J Physiol* 586: 2293–2306, 2008. doi:10.1113/jphysiol.2008.151795.

Harris KD, Mrsic-Flogel TD. Cortical connectivity and sensory coding. *Nature* 503: 51–58, 2013. doi:10.1038/nature12654.

Hellwig B. A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biol Cybern* 82: 111–121, 2000. doi:10.1007/PL00007964.

Hestrin S. Different glutamate receptor channels mediate fast excitatory synaptic currents in inhibitory and excitatory cortical neurons. *Neuron* 11: 1083–1091, 1993. doi:10.1016/0896-6273(93)90221-C.

Heyselaar E, Johnston K, Paré M. A change detection approach to study visual working memory of the macaque monkey. *J Vis* 11: 11, 2011. doi:10.1167/11.3.11.

Hull C, Isaacson JS, Scanziani M. Postsynaptic mechanisms govern the differential excitation of cortical neurons by thalamic inputs. *J Neurosci* 29: 9127–9136, 2009. doi:10.1523/JNEUROSCI.5971-08.2009.

Jahr CE, Stevens CF. Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics. *J Neurosci* 10: 3178–3182, 1990. doi:10.1523/JNEUROSCI.10-09-03178.1990.

Johnston KD, Brunamonti E, Thomas NW, Pare M. Posterior parietal cortex persistent activity and visuospatial working memory (Abstract). *Neuroscience 2009*. Chicago, IL, 2009, 356.16.

Kara P, Reinagel P, Reid RC. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron* 27: 635–646, 2000. doi:10.1016/S0896-6273(00)00072-6.

Kisvárday ZF, Beaulieu C, Eysel UT. Network of GABAergic large basket cells in cat visual cortex (area 18): implication for lateral disinhibition. *J Comp Neurol* 327: 398–415, 1993. doi:10.1002/cne.903270307.

Krimer LS, Goldman-Rakic PS. Prefrontal microcircuits: membrane properties and excitatory input of local, medium, and wide arbor interneurons. *J Neurosci* 21: 3788–3796, 2001. doi:10.1523/JNEUROSCI.21-11-03788.2001.

Krimer LS, Zaitsev AV, Czanner G, Kröner S, González-Burgos G, Povysheva NV, Iyengar S, Barrionuevo G, Lewis DA. Cluster analysis-based physiological classification and morphological properties of inhibitory neurons in layers 2-3 of monkey dorsolateral prefrontal cortex. *J Neurophysiol* 94: 3009–3022, 2005. doi:10.1152/jn.00156.2005.

Lara AH, Wallis JD. Capacity and precision in an animal model of visual short-term memory. *J Vis* 12: 13, 2012. doi:10.1167/12.3.13.

Leszczyński M, Fell J, Axmacher N. Rhythmic working memory activation in the human hippocampus. *Cell Reports* 13: 1272–1282, 2015. doi:10.1016/j.celrep.2015.09.081.

Lew TF, Vul E. Ensemble clustering in visual working memory biases location memories and reduces the Weber noise of relative positions. *J Vis* 15: 10, 2015. doi:10.1167/15.4.10.

Linden DE, Bittner RA, Muckli L, Waltz JA, Kriegeskorte N, Goebel R, Singer W, Munk MH. Cortical capacity constraints for visual working memory: dissociation of fMRI load effects in a fronto-parietal network. *Neuroimage* 20: 1518–1530, 2003. doi:10.1016/j.neuroimage.2003.07.021.

Linke AC, Vicente-Grabovetsky A, Mitchell DJ, Cusack R. Encoding strategy accounts for individual differences in change detection measures of VSTM. *Neuropsychologia* 49: 1476–1486, 2011. doi:10.1016/j.neuropsychologia.2010.11.034.

Lisman JE, Idiart MA. Storage of 7 +/− 2 short-term memories in oscillatory subcycles. *Science* 267: 1512–1515, 1995. doi:10.1126/science.7878473.

