

Stochastic sampling provides a unifying account of visual working memory limits

Sebastian Schneegans^a , Robert Taylor^a , and Paul M. Bays^{a,1} 

^aDepartment of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom

Edited by Wilson S. Geisler, The University of Texas at Austin, Austin, TX, and approved July 14, 2020 (received for review March 6, 2020)

Research into human working memory limits has been shaped by the competition between different formal models, with a central point of contention being whether internal representations are continuous or discrete. Here we describe a sampling approach derived from principles of neural coding as a framework to understand working memory limits. Reconceptualizing existing models in these terms reveals strong commonalities between seemingly opposing accounts, but also allows us to identify specific points of difference. We show that the discrete versus continuous nature of sampling is not critical to model fits, but that, instead, random variability in sample counts is the key to reproducing human performance in both single- and whole-report tasks. A probabilistic limit on the number of items successfully retrieved is an emergent property of stochastic sampling, requiring no explicit mechanism to enforce it. These findings resolve discrepancies between previous accounts and establish a unified computational framework for working memory that is compatible with neural principles.

visual working memory | population coding | resource model | capacity limits

Working memory refers to the nervous system's ability to form stable internal representations that can be actively manipulated in the pursuit of behavioral goals. A classical view of visual working memory (VWM) held that it was organized into a limited number of memory slots, each capable of holding a single object (1, 2). This model was subsequently modified to allow multiple slots to hold the same object and be combined on retrieval to achieve higher precision (3). This "slots+averaging" model incorporated aspects of an alternative view, which holds that VWM is better conceptualized as a continuous resource that can be flexibly distributed between different objects or visual elements (4, 5), accounting for set size effects in delayed reproduction tasks (6) (Fig. 1A) and flexibility in prioritizing representations (7). Variable precision models (8, 9) additionally proposed that the amount of memory resource is not fixed but varies randomly from item to item and trial to trial. An alternative approach (10) sought to explain VWM errors from neural principles as decoding variability in population representations (11), with the limited memory resource equated to the total neural activity dedicated to storage. Here we show that each of these influential accounts of VWM can be interpreted within a common framework based on the statistical principle of sampling (12–18).

Sampling Interpretation of Population Coding

We first show how a population coding model (10) can, with some simplifying assumptions, be reinterpreted in terms of sampling (Fig. 1A–C). We consider a mathematically idealized population of independent neurons encoding a one-dimensional (1D) stimulus feature θ , where the amplitude of each cell's activity is determined by its individual tuning function. Neurons are assumed to share the same tuning function, merely shifted so the peak lies at each neuron's preferred feature value φ_i ,

$$f_i(\theta) = f(\theta - \varphi_i). \quad [1]$$

Discrete spikes are generated from the cells' activity via independent Poisson processes. If we pick, at random, any spike generated by the neural population in response to a stimulus value θ , we can determine the probability that it was produced by a neuron with preferred feature value φ . If we assume dense uniform coverage of the underlying feature space by neural tuning curves, this yields a continuous probability distribution $p(\varphi)$ over the space of preferred feature values (Fig. 1C). This distribution has the same shape as the neural tuning curves and is centered on the true stimulus value,

$$p(\varphi) \propto f(\theta - \varphi). \quad [2]$$

Thus, if we associate each spike with the preferred feature value of the neuron that generated it (the principle of population vector decoding; ref. 19), we can interpret the spiking activity of the population as a set of noisy samples of the true stimulus value, drawn from the distribution $p(\varphi)$.

Retrieval of a feature value is modeled as decoding of the spikes generated within a fixed time window. In the idealized case with Gaussian tuning functions, the maximum likelihood (ML) decoder generates an estimate by simple averaging of the spike values,

$$\hat{\theta}_{ML} = \frac{1}{n} \sum_j \varphi_{(j)}, \quad [3]$$

where $\varphi_{(j)}$ is the preferred feature value of the neuron that generated the j th spike.

Due to the superposition property of Poisson processes, the number of spikes—or samples—generated by the neural

Significance

We demonstrate that three of the most prominent accounts of visual working memory in the psychology and neuroscience literature—the slots+averaging model, the variable precision model, and the population coding model—can all be expressed in the common mathematical framework of sampling. This reformulation allows us to pinpoint the key differences between these models, and to determine which factors are critical to account for the observed patterns of recall errors across different human psychophysical experiments. Moreover, the sampling framework provides a possible neural grounding for these models in the spiking activity of neuronal populations, as well as a link to existing theories of capacity limits in visual attention.

