Supplementary Information
Stochastic sampling provides a unifying account of working memory limits

Sebastian Schneegans, Robert Taylor & Paul M Bays

Department of Psychology, University of Cambridge, Cambridge, UK

1 Behavioral data

To evaluate different models of visual working memory, we compared the quality of model fits for a large set of behavioral data from continuous report tasks. This dataset is compiled of 15 experiments with over 190,000 trials in total. We included available data from published continuous report experiments (either single-report or whole-report tasks) that have the following characteristics: They test recall performance for at least two different set sizes with a fixed delay duration, all items are presented simultaneously and equally likely to be tested, and the reported feature is either color or orientation. The target item can be indicated either by a location cue or a categorical color cue. The dataset for single-report tasks (Table S1) is similar to the dataset used in a previous model comparison (van den Berg et al., 2014), but we excluded experiments from two studies that have in the meantime been retracted, and added several more recent studies. We also fit behavioral data from four whole-report tasks (Table S2), in which participants had to report the feature values of all items presented in the sample array, with the order of responses either freely chosen by the participant, or determined randomly by the experiment software (Adam et al., 2017). In the whole-report tasks, items were always cued or selected via their location.

2 Models

2.1 General assumptions and notations

For a single trial in a continuous report task, we denote the set size of the memory sample array with N, and the feature values of the sample items with $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$. For classical continuous report tasks with a single report, we denote the reported feature value with ψ. For whole-report tasks, the sequence of reported feature value is $\boldsymbol{\psi} = (\psi_1, \dots, \psi_N)$.

Each model defines a response probability distribution $p(\psi|\theta)$ assuming that the response is generated based on a sample item with true feature value θ. For all model fits, we incorporate swap errors, which we found to consistently improve the quality of fits (see Section 4.3). We assume that response distributions around the selected

No	Study	Feature	Set Sizes	Participants	Trials
1	Zhang and Luck (2008)	Color	1, 2, 3, 6	8	125
2	Bays et al. (2009)	Color	1, 2, 4, 6	12	200
3	van den Berg et al. (2012), Exp 1	Color	1 – 8	13	216
4	van den Berg et al. (2012), Exp 2	Orientation	1 – 8	6	320
5	Rademaker et al. (2012)	Orientation	3, 6	6	800
6	Bays (2014), Exp 1	Orientation	1, 2, 4, 8	8	230
7	Bays et al. (2011a), Exp 1	Orientation	1, 2, 4, 6	8	800
8	Bays et al. (2011b)	Orientation	1, 6	10	50, 250
9	Bays et al. (2011b)	Color	1, 6	10	50, 250
10	Gorgoraptis et al. (2011), Exp 2	Orientation	1 – 5	8	100
11	Pratte et al. (2017)	Orientation	1, 2, 3, 6	12	640

Table S1: Single-report experiments used for model comparison in this study. The Trials column denotes the number of trials each participant completed per set size.

No	Feature	Report Order	Set Sizes	Participants	Trials
1	Color	Free	1, 2, 3, 4, 6	22	99
2	Orientation	Free	1, 2, 3, 4, 6	20	200
3	Color	Random	1, 2, 3, 4, 6	17	99
4	Orientation	Random	1, 2, 3, 4, 6	19	200

Table S2: Whole-report experiments used for model comparison. All experiments are taken from Adam et al. (2017).

feature value are identical for target and non-target features, and that each non-target feature value has a fixed probability p_{NT} of being used as the basis for response generation, so the total proportion of swap errors increases linearly with set size (this is of course not tenable for very large N, but suffices for studies in our dataset, $N \leq 8$). For the response ψ_i corresponding to a cued item θ_i, we then obtain the probability distribution

$$p(\psi_i|\boldsymbol{\theta}) = p_T p(\psi_i|\theta_i) + p_{\mathrm{NT}} \sum_{j \in \{1,\dots,N\}, j \neq i} p(\psi_i|\theta_j), \tag{1}$$

where $p_T = 1 - (N-1)p_{\mathrm{NT}}$ is the probability that the response is based on the feature value of the target item.

In the whole-report tasks with freely chosen response order, we assume for all models that responses are ordered by precision (either expressed as the number of samples assigned to them or as a continuous precision value), starting with the highest precision item. We further assume that this ordering is still maintained if a swap error occurs. We reason that the item to report is selected based on the precision with which its reported feature value is represented, and a swap error occurs when the location of that item is chosen incorrectly. We consider this to be more plausible than the possibility that a location is selected first based on the precision of the associated feature value, and then a different (lower precision) feature value from a different item is reported.

In all the studies making up our experimental dataset, the space of reported features was circular. We therefore used von Mises distributions (a circular analogue of Gaussian) in the definitions of our models, and we measure precision as Fisher Information (as in van den Berg et al., 2012). Our conclusions would not have been significantly affected had we used wrapped normal distributions, or defined precision as the inverse square of circular standard deviation (as in e.g. Bays, 2014).

2.2 Stochastic sampling model

The stochastic sampling model assumes that each memorized feature value is represented by a varying number of discrete samples with fixed precision. The free parameters of this model are the sample precision ω_1 and the mean total number of samples γ. The number of samples that contributes to the representation of each individual item is drawn independently from a Poisson distribution with mean γ/N. The resulting response distribution is then a mixture of von Mises distributions with different precisions, each corresponding to a certain number of samples and weighted with the probability of obtaining that sample count:

$$p(\psi|\theta) = \sum_{k=0}^{\infty} \mathrm{Pr}_{\mathrm{Poisson}}\left(k; \frac{\gamma}{N}\right) \phi_{\circ}(\psi; \theta, \kappa(k\omega_1)) \tag{2}$$

Here, $\mathrm{Pr}_{\mathrm{Poisson}}$ is the Poisson distribution,

$$\mathrm{Pr}_{\mathrm{Poisson}}(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{3}$$

and $\phi_{\circ}$ is the von Mises distribution,

$$\phi_{\circ}(\psi; \theta, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\psi-\theta)}. \tag{4}$$

The term $\kappa(\omega)$ signifies the concentration parameter that yields a von Mises distribution with precision ω. If precision is expressed as Fisher Information, the corresponding value κ can be obtained by numerically inverting the relationship $\omega = \kappa\frac{I_1(\kappa)}{I_0(\kappa)}$. I_n is the modified Bessel function of the first kind. For fitting the model to data, we only compute the sum in the above equation over sample counts k for which $\mathrm{Pr}_{\mathrm{Poisson}}(k; \frac{\gamma}{N}) \geq 10^{-5}$.