Louie K, Grattan LE, Glimcher PW. Reward value-based gain control: divisive normalization in parietal cortex. *J Neurosci* 31: 10627–10639, 2011. doi:10.1523/JNEUROSCI.1237-11.2011.

Luck SJ, Vogel EK. The capacity of visual working memory for features and conjunctions. *Nature* 390: 279–281, 1997. doi:10.1038/36846.

Luck SJ, Vogel EK. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends Cogn Sci* 17: 391–400, 2013. doi:10.1016/j.tics.2013.06.006.

Ma WJ, Husain M, Bays PM. Changing concepts of working memory. *Nat Neurosci* 17: 347–356, 2014. doi:10.1038/nn.3655.

Macoveanu J, Klingberg T, Tegnér J. A biophysical model of multiple-item working memory: a computational and neuroimaging study. *Neuroscience* 141: 1611–1618, 2006. doi:10.1016/j.neuroscience.2006.04.080.

Maier A, Adams GK, Aura C, Leopold DA. Distinct superficial and deep laminar domains of activity in the visual cortex during rest and stimulation. *Front Syst Neurosci* 4: 1–11, 2010.

Maimon G, Assad JA. Beyond Poisson: increased spike-time regularity across primate parietal cortex. *Neuron* 62: 426–440, 2009. doi:10.1016/j.neuron.2009.03.021.

Markram H, Toledo-Rodriguez M, Wang Y, Gupta A, Silberberg G, Wu C. Interneurons of the neocortical inhibitory system. *Nat Rev Neurosci* 5: 793–807, 2004. doi:10.1038/nrn1519.

Matsuyoshi D, Osaka M, Osaka N. Age and individual differences in visual working memory deficit induced by overload. *Front Psychol* 5: 384, 2014. doi:10.3389/fpsyg.2014.00384.

McBain CJ, Fisahn A. Interneurons unbound. *Nat Rev Neurosci* 2: 11–23, 2001. doi:10.1038/35049047.

Miller GA. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol Rev* 63: 81–97, 1956. doi:10.1037/h0043158.

Nassar MR, Helmers JC, Frank MJ. Chunking as a rational strategy for lossy data compression in visual working memory. *Psychol Rev* 125: 486–511, 2018. doi:10.1037/rev0000101.

Nawrot MP. Analysis and interpretation of interval and count variability in neural spike trains. In: *Analysis of Parallel Spike Trains*, edited by Grün S, Rotter S. New York: Springer, 2010, p. 37–58. doi:10.1007/978-1-4419-5675-0_3.

Nawrot MP, Boucsein C, Rodriguez Molina V, Riehle A, Aertsen A, Rotter S. Measurement of variability dynamics in cortical spike trains. *J Neurosci Methods* 169: 374–390, 2008. doi:10.1016/j.jneumeth.2007.10.013.

Oemisch M, Johnston K, Paré M. Methylphenidate does not enhance visual working memory but benefits motivation in macaque monkeys. *Neuropharmacology* 109: 223–235, 2016. doi:10.1016/j.neuropharm.2016.06.019.

Okimura T, Tanaka S, Maeda T, Kato M, Mimura M. Simulation of the capacity and precision of working memory in the hypodopaminergic state: relevance to schizophrenia. *Neuroscience* 295: 80–89, 2015. doi:10.1016/j.neuroscience.2015.03.039.

Orhan AE, Jacobs RA. A probabilistic clustering theory of the organization of visual short-term memory. *Psychol Rev* 120: 297–328, 2013. doi:10.1037/a0031541.

Packer AM, Yuste R. Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: a canonical microcircuit for inhibition? *J Neurosci* 31: 13260–13271, 2011. doi:10.1523/JNEUROSCI.3131-11.2011.

Palva JM, Monto S, Kulashekhar S, Palva S. Neuronal synchrony reveals working memory networks and predicts individual memory capacity. *Proc Natl Acad Sci USA* 107: 7580–7585, 2010. doi:10.1073/pnas.0913113107.