Author contributions: S.S., R.T., and P.M.B. performed research; S.S., R.T., and P.M.B. developed the theory; S.S. and P.M.B. carried out the analysis; S.S. and P.M.B. wrote the paper; and P.M.B. conceived and directed the project.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).

¹To whom correspondence may be addressed. Email: pmb20@cam.ac.uk.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2004306117/-DCSupplemental>.

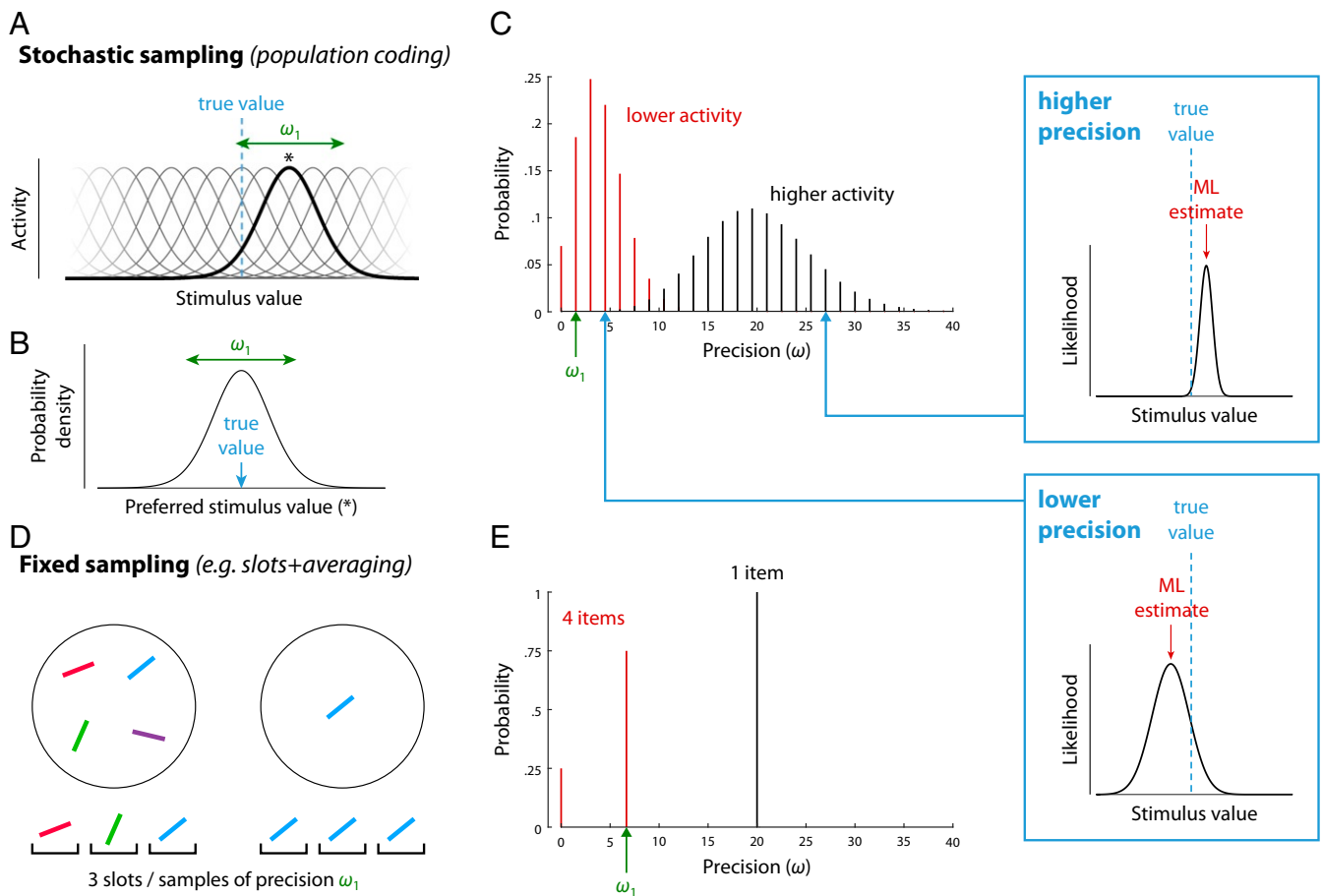


Fig. 1. Sampling interpretation of working memory models. (A–C) A theoretical account of neural population coding can be reinterpreted as sampling. (A) The stimulus-evoked response of spiking neurons in an idealized population depends on their individual tuning (one neuron's tuning function and preferred value [*] is highlighted). (B) Probability distribution over stimulus space obtained by associating a spike with the preferred stimulus of the neuron that generated it. (C) Precision of maximum likelihood estimates based on spikes emitted in a fixed decoding window. Precision, defined as the width of the likelihood function (*insets*), is discretely distributed as a product of the tuning precision (ω_1) and the number of spikes, which varies stochastically. Assuming normalization of total activity encoding multiple items, larger set sizes correspond to less mean activity per item. (D and E) An account based on averaging limited memory slots can also be described as sampling. (D) Allocation of a fixed number of samples or slots (here, three) to memory displays of different sizes. (E) Precision is discretely distributed as a product of the tuning width, ω_1 , and the number of samples allocated per item.

population within the decoding window is also a Poisson random variable. If the total spike rate in the neural population is normalized (20), or fixed at a population level γ , it implements a form of limited resource (10). This resource is continuous—unlike the discrete number of samples—and can be distributed