In the whole-report task with random report order, the stochastic sampling model predicts that there are no response correlations, because the number of samples is drawn independently for each item. The response distribution for this case is given by

$$p(\boldsymbol{\psi}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\psi_i|\boldsymbol{\theta}), \tag{5}$$

where $p(\psi_i|\boldsymbol{\theta})$ is the probabilty distribution including swap errors as defined in Eq. 1.

In the free response order condition, responses are ordered by the number of samples that represent each item. This sorting induces positive correlations between response errors for consecutive responses within a trial. To compute the response probability for this condition, we determine all possible ordered sequences of sample counts $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N), \lambda_i \geq \lambda_j \, \forall \, i < j$. The probability that each such sequence of sample counts will be generated by the model is

$$\Pr(\boldsymbol{\lambda}) = \prod_{k=0}^{\max(\boldsymbol{\lambda})} \Pr_{\text{Poisson}}\left(k; \frac{\gamma}{N}\right)^{n_k(\boldsymbol{\lambda})} \binom{N - \sum_{j=0}^{k-1} n_j(\boldsymbol{\lambda})}{n_k(\boldsymbol{\lambda})}, \tag{6}$$

where $n_k(\boldsymbol{\lambda})$ is the number of entries in $\boldsymbol{\lambda}$ with $\lambda_i = k$. The probability distribution for a sequence of responses (taking into account swap errors) is then

$$p(\boldsymbol{\psi}|\boldsymbol{\theta}) = \sum_{\boldsymbol{\lambda}} \Pr(\boldsymbol{\lambda}) \prod_{i=1}^{N} \left(p_T \phi_\circ(\psi_i; \theta_i, \kappa(\lambda_i \omega)) + p_{\text{NT}} \sum_{j \in \{1,\ldots,N\}, j \neq i} \phi_\circ(\psi_i; \theta_j, \kappa(\lambda_i \omega)) \right) \tag{7}$$

For the implementation, we again consider only sample counts k with $\Pr_{\text{Poisson}}(k; \frac{\gamma}{N}) \geq 10^{-5}$, and we exclude the least likely sequences $\boldsymbol{\lambda}$ up to a cumulative probability of 10^{-3}.

2.3 Fixed sampling model

The fixed sampling model assumes that a fixed number K of samples, each with a fixed precision ω, is distributed as evenly as possible among the memory items in each trial. The probability distribution of the response is given by

$$p(\psi|\theta) = \frac{K \bmod N}{N} \phi_\circ\left(\psi; \theta, \kappa\left(\left\lceil \frac{K}{N} \right\rceil \omega\right)\right) + \left(1 - \frac{K \bmod N}{N}\right) \phi_\circ\left(\psi; \theta, \kappa\left(\left\lfloor \frac{K}{N} \right\rfloor \omega\right)\right) \tag{8}$$

For the whole-report task, we again assume that responses are ordered by sample count. In the free response order condition, the response probability distribution is described by

$$p(\boldsymbol{\psi}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\psi_i|\boldsymbol{\theta}) \tag{9}$$

with

$$p(\psi|\theta) = \begin{cases} \phi_\circ\big(\psi; \theta, \kappa\big(\lceil \frac{K}{N} \rceil \omega\big)\big), & \text{if } i \leq K \bmod N \\ \phi_\circ\big(\psi; \theta, \kappa\big(\lfloor \frac{K}{N} \rfloor \omega\big)\big), & \text{otherwise.} \end{cases} \tag{10}$$

In the random response order condition, the fixed sampling model predicts negative correlations between response errors in a single trial. We determine all possible unordered sequences of sample counts $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ with $\sum_{i=1}^{N} \lambda_i = K$, where the only possible values for each λ_i are $\lfloor \frac{K}{N} \rfloor$ and $\lceil \frac{K}{N} \rceil$. Each such sequence will occur with equal probability $\Pr(\boldsymbol{\lambda}) = \binom{N}{K \bmod N}$. The response probability $p(\boldsymbol{\psi}|\boldsymbol{\theta})$ can then be expressed in the same way as in Eq. 7.

2.4 Fixed sampling model with random allocation

As a variant of the fixed sampling model described above, we considered a model in which the total number of samples, K, is fixed, but each sample is randomly and independently assigned to one of the N memory items with equal probability. The probability of obtaining a certain number k of samples for a single item is then given by the binomial distribution,

$$\mathrm{Pr}_{\mathrm{Binom}}\left(k; K, \frac{1}{N}\right) = \binom{K}{k}\left(\frac{1}{N}\right)^{k}\left(1-\frac{1}{N}\right)^{K-k}, \tag{11}$$

and the response probability distribution is:

$$p(\psi|\theta) = \sum_{k=0}^{K} \mathrm{Pr}_{\mathrm{Binom}}\left(k; K, \frac{1}{N}\right) \phi_{\circ}(\psi; \theta, \kappa(k\omega)) \tag{12}$$

This model predicts correlations between response errors within a trial of the whole-report task both in the free response order and the random response order conditions. For the free response order condition, the possible sequences of ordered sample counts are $\{\boldsymbol{\lambda}| \sum_{i=1}^{N} \lambda_i = K, \lambda_i \geq \lambda_j \forall i < j\}$. For the random response order condition, the set of possible sample count sequences is $\{\boldsymbol{\lambda}| \sum_{i=1}^{N} \lambda_i = K\}$. In both cases, we can compute the probability of each sequence as

$$\mathrm{Pr}(\boldsymbol{\lambda}) = \frac{\widetilde{\mathrm{Pr}}(\boldsymbol{\lambda})}{\sum_{\boldsymbol{\lambda}'} \widetilde{\mathrm{Pr}}(\boldsymbol{\lambda}')} \tag{13}$$

with

$$\widetilde{\mathrm{Pr}}(\boldsymbol{\lambda}) = \prod_{i=1}^{N} \binom{K - \sum_{j=1}^{i-1} \lambda_j}{\lambda_i}. \tag{14}$$

The response probability $p(\boldsymbol{\psi}|\boldsymbol{\theta})$ in the whole-report task can again be expressed as in Eq. 7.