Paré M, Wurtz RH. Monkey posterior parietal cortex neurons antidromically activated from superior colliculus. *J Neurophysiol* 78: 3493–3497, 1997. doi:10.1152/jn.1997.78.6.3493.

Povysheva NV, González-Burgos G, Zaitsev AV, Kröner S, Barrionuevo G, Lewis DA, Krimer LS. Properties of excitatory synaptic responses in fast-spiking interneurons and pyramidal cells from monkey and rat prefrontal cortex. *Cereb Cortex* 16: 541–552, 2006. doi:10.1093/cercor/bhj002.

Prescott SA, De Koninck Y. Gain control of firing rate by shunting inhibition: roles of synaptic noise and dendritic saturation. *Proc Natl Acad Sci USA* 100: 2076–2081, 2003. doi:10.1073/pnas.0337591100.

Raffone A, Wolters G. A cortical mechanism for binding in visual working memory. *J Cogn Neurosci* 13: 766–785, 2001. doi:10.1162/08989290152541430.

Rajdl K, Lansky P. Fano factor estimation. *Math Biosci Eng* 11: 105–123, 2014. doi:10.3934/mbe.2014.11.105.

Riley MR, Constantinidis C. Role of prefrontal persistent activity in working memory. *Front Syst Neurosci* 9: 181, 2016. doi:10.3389/fnsys.2015.00181.

Roggeman C, Klingberg T, Feenstra HE, Compte A, Almeida R. Trade-off between capacity and precision in visuospatial working memory. *J Cogn Neurosci* 26: 211–222, 2014. doi:10.1162/jocn_a_00485.

Rolls ET, Dempere-Marco L, Deco G. Holding multiple items in short term memory: a neural mechanism. *PLoS One* 8: e61078, 2013. doi:10.1371/journal.pone.0061078.

Salazar RF, Dotson NM, Bressler SL, Gray CM. Content-specific fronto-parietal synchronization during visual working memory. *Science* 338: 1097–1100, 2012. doi:10.1126/science.1224000.

Salin PA, Prince DA. Spontaneous $GABA_A$ receptor-mediated inhibitory currents in adult rat somatosensory cortex. *J Neurophysiol* 75: 1573–1588, 1996. doi:10.1152/jn.1996.75.4.1573.

Schneegans S, Bays PM. No fixed item limit in visuospatial working memory. *Cortex* 83: 181–193, 2016. doi:10.1016/j.cortex.2016.07.021.

Self MW, Kooijmans RN, Supèr H, Lamme VA, Roelfsema PR. Different glutamate receptors convey feedforward and recurrent processing in macaque V1. *Proc Natl Acad Sci USA* 109: 11031–11036, 2012. doi:10.1073/pnas.1119527109.

Shadlen MN, Newsome WT. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J Neurosci* 18: 3870–3896, 1998. doi:10.1523/JNEUROSCI.18-10-03870.1998.

Siegel M, Warden MR, Miller EK. Phase-dependent neuronal coding of objects in short-term memory. *Proc Natl Acad Sci USA* 106: 21341–21346, 2009. doi:10.1073/pnas.0908193106.

Snodderly DM, Gur M. Organization of striate cortex of alert, trained monkeys (*Macaca fascicularis*): ongoing activity, stimulus selectivity, and widths of receptive field activating regions. *J Neurophysiol* 74: 2100–2125, 1995. doi:10.1152/jn.1995.74.5.2100.

Somers DC, Nelson SB, Sur M. An emergent model of orientation selectivity in cat visual cortical simple cells. *J Neurosci* 15: 5448–5465, 1995. doi:10.1523/JNEUROSCI.15-08-05448.1995.

Standage D, Blohm G, Dorris MC. On the neural implementation of the speed-accuracy trade-off. *Front Neurosci* 8: 236, 2014. doi:10.3389/fnins.2014.00236.

Standage D, Paré M. Persistent storage capability impairs decision making in a biophysical network model. *Neural Netw* 24: 1062–1073, 2011. doi:10.1016/j.neunet.2011.05.004.