between memory items, depending on task demands (e.g., prioritizing one item that is cued as a likely target). We will focus on the simplest case, in which the total spike rate is distributed evenly among all memory items, resulting in a mean number of samples available for decoding each stimulus that is inverse to the set size N . This has been shown to quantitatively capture the set size effect in single-report delayed reproduction tasks (Fig. 24).

The actual number of samples available in this model for decoding each item in a single trial, n_k , is a discrete random variable independently drawn from a Poisson distribution, with its mean determined by the spike rate for that item,

$$n_k \sim \text{Pois} \left(\frac{\gamma}{N} \right). \quad [4]$$

The neural population model can therefore be interpreted as a *stochastic sampling model*.

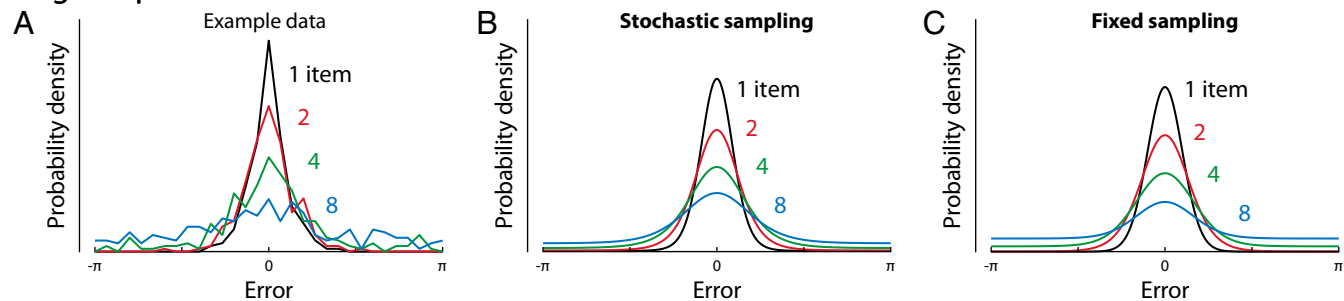
Fixed Sampling Models

The most prominent discrete representation account of VWM, the slots+averaging model (3), can also readily be interpreted in terms of sampling (Fig. 1 D and E). Each slot is postulated to hold a representation of a single item with a fixed precision, and so provides a noisy sample of the item's feature value (or values; the sampling interpretation is agnostic as to feature- vs. object-based views of VWM; refs. 21 and 22). Multiple slots, or samples, that correspond to the same object are averaged at retrieval to enhance the precision of the estimated stimulus feature. Thus, the format of representation and the decoding mechanism are identical to the stochastic sampling model. There is one critical difference, however: The slots+averaging model assumes that the total number of samples available for representing multiple items is fixed, that is,

$$\sum_k^N n_k = K. \quad [5]$$

This has also been the most common assumption in previous sampling-based models in the attentional and memory literature (refs. 12–14, but see ref. 23). We will refer to this as a *fixed sampling model*.