2.5 Stochastic sampling model with even allocation

A second model variant assumes that the total number of samples varies from trial to trial (as in the stochastic sampling model), but these samples are distributed across memory items as evenly as possible (as in the fixed sampling model). For each trial, the total number of samples is drawn from a Poisson distribution with mean γ. The probability distribution for a single response can then be given as weighted sum of probabilities from the fixed sampling model with different numbers of samples k:

$$p(\psi|\theta) = \sum_{k=0}^{\infty} \mathrm{Pr}_{\mathrm{Poisson}}(k; \gamma) p_{\mathrm{fs}}(\psi|\theta; k) \tag{15}$$

The response probabilities for the whole-report task can be determined in the same fashion as mixtures of the corresponding response probabilities in the fixed sampling model. We note that this introduces error correlations even in the case of the free response order condition, in which there are no correlations in the fixed sampling model.

2.6 Generalized stochastic sampling model

In the generalized stochastic sampling model, the Poisson distribution over precision values is replaced by a negative binomial distribution with an additional discretization parameter p. The distribution of response errors is then given by

$$p(\psi|\theta) = \sum_{k=0}^{\infty} \mathrm{Pr}_{\mathrm{NegBin}}\left(k; \frac{\gamma}{(1-p)N}, p\right) \phi_{\circ}(\psi; \theta, \kappa(k\omega_1 p)) \tag{16}$$

with

$$\mathrm{Pr}_{\mathrm{NegBin}}(k; r, p) = \frac{\Gamma(k+r)}{k!\Gamma(r)} p^r (1-p)^k \tag{17}$$

for $0 < p < 1$. For fitting the model to data, we only compute the sum over sample counts k for which $\mathrm{Pr}_{\mathrm{NegBin}}(k; \frac{\gamma}{(1-p)N}, p) \geq 0.5 \cdot 10^{-4}$. We did not attempt to fit this model to whole-report data, as the number of combinatorial possibilities quickly becomes computationally infeasible as p gets small.

2.7 Gamma model

The Gamma model assumes that recall precision for each item is drawn independently from a Gamma distribution with shape parameter $\frac{\gamma}{N}$ and scale parameter ω_1. This model constitutes the limit case of the generalized stochastic sampling model for $p \to 0$ (see Section 5.2), and has previously been proposed independently by van den Berg et al. (2012) and Fougnie et al. (2012). In the formulation of van den Berg et al., the precision is distributed as $J \sim \mathrm{Gamma}(\bar{J}_1/N^{\alpha}, \tau)$, which is identical to the model described here for $\bar{J}_1 = \gamma$, $\omega_1 = \tau$, and $\alpha = 1$.

The response probability distribution in the Gamma model is described as a continuous mixture of von Mises distributions:

$$p(\psi|\theta) = \int_{\omega=0}^{\infty} p_{\mathrm{Gamma}}\left(\omega; \frac{\gamma}{N}, \omega_1\right) \phi_{\circ}(\psi; \theta, \kappa(\omega)) d\omega \tag{18}$$

with

$$p_{\mathrm{Gamma}}(\omega, k, \theta) = \frac{1}{\Gamma(k)\theta^k} \omega^{k-1} e^{-\frac{\omega}{\theta}}, \tag{19}$$

where Γ is the gamma function. For model fitting, the integral is computed numerically with 1000 possible values of ω, which cover the range of precision values with cumulative probabilities of the gamma distribution from 10^{-5} to $1 - 10^{-5}$.

In the whole-report task, the variable precision model predicts similar correlation patterns as the stochastic sampling model (which likewise draws precision values independently for each item). In the random report order condition, response errors within a trial are uncorrelated, and the precision distribution is given by

$$p(\boldsymbol{\psi}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\psi_i|\boldsymbol{\theta}). \tag{20}$$

In the free report order condition, the probability distribution for a sequence of responses can be described as

$$p(\boldsymbol{\psi}|\boldsymbol{\theta}) = \int_{\omega_1=0}^{\infty} \cdots \int_{\omega_N=0}^{\infty} p(\boldsymbol{\omega}) \cdot \\ \prod_{i=1}^{N} \left(p_T \phi_\circ(\psi_i; \theta_i, \kappa(\omega_i)) + p_{\mathrm{NT}} \sum_{j\in\{1,\ldots,N\}, j\neq i} \phi_\circ(\psi_i; \theta_j, \kappa(\omega_i)) \right) d\omega_1 \ldots d\omega_N, \quad (21)$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ is the sequence of ordered precision values for the responses, and $p(\boldsymbol{\omega})$ is the probability of obtaining such a sequence if each precision value is drawn independently from a gamma distribution. Evaluating this equation is challenging, and in order to obtain an approximation, we discretize the range of possible precision values for each individual response into $m = 12$ bins of equal probability. We can then determine the probability of obtaining a sequence of ordered precision bins $\boldsymbol{b} = (b_1, \ldots, b_N)$ as

$$\Pr(\boldsymbol{b}) = m^{-N} \prod_{k=1}^{m} \binom{N - \sum_{j=0}^{k-1} n_j(\boldsymbol{b})}{n_k(\boldsymbol{b})}. \quad (22)$$

Here, $n_k(\boldsymbol{b})$ denotes the number of entries in $\boldsymbol{b}$ with $b_i = k$, analogously to its use in Eq. 6. The response probability for a sequence of responses can then be given as weighted sum over all possible sequences $\boldsymbol{b}$,

$$p(\boldsymbol{\psi}|\boldsymbol{\theta}) \approx \sum_{\boldsymbol{b}} \Pr(\boldsymbol{b}) \prod_{i=1}^{N} \int_{\omega=\omega_{\mathrm{low}}(b_i)}^{\omega_{\mathrm{high}}(b_i)} p_{\mathrm{Gamma}}\left(\omega; \frac{\gamma}{N}, \omega_1\right) \cdot \\ \left(p_T \phi_\circ(\psi_i; \theta_i, \kappa(\omega)) + p_{\mathrm{NT}} \sum_{j\in\{1,\ldots,N\}, j\neq i} \phi_\circ(\psi_i; \theta_j, \kappa(\omega)) \right) d\omega. \quad (23)$$

Here, $\omega_{\mathrm{low}}(b)$ and $\omega_{\mathrm{high}}(b)$ are the boundaries of the precision bin with index b. Evaluating this form is more feasible, since the integrals over each precision bin can be computed independently (using the same numerical method with a total of 1000 sampling points as above) and then combined.