Standage D, You H, Wang DH, Dorris MC. Trading speed and accuracy by coding time: a coupled-circuit cortical model. *PLOS Comput Biol* 9: e1003021, 2013. doi:10.1371/journal.pcbi.1003021.

Tanaka S. Multi-directional representation of spatial working memory in a model prefrontal cortical circuit. *Neurocomputing* 44–46: 1001–1008, 2002. doi:10.1016/S0925-2312(02)00503-9.

Thomas NW, Paré M. Temporal processing of saccade targets in parietal cortex area LIP during visual search. *J Neurophysiol* 97: 942–947, 2007. doi:10.1152/jn.00413.2006.

Thompson KG, Hanes DP, Bichot NP, Schall JD. Perceptual and motor processing stages identified in the activity of macaque frontal eye field neurons during visual search. *J Neurophysiol* 76: 4040–4055, 1996. doi:10.1152/jn.1996.76.6.4040.

Todd JJ, Marois R. Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature* 428: 751–754, 2004. doi:10.1038/nature02466.

Troyer TW, Miller KD. Physiological gain leads to high ISI variability in a simple model of a cortical regular spiking cell. *Neural Comput* 9: 971–983, 1997. doi:10.1162/neco.1997.9.5.971.

Tuckwell H. *Introduction to Theoretical Neurobiology*. Cambridge, UK: Cambridge Univ. Press, 1988.

van den Berg R, Ma WJ. "Plateau"-related summary statistics are uninformative for comparing working memory models. *Atten Percept Psychophys* 76: 2117–2135, 2014. doi:10.3758/s13414-013-0618-7.

van den Berg R, Shin H, Chou W-C, George R, Ma WJ. Variability in encoding precision accounts for visual short-term memory limitations. *Proc Natl Acad Sci USA* 109: 8780–8785, 2012. doi:10.1073/pnas.1117465109.

Vogel EK, Awh E. How to exploit diversity for scientific gain: using individual differences to constrain cognitive theory. *Curr Dir Psychol Sci* 17: 171–176, 2008. doi:10.1111/j.1467-8721.2008.00569.x.

Vogel EK, Machizawa MG. Neural activity predicts individual differences in visual working memory capacity. *Nature* 428: 748–751, 2004. doi:10.1038/nature02447.

Voges N, Schüz A, Aertsen A, Rotter S. A modeler's view on the spatial structure of intrinsic horizontal connectivity in the neocortex. *Prog Neurobiol* 92: 277–292, 2010. doi:10.1016/j.pneurobio.2010.05.001.

Wang XJ. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J Neurosci* 19: 9587–9603, 1999. doi:10.1523/JNEUROSCI.19-21-09587.1999.

Wang XJ. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci* 24: 455–463, 2001. doi:10.1016/S0166-2236(00)01868-3.

Wei Z, Wang XJ, Wang DH. From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *J Neurosci* 32: 11228–11240, 2012. doi:10.1523/JNEUROSCI.0735-12.2012.

Wilken P, Ma WJ. A detection theory account of change detection. *J Vis* 4: 1120–1135, 2004. doi:10.1167/4.12.11.

Wilson HR, Cowan JD. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik* 13: 55–80, 1973. doi:10.1007/BF00288786.

Wimmer K, Nykamp DQ, Constantinidis C, Compte A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat Neurosci* 17: 431–439, 2014. doi:10.1038/nn.3645.

Xiang Z, Huguenard JR, Prince DA. $GABA_A$ receptor-mediated currents in interneurons and pyramidal cells of rat visual cortex. *J Physiol* 506: 715–730, 1998. doi:10.1111/j.1469-7793.1998.715bv.x.

Xu Y. The role of the superior intraparietal sulcus in supporting visual short-term memory for multifeature objects. *J Neurosci* 27: 11676–11686, 2007. doi:10.1523/JNEUROSCI.3545-07.2007.

Zhang W, Luck SJ. Discrete fixed-resolution representations in visual working memory. *Nature* 453: 233–235, 2008. doi:10.1038/nature06860.