Single Report



Whole report

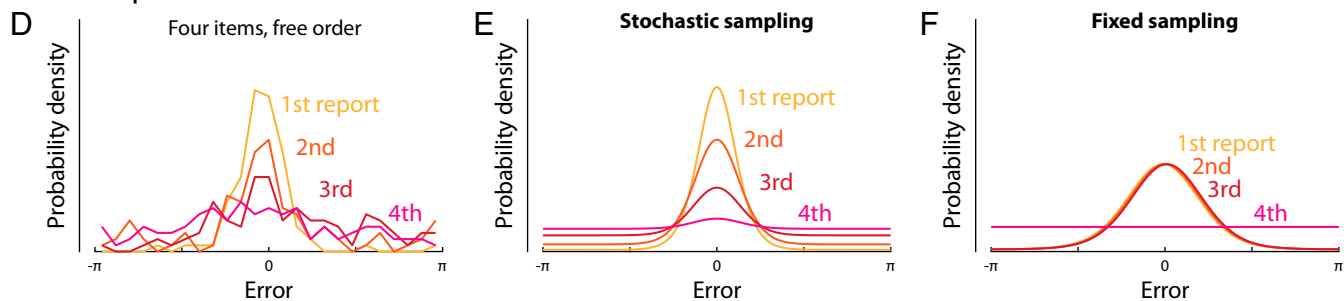


Fig. 2. Response distributions and model fits in delayed reproduction tasks. (A) Distributions of response errors in a single-report task for a representative participant at different set sizes (10). (B and C) ML fits of the data in A with the stochastic sampling model and fixed sampling model, respectively. (D) Distributions of response errors in a whole-report task for a representative participant at set size four, showing how errors increase with the (freely chosen) order of sequential report (24). (E and F) ML fits of the participant's data with the stochastic sampling model and fixed sampling model, respectively. Fits are based on results from all set sizes, not only the single set size shown in D.

Predictions for Precision and Error Distributions

We now consider the distribution of representational precision in these models. For any particular set of samples, ϕ , the information they provide about the stimulus is described by the likelihood function, $\mathcal{L}(\theta; \phi) = p_{\theta}(\phi|\theta)$, equivalent to the conditional probability of obtaining those samples given different values of the stimulus. The width of the likelihood function is a measure of uncertainty in the estimate: A set of samples with a broad likelihood function (Fig. 1C, *Bottom Inset*) is compatible with many different feature values, whereas a narrow likelihood function (Fig. 1C, *Top Inset*) identifies a value more precisely. While a pattern of samples may have a sharp likelihood function with a peak far from the true estimate (a kind of “false alarm”), statistically, this is unlikely.

If the sample values follow a normal distribution with variance σ^2 centered on the true stimulus value, then the likelihood function is also normal, with a width that depends only on the number of samples available for decoding,

$$\mathcal{L}(\theta; \phi) \propto \phi\left(\theta; \hat{\theta}_{ML}, \frac{\sigma^2}{n_k}\right). \quad [6]$$

Furthermore, for a specified number of samples, the ML estimate is distributed around the true stimulus value as a normal with the same width as the likelihood,

$$\hat{\theta}_{ML}|n_k \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n_k}\right). \quad [7]$$

This correspondence between uncertainty, as expressed in the likelihood width, and trial-to-trial variability is not universal, but does apply to all of the models considered in this study, and justifies defining the precision of an individual estimate (which we will denote ω) as the precision of its corresponding likelihood function (see *SI Appendix, Fig. S1* for a detailed illustration).

Adopting this definition explicitly (see also refs. 25 and 26) allows us to treat precision as a random variable with a defined probability function, describing variation in the reliability of estimates while also predicting the distribution of errors across trials. This will prove critical in fitting data from whole-report tasks (Fig. 2D and below).

For the stochastic sampling model based on population coding, likelihood precision has a Poisson distribution (Fig. 1C), scaled by the precision of a single sample which is determined by the neural tuning function, $\omega_1 = 1/\sigma^2$,

$$\frac{\omega}{\omega_1} \sim \text{Poisson}\left(\frac{\gamma}{N}\right). \quad [8]$$

Example distributions of decoding error are shown in Fig. 2B and E, where we have made a transition from 1D Euclidean to a circular stimulus space, corresponding more closely to the feature dimensions (e.g., orientation, hue) commonly used experimentally. The distribution of errors can be described as a scale mixture of normal distributions with precision proportional to the sample count (*SI Appendix, Fig. S1*; due to the circular stimulus space, this is a close approximation rather than exact; see *SI Appendix, Supporting Information Text*). The dispersion of errors increases with decreasing activity (e.g., as a result of increasing set size; Fig. 2B), and the distribution deviates from normality, with this effect being particularly evident at lower activity levels (blue curve) where long tails are observed.

For the fixed sampling model, making the common assumption that samples are distributed as evenly as possible among items (9, 27), we obtain a discrete distribution over, at most, two precision values (Fig. 1E), which are multiples of the precision of one sample, ω_1 . As in the stochastic model, mean precision is inversely proportional to set size, but, because the distributions over precision differ, the fixed and stochastic models make distinct, testable predictions for error distributions (Fig. 2).