2.8 Neural population model with heterogeneous tuning curves

We tested a variant of the neural population model that incorporates heterogeneity in the cells' tuning functions of the kind observed in electrophysiological recordings. Specifically, the model takes into account that neurons differ in their minimum (baseline) and maximum (peak) levels of activity, as well as in tuning width. The model also relaxes the assumption that the feature space is covered homogeneously by neural tuning curves, instead selecting neurons' preferred values at random from a uniform distribution. As in the original implementation of the neural population model (Bays, 2014), we assume that each of N feature values in the memory sample array is encoded by a different population of M neurons. The tuning curve of neuron i encoding a feature value θ is given by a scaled von Mises distribution function plus a baseline,

$$f_i(\theta) = \alpha_i + \beta_i \exp\left(\kappa_i(\cos(\theta - \varphi_i) - 1)\right). \quad (24)$$

Here, α_i is the amplitude of the neuron's baseline activity, β_i is the gain of the neuron, κ_i is the von Mises concentration parameter which determines the tuning width, and φ_i is the neuron's preferred feature value. These parameters are chosen randomly for each simulated neuron, with the degree of interneuron variability determined by a global heterogeneity parameter ν.

For $\nu = 0$, the tuning parameters of all neurons are identical (with no baseline activity and homogeneous coverage of the feature space), making the model identical to the standard neural population model described in Bays (2014), and an exact circular analogue of the population model described in the main manuscript. The distributions of parameter values were chosen such that for $\nu = 1$, the population has approximately the heterogeneity observed in orientation-selective neurons in cortical area V1 (Ecker et al., 2010). For $\nu > 1$, individual neurons' parameters vary over wider ranges than observed in these biological populations.

Concretely, the parameters for each neuron are drawn from the following distributions:

$$\log \kappa_i \sim \mathcal{N}(\log \tilde{\kappa}, \nu^2) \tag{25}$$

$$\log \beta_i \sim \mathcal{N}(\log 1, \nu^2) \tag{26}$$

$$\log \alpha_i \sim \mathcal{N}(\log (0.04\nu\beta), \nu^2) \tag{27}$$

$$\varphi_i \sim \mathcal{N}\left(\frac{2\pi}{M}(i-1), \nu^2\right) \bmod 2\pi \tag{28}$$

As in previous versions of the population model, we scaled the total expected activity of all neurons encoding all items with a population gain parameter, γ, which was fixed across changes in set size. However, in the heterogeneous model the information capacity of a neural population varied not only as a function of γ but also all the individual tuning parameters of all the component neurons. In order to equate populations with different randomly-drawn tuning parameters, instead of treating γ as a free parameter for model fitting, we instead used the expected precision of a decoded estimate as the free parameter, and set the population gain γ to a value that would achieve it.

Specifically, we first normalized the tuning curves such that the population would on average fire exactly one spike within the decoding time interval:

$$\check{f}_i(\theta) = \frac{f_i(\theta)}{\frac{1}{2\pi}\sum_{j=1}^{M}\int_{-\pi}^{\pi} f_j(\theta)d\theta} \tag{29}$$

The mean precision of maximum likelihood decoding from this population (assuming full knowledge of the tuning curves) was determined by the expected Fisher Information,

$$\check{\mathcal{I}} = \frac{1}{2\pi}\sum_{i=1}^{M}\int_{-\pi}^{\pi}\left(\frac{d\log \check{f}_i(\theta)}{d\theta}\right)^2 \check{f}_i(\theta)d\theta. \tag{30}$$

The mean decoding precision scales linearly with the number of spikes available for decoding, so in order to achieve the desired precision $\bar{\mathcal{I}}$, we set the global gain parameter γ to

$$\gamma = \bar{\mathcal{I}}/\check{\mathcal{I}} \tag{31}$$

The spike count r_i of neuron i encoding feature value θ in a trial with set size N was then drawn from a Poisson distribution,

$$r_i \sim \mathrm{Poiss}\left(\frac{\gamma}{N}\check{f}_i(\theta)\right). \tag{32}$$

The log likelihood of stimulus feature θ' for a given set of spikes $\boldsymbol{r}$ is (up to addition by a constant),

$$\log \mathcal{L}(\theta'|\boldsymbol{r}) = \sum_{i=1}^{M} r_i \log \check{f}_i(\theta') - \check{f}_i(\theta'). \tag{33}$$

Decoded estimates were obtained as the maximum of this function, and their precision as the width of the likelihood function measured in terms of Fisher Information.

The heterogeneous model therefore has three free (global) parameters that determine the distributions of the single neuron parameters and thereby the predicted error distributions of decoded estimates: the median tuning curve width, $\tilde{\kappa}$, the heterogeneity parameter, ν, and the mean precision for a single stored item, $\bar{\mathcal{I}}$. We estimated the response distributions for this model by sampling. For each combination of values for the parameters $\tilde{\kappa}$, ν, and $\bar{\mathcal{I}}$ on a search grid (described in Section 3), we randomly drew 100 sets of single-neuron parameters for M = 1000 neurons[1] from the distributions specified in Equations 25 to 28. For each set of single-neuron parameters, we generated a neural spiking pattern in response to 1000 randomly chosen feature values θ, and obtained likelihood functions and maximum likelihood estimates $\hat{\theta}$ as described above. The response error distribution is then approximated by a histogram over the decoding errors, $\hat{\theta} - \theta$, averaged over all sets of single-neuron parameters.

3 Fitting procedure

We fit models separately to the behavioral data of each participant in each experiment of the single-report and whole-report dataset. Data of each participant across all set sizes was fit with a single set of parameter values. We employed two different methods to determine the ML parameter values, namely the Nelder-Mead simplex algorithm and grid search over the parameter space.

We used the Nelder-Mead simplex algorithm to fit all models except for the generalized stochastic model and the neural population model with heterogeneous tuning curves, for both single-report and whole-report data. We defined a limited grid of initial parameter values, and ran the fitting algorithm (function *fminsearch* in Matlab) with each possible combination of initial values until a termination tolerance of 0.01 was reached for both the fitted parameter values and the resulting likelihood value. Possible initial values for the sample precision ω_1 were $2^0, 2^2, 2^4$. For stochastic sampling models and gamma model, we first defined initial values for the mean precision at set size one, $E[\omega]$, as $2^2, 2^4, 2^6$, then determined initial values of γ as $\gamma = \frac{E[\omega]}{\omega_1}$. For fixed sampling models, we obtained separate fits for all integer values of K in the range $(0, 25)$, and selected the fit with the highest likelihood. Initial values for p_{NT} were $0.01, 0.05, 0.1$, and for α they were $2^{-0.5}, 2^0, 2^{0.5}$. In variants where these parameters were not used they were fixed at $p_{\mathrm{NT}} = 0$ and $\alpha = 1$, respectively.

We used the grid search to fit the generalized stochastic model (separately for different values of the discretization parameter p) and the neural population model with heterogeneous tuning curves to single-report data. For the latter, likelihood values

[1] The precise number of neurons has very little influence on the response distributions once the average spacing between neurons' preferred values is significantly smaller than the tuning curve width, and therefore we do not treat M as a free parameter in this model.

were determined by sampling, as no closed form solution is available. We also obtained additional fits for the models described in the main text (stochastic sampling, fixed sampling, random-fixed, even-stochastic, and gamma) to verify that the Nelder-Mead simplex algorithm for these models terminated in the global rather than a local maximum of the likelihood function. The parameter grid was spanned by 50 possible values of each model parameter. Values for sample precision, ω_1, were spaced logarithmically in the range $[2^{-4}, 2^5]$, and values for mean precision, $\bar{\omega}$, in the range $[2^{-2}, 2^9]$. For the fixed sampling models, the parameter K took all integer values in the range $[0, 50)$. The values for proportion of non-target responses, p_{NT}, were evenly spaced in the range $[0.0, 0.14]$ for all models. To compute likelihood values in the grid search, response errors were discretized into 101 evenly spaced bins for all models as well as for the behavioral data (ensuring fair comparison between models with closed-form likelihood function and the heterogeneous neural model which requires sampling).

In both fitting methods, we determined a maximum likelihood value L and an associated set of parameters. For comparison between models that differed in the number of free parameters, we computed Akaike information criterion (AIC) scores,

$$\mathrm{AIC} = 2k - 2\log(L), \tag{34}$$

and Bayesian information criterion (BIC) scores,

$$\mathrm{BIC} = \log(n)k - 2\log(L). \tag{35}$$

Here, k is the number of free parameters in each model, and n is the number of data points (number of trials in the single-report tasks, and number of individual responses in the whole-report tasks). AIC and BIC differences for the models described in the main text are depicted in Fig. S1A and E for single-report and whole-report data, respectively. ML fit values of free parameters are reported in Tables S3 and S4.

4 Additional model comparisons

4.1 Heterogeneous population model

The stochastic sampling model is derived from a mathematical idealization of neural population coding, in which the stimulus space is evenly covered with tuning functions of identical width and amplitude (as in Fig. S2A). In reality, neurons selective for the same feature in the same region of the brain vary greatly in their tuning characteristics. To examine the impact of this heterogeneity on the precision of decoded estimates we fit the single-report data with an extension of the standard model in which neurons varied in their tuning, to a degree set by a heterogeneity parameter (ν; see Supplementary Methods for details).

The results indicated that adding heterogeneity to the simulated population improved fits to data (Fig. S2C; ΔAIC = 5.8 ± 1.5, ΔBIC = 1.0 ± 1.5, compared to model with fixed ν = 0; ΔAIC = 8.3 ± 1.8, ΔBIC = 3.4 ± 1.7, compared to stochastic sampling model which incorporates additional simplifications). Fig. S2B shows an illustrative set of neural tuning functions corresponding to the mean fitted heterogeneity parameter value (ν = 0.66 ± 0.08; note that ν = 1 was approximately matched to heterogeneity of orientation-selective neurons in recordings from V1; other parameters: $\tilde{\kappa}$ = 1.53 ± 0.15; $\bar{\mathcal{I}}$ = 18.6 ± 1.0).

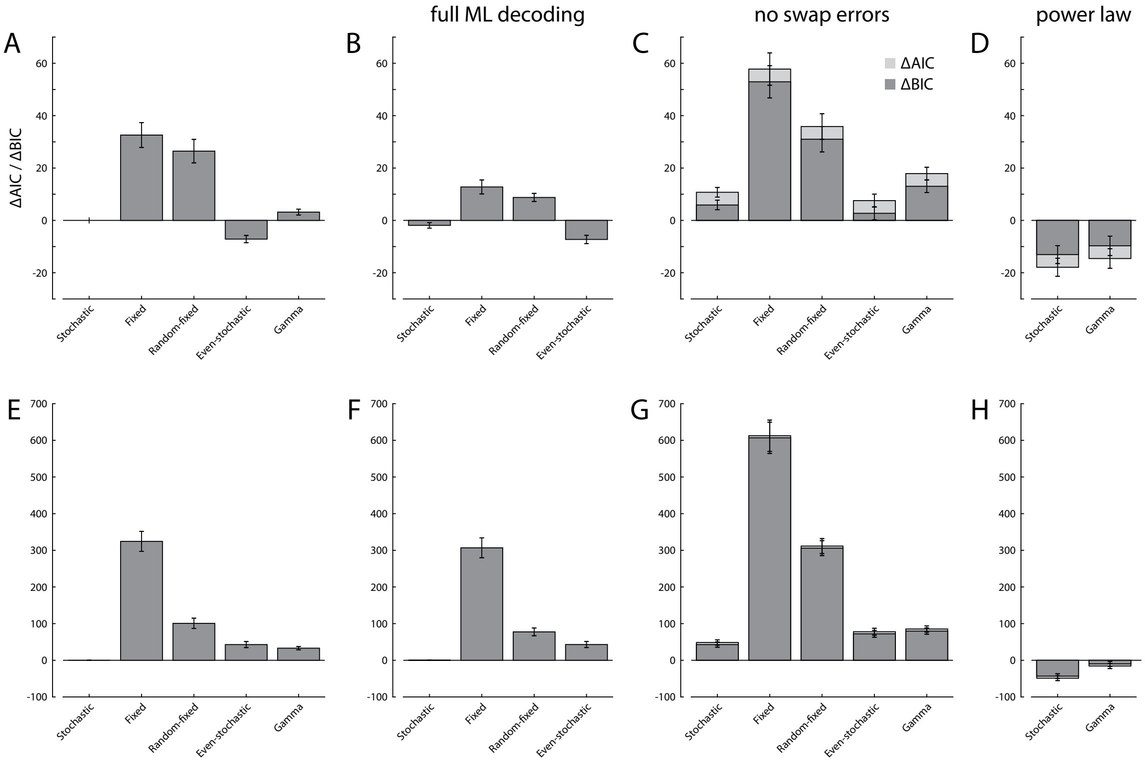

Figure S1: Model comparison for additional model variants. Mean differences in AIC (light gray) and BIC values (dark gray) relative to the stochastic sampling model with swap errors are shown for single-report data (A-D) and whole-report data (E-H). Better models have lower values. Error bars indicate ± 1 SE. (A, E) Models as described in the main text, including a fixed probability of swap errors per memory item. (B, F) Variant of discrete sampling models with exact maximum likelihood decoding from samples in circular feature space. (C, G) Model variants without swap errors (p_{NT} = 0). (D, H) Model variants with power law relationship for set size effect, with exponent α as additional free parameter.

Model	γ or K	ω_1	p_{NT}	α
Stochastic	13.2 ± 1.7	0.241 ± 0.044	0.0245 ± 0.0023	
Fixed	4.90 ± 0.21	1.25 ± 0.13	0.0265 ± 0.0023	
Random-fixed	11.4 ± 0.7	0.579 ± 0.080	0.0238 ± 0.0023	
Even-stochastic	7.31 ± 1.47	2.49 ± 0.15	0.0287 ± 0.0024	
Gamma	8.63 ± 1.94	5.00 ± 0.36	0.0281 ± 0.0024	
full ML decoding for circular feature spaces				
Stochastic	35.0 ± 6.3	0.455 ± 0.068	0.0261 ± 0.0024	
Fixed	11.7 ± 0.9	1.19 ± 0.15	0.0305 ± 0.0027	
Random-fixed	13.8 ± 0.8	0.908 ± 0.096	0.0262 ± 0.0024	
Even-stochastic	15.9 ± 3.3	2.93 ± 0.17	0.0289 ± 0.0024	
excluding swap errors				
Stochastic	8.35 ± 0.81	0.525 ± 0.086		
Fixed	3.67 ± 0.13	1.63 ± 0.16		
Random-fixed	6.91 ± 0.34	1.03 ± 0.11		
Even-stochastic	4.49 ± 0.14	2.99 ± 0.17		
Gamma	3.95 ± 0.73	8.16 ± 0.62		
power law for set size effects				
Stochastic	8.46 ± 0.51	1.41 ± 0.10	0.0310 ± 0.0026	0.809 ± 0.031
Gamma	4.35 ± 0.37	5.60 ± 0.40	0.0339 ± 0.0026	0.809 ± 0.032

Table S3: Parameter values of ML fits for single-report data (mean ± 1 SE across participants and experiments). Outliers with deviation from mean greater than 3 SD were excluded (at most 4 out of 101 individual fit values for each parameter).

Model	γ or K	ω_1	p_{NT}	α
Stochastic	4.30 ± 0.14	2.98 ± 0.26	0.0367 ± 0.0041	
Fixed	2.72 ± 0.10	3.25 ± 0.31	0.0529 ± 0.0036	
Random-fixed	4.21 ± 0.16	3.26 ± 0.28	0.0481 ± 0.0039	
Even-stochastic	2.87 ± 0.07	6.67 ± 0.32	0.0292 ± 0.0039	
Gamma	1.40 ± 0.06	23.2 ± 1.6	0.0384 ± 0.0043	
full ML decoding for circular feature spaces				
Stochastic	4.31 ± 0.14	3.54 ± 0.26	0.0367 ± 0.0041	
Fixed	6.08 ± 0.88	3.57 ± 0.33	0.0551 ± 0.0037	
Random-fixed	4.19 ± 0.17	3.86 ± 0.28	0.0491 ± 0.0040	
Even-stochastic	2.87 ± 0.07	7.23 ± 0.32	0.0292 ± 0.0039	
excluding swap errors				
Stochastic	3.57 ± 0.11	3.48 ± 0.25		
Fixed	2.42 ± 0.06	1.45 ± 0.19		
Random-fixed	3.61 ± 0.12	2.27 ± 0.23		
Even-stochastic	2.56 ± 0.06	6.91 ± 0.32		
Gamma	1.07 ± 0.04	31.7 ± 2.1		
power law for set size effects				
Stochastic	7.06 ± 0.39	4.10 ± 0.21	0.0239 ± 0.0032	1.37 ± 0.03
Gamma	2.33 ± 0.12	19.5 ± 1.4	0.0258 ± 0.0034	1.38 ± 0.03

Table S4: Parameter values of ML fits for whole-report data (mean ± 1 SE across participants and experiments). Outliers with deviation from mean greater than 3 SD were excluded (at most 4 out of 78 individual fit values for each parameter).

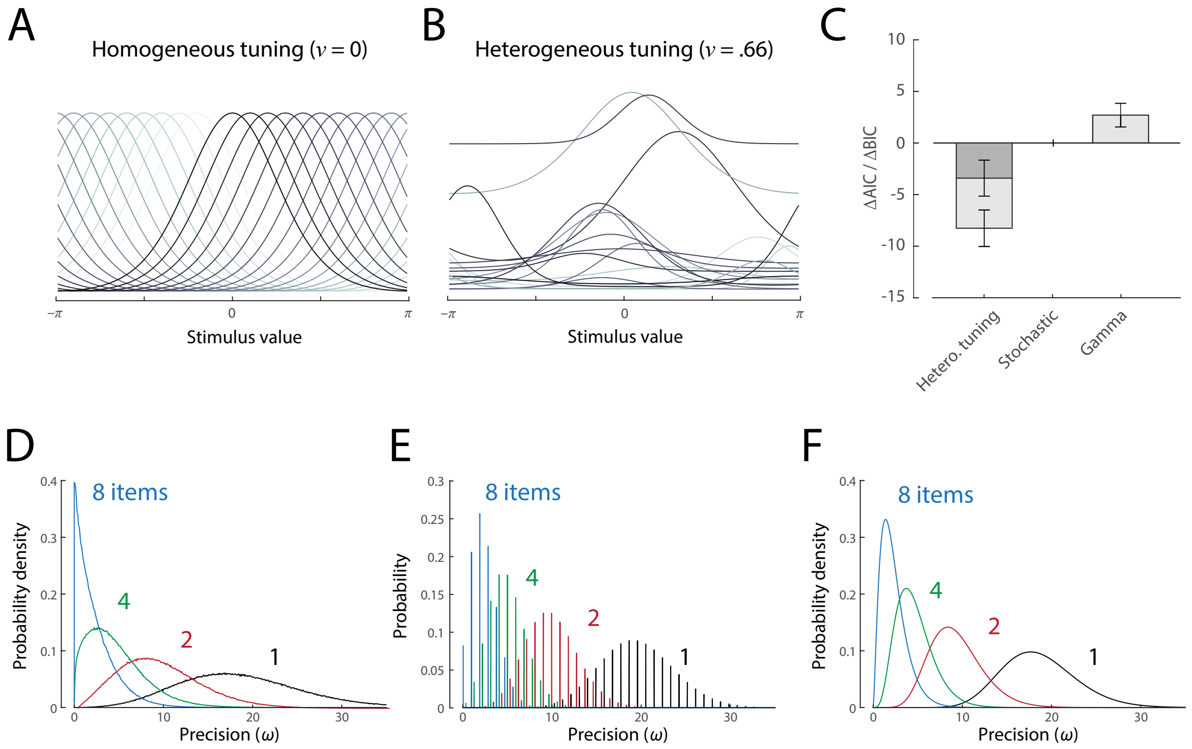

Figure S2: Population coding model with heterogeneous tuning. (A) Examples of tuning functions in the idealized neural population underlying the stochastic sampling model. (B) Example tuning functions corresponding to mean parameters of the fitted heterogeneous model. (C) Results of model comparison on single-report data. The heterogeneous model out-performed the stochastic sampling model, and the model with Gamma-distributed precision, according to both AIC and BIC measures. (D) Distributions of precision of decoded estimates in the heterogeneous model, for different set sizes, based on mean parameters of best fit. (E) Distributions of precision for the stochastic sampling model with matched parameters, for comparison. (F) Distributions of precision for the Gamma model with matched parameters.

The consequences of heterogeneity for the precision of decoded estimates is illustrated in Fig. S2D, again based on best-fitting parameters of the heterogeneous model. The interneuronal variation in tuning results in variability in the precision associated with each spike, which has the effect of making precision distributions continuous (at least in the range of population activity levels that correspond to typical experimental set sizes). Unlike the discrete distribution predicted by the (homogeneous) stochastic sampling model (Fig. S2E), there is no probability of zero precision decoding. This is because the unevenness in coverage of the stimulus space makes no spikes a more probable response to some feature values than others, meaning it is no longer uninformative about the stimulus. However, at lower set sizes there is a sharp increase in probability of very low precision estimates that could not in practice be discriminated from zero (blue curve in Fig. S2D).

Example precision distributions from the Gamma (variable precision; Fougnie et al., 2012; van den Berg et al., 2012) model are shown in Fig. S2F. Interestingly, heterogeneity provides a second putative connection between population coding and Gamma-distributed precision, in addition to the one set out in the main text. This is because a Gamma process (the random process whose marginal distribution at each moment in time is a Gamma distribution) can be constructed from an infinite superposition of different Poisson processes, varying in their rate and, inversely, in their amplitude (i.e. Lévy jump size). So, with the right kind of heterogeneity, a population model with Poisson spiking could theoretically result in estimates with exactly Gamma-distributed precision.

4.2 Full maximum likelihood decoding for circular feature spaces

The assumption that the decoding precision in sampling models increases in equal discrete steps with the number of samples is only strictly true for certain cases. For circular feature spaces with samples drawn from a von Mises distributions, it is only an approximation. An exact method to compute the distribution of response errors arising from ML decoding in circular space was derived in Bays (2014) and Bays (2016). For a given number of samples, m, that are drawn independently from the same von Mises distribution with concentration parameter $\kappa_1 = \kappa(\omega_1)$, the resulting distribution of decoding error can be described as a continuous scale mixture of von Mises distributions,

$$p(\psi|\theta, m) = \int p(r|m, \kappa_1) p(\phi_\circ(\psi; \theta, r\kappa_1) dr \tag{36}$$

with

$$p(r|m, \kappa) = \frac{I_0(\kappa r)}{(I_0(\kappa))^m} r\psi_m(r). \tag{37}$$

Here, $r\psi_m(r)$ is the probability density function for the resultant length r of a uniform random walk of m steps. The distribution of response errors in each sampling model is then a mixture of probability distributions $p(\psi|\theta, m)$, weighted with the probability of obtaining m samples for an item.

We obtained ML fits using this method to determine response error distributions for the stochastic sampling model, fixed sampling model, and random-fixed and even-stochastic variants (the method is not compatible with the Gamma model, since this model does not use discrete samples). The quality of fit was improved for all models

(Fig. S1B and F), with the largest changes for fixed sampling model and random-fixed model fits to single-report data. However, the overall pattern of results did not change when employing exact ML decoding instead of the simpler approximation. For the stochastic sampling model, the change in quality of fit was minimal, supporting our assumption that the simpler form provides a close approximation to full maximum likelihood decoding for this model.

4.3 Excluding swap errors

We obtained ML fits of the behavioral data for all models without swap errors by keeping the parameter p_{NT} fixed at zero. The quality of fit for all models decreased substantially in this variant, independent of whether we measured it via AIC or BIC values (which differ in how strongly they penalize additional free parameters; Fig. S1C and G). This is consistent with previous findings that inclusion of swap errors improves model fit (e.g. Bays et al., 2009; van den Berg et al., 2014).

4.4 Power law for set size effects

The variable precision model of van den Berg et al. (2012) proposed that the effect of set size on mean recall precision is best explained by a power law of the form $E[\omega] \propto N^{-\alpha}$, with a free parameter α. We added this parameter to the formulations of the stochastic sampling model and the Gamma model (the other models assume that a certain number of samples is distributed between all items, thus the power law is not readily applicable). Quality of fit was improved for both models (Fig. S1 D and E), with the stochastic sampling model still providing better quality of fit for both single-report and whole-report datasets.

The distribution of parameter values for α in ML fits was very similar between the two models, but differed markedly between single-report and whole-report data (Tables S3 and S4). The mean value was less than one for the former, but greater than one in the latter, indicating an additional penalty for recall performance at higher set sizes (beyond what would be expected by even distribution of a fixed amount of memory resources between items). This may be explained by the requirement to sequentially report all items in the sample array, which increases the effective delay between sample array and report and may cause moderate interference effects on VWM representations for later-reported items.

5 Negative binomial distribution

In the generalized stochastic sampling model, precision follows a negative binomial distribution,

$$\frac{\omega}{\omega_1 p} \sim \mathrm{NegBin}\left(\frac{\xi}{1-p}, p\right), \tag{38}$$

with parameters $\xi = \frac{\gamma}{N} > 0$, $\omega_1 > 0$, and $0 < p < 1$. The probability of a precision value ω with $\frac{\omega}{\omega_1 p} \in \mathbb{Z}^{\geq 0}$ is determined as

$$\Pr(\omega) = \frac{\Gamma(\frac{\omega}{\omega_1 p} + \frac{\xi}{1-p})}{(\frac{\omega}{\omega_1 p})!\,\Gamma(\frac{\xi}{1-p})} p^{\frac{\xi}{1-p}} (1-p)^{\frac{\omega}{\omega_1 p}}. \tag{39}$$

This probability distribution has mean $E[\omega] = \xi\omega_1$ and variance $Var[\omega] = \xi{\omega_1}^2$. Interpreted as a discrete sampling model, the expected number of samples per item is ξ/p with variance ξ/p^2.

5.1 Poisson distribution as limit case

The Poisson distribution is the limit case of the negative binomial distribution for $p \to 1$. Using

$$\frac{\omega}{\omega_1 p} \overset{p\to 1}{\longrightarrow} \frac{\omega}{\omega_1}, \tag{40}$$

and

$$p^{\frac{a}{1-p}} = e^{a\frac{\ln p}{1-p}} \overset{p\to 1}{\longrightarrow} e^{-a} \tag{41}$$

we obtain

$$\lim_{p\to 1} \Pr(\omega) = \lim_{p\to 1}\left(\frac{\Gamma(\frac{\omega}{\omega_1 p} + \frac{\xi}{1-p})}{(\frac{\omega}{\omega_1 p})!\;\Gamma(\frac{\xi}{1-p})} p^{\frac{\xi}{1-p}}(1-p)^{\frac{\omega}{\omega_1 p}}\right) \tag{42}$$

$$= \lim_{p\to 1}\left(\frac{\Gamma(\frac{\omega}{\omega_1} + \frac{\xi}{1-p})}{(\frac{\omega}{\omega_1})!\;\Gamma(\frac{\xi}{1-p})} e^{-\xi}(1-p)^{\frac{\omega}{\omega_1}}\right). \tag{43}$$

For $\xi > 0$ and $\omega_1 > 0$, we have

$$\frac{\frac{\xi}{1-p}}{\frac{\omega}{\omega_1}} \overset{p\to 1}{\longrightarrow} \infty, \tag{44}$$

and can apply the rule

$$\frac{\Gamma(x+a)}{\Gamma(x)} \overset{x/a\to\infty}{\longrightarrow} x^a. \tag{45}$$

This yields

$$\lim_{p\to 1} \Pr(\omega) = \lim_{p\to 1}\left(\frac{1}{(\frac{\omega}{\omega_1})!}\left(\frac{\xi}{(1-p)}\right)^{\frac{\omega}{\omega_1}} e^{-\xi}(1-p)^{\frac{\omega}{\omega_1}}\right) \tag{46}$$

$$= \frac{\xi^{\frac{\omega}{\omega_1}} e^{-\xi}}{(\frac{\omega}{\omega_1})!}, \tag{47}$$

showing that for this limit case $\frac{\omega}{\omega_1}$ is Poisson distributed,

$$\frac{\omega}{\omega_1} \sim \mathrm{Poisson}(\xi). \tag{48}$$

5.2 Gamma distribution as limit case

The continuous Gamma distribution can be shown to be the limit case of the negative binomial distribution for $p \to 0$. For this case, we have

$$\frac{\xi}{1-p} \overset{p\to 0}{\longrightarrow} \xi, \tag{49}$$

and we can use the limit rule

$$(1-p)^{a/p} = e^{a\frac{\ln(1-p)}{p}} \overset{p\to 0}{\longrightarrow} e^{-a}. \tag{50}$$

We can therefore write the limit of the negative binomial distribution as

$$\lim_{p\to 0}\mathrm{Pr}(\omega) = \lim_{p\to 0}\left(\frac{\Gamma(\frac{\omega}{\omega_1 p}+\frac{\xi}{1-p})}{(\frac{\omega}{\omega_1 p})!\;\Gamma(\frac{\xi}{1-p})}p^{\frac{\xi}{1-p}}(1-p)^{\frac{\omega}{\omega_1 p}}\right) \tag{51}$$

$$= \lim_{p\to 0}\left(\frac{\Gamma(\frac{\omega}{\omega_1 p}+\xi)}{(\frac{\omega}{\omega_1 p})!\;\Gamma(\xi)}p^{\xi}e^{-\frac{\omega}{\omega_1}}\right) \tag{52}$$

Furthermore, for any $\xi > 0$, we have

$$\frac{\frac{\omega}{\omega_1 p}}{\xi} \overset{p\to 0}{\longrightarrow} \infty, \tag{53}$$

and we can apply the rule

$$\frac{\Gamma(x+a)}{x!} \overset{x/a\to\infty}{\longrightarrow} x^{a-1}. \tag{54}$$

We obtain

$$\lim_{p\to 0}\mathrm{Pr}(\omega) = \frac{1}{\Gamma(\xi)}\left(\frac{\omega}{\omega_1 p}\right)^{\xi-1}p^{\xi}e^{-\frac{\omega}{\omega_1}} \tag{55}$$

$$= p\omega_1\frac{1}{\Gamma(\xi)\omega_1^{\xi}}\omega^{\xi-1}e^{-\frac{\omega}{\omega_1}} \text{ for } \frac{\omega}{\omega_1 p} \in \mathbb{Z}^{\geq 0}. \tag{56}$$

We can now make the transition to the continuous probability density,

$$p(\omega) = \frac{1}{\Gamma(\xi)\omega_1^{\xi}}\omega^{\xi-1}e^{-\frac{\omega}{\omega_1}}, \tag{57}$$

which matches the Gamma distribution with

$$\omega \sim \mathrm{Gamma}(\xi,\omega_1). \tag{58}$$

References

Adam, K. C., Vogel, E. K., and Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, 97:79–97.

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, 34(10):3632–3645.

Bays, P. M. (2016). A signature of neural coding at human perceptual limits. *Journal of vision*, 16(11):4–4.

Bays, P. M., Catalao, R. F., and Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of vision*, 9(10):7–7.

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., and Husain, M. (2011a). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of vision*, 11(10):6–6.

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

van den Berg, R., Awh, E., and Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological review*, 121(1):124.

van den Berg, R., Shin, H., Chou, W.-C., George, R., and Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, 109(22):8780–8785.

Ecker, A. S., Berens, P., Keliris, G. A., Bethge, M., Logothetis, N. K., and Tolias, A. S. (2010). Decorrelated neuronal firing in cortical microcircuits. *Science*, 327(5965):584–587.

Fougnie, D., Suchow, J. W., and Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature communications*, 3:1229.

Gorgoraptis, N., Catalao, R. F., Bays, P. M., and Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience*, 31(23):8502–8511.

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

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

Zhang, W. and Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192):233–235.