Response Errors Discriminate between Models

We tested the ability of stochastic and fixed sampling models to capture response errors in delayed reproduction tasks (*SI Appendix*, Fig. S2). We fit the models to a large dataset of single-report tasks originating from different laboratories (*SI Appendix*, Table S1) and also to a set of whole-report experiments (24), in which participants reported the feature values of all items in a sample array, either in a prescribed random order or in an order freely chosen by each participant on each trial. While only a single study, the whole-report results include information regarding correlations in errors between items represented simultaneously in VWM that could differentiate the models. On free choice trials, we assumed that participants gave their responses in order of decreasing precision (corresponding to decreasing number of samples and increasing likelihood width). This assumption is supported by previous findings that humans have knowledge about the uncertainty with which individual items are recalled (8, 25).

Overall, the stochastic model fit data substantially better than the fixed sampling model for both single-report (Fig. 3A; difference in log likelihood per participant, $\Delta LL = 16.3 \pm 2.37$ [M \pm SE]) and whole-report tasks (Fig. 3B; $\Delta LL = 162 \pm 13.6$), indicating that stochasticity is critical for capturing behavioral performance (see also *SI Appendix*, Fig. S3). The response error distributions in the whole-report task with freely chosen response order have previously been argued to provide evidence for a fixed item limit (24), since they approach uniform distributions for the later responses at high set sizes (Fig. 2D; see *SI Appendix*, Figs. S4 and S5 for full behavioral results and model fits). However, this qualitative observation is also predicted by the stochastic sampling model with responses ordered by precision, as the lowest precision retrievals will be based on few, or no, samples (Fig. 2E). Quality of fits could be further improved by taking into account deterioration of recall precision with increasing retention intervals (*SI Appendix*, Figs. S3J and S5), modeled as random drift of encoded feature values over time (28) (*SI Appendix*).

In contrast, the quantitative changes in error distribution with response order and set size were relatively poorly fit by the fixed sampling model (Fig. 2F). In particular, when the set size exceeds the fixed sample count, each item is represented by either one or zero samples, so this model cannot reproduce the graded decline in precision with response order that is also present in individual participants' data (and does not merely arise at the group level due to averaging across participants with different capacities).

We tested two intermediate model versions in order to further dissociate the specific aspects in which the fixed and stochastic sampling models differ, and determine the significance of each for capturing human performance. In the *random-fixed* model, the total number of samples was fixed but distributed randomly between items. This model provided an improved fit to data compared to the fixed model with even allocation (moderately for single-report, $\Delta LL = 3.07 \pm 1.10$; strongly for whole-report, $\Delta LL = 112 \pm 11.9$), but was still substantially worse than the stochastic model in both cases (single-report, $\Delta LL = 13.2 \pm 2.24$; whole-report, $\Delta LL = 50.4 \pm 7.03$). In the *even-stochastic* model, the total number of samples was a Poisson random variable, but the samples were distributed as evenly as possible between items. This model achieved a better fit to single-report data than the stochastic model with independent sample counts for each item ($\Delta LL = 3.57 \pm 0.697$), but provided a much worse fit to whole-report data ($\Delta LL = 21.4 \pm 4.12$). Combining evidence across all participants and tasks, the stochastic model with independent sample counts was strongly preferred over this and the other alternative models (total $\Delta LL > 1,450$; Fig. 3C).

Generalizing the Stochastic Model

For the models examined above, typical fitted parameters indicate that estimates are based on relatively small numbers of

samples (e.g., mean of ~ 13 samples based on fits to single-report data). One result is that the precision of decoded estimates could take on only a limited set of possible values, and error distributions reflect a discrete mixture of distributions with different widths. From a neural perspective, while consistent with the remarkable fidelity with which single neurons' activity encodes visual stimuli (29, 30), such small sample counts nonetheless seem unlikely when interpreted as spike counts (see *Toward Biophysically Realistic Models*). To investigate whether discreteness and/or low numbers of samples are important for reproducing human performance, we therefore implemented a generalization of the stochastic model in which the number of samples was free to vary.

The distribution over precision values in the generalized stochastic model was obtained as a scaling of the negative binomial distribution,

$$\frac{\omega}{\omega_1 p} \sim \text{NegBin} \left(\frac{\gamma}{N} \frac{1}{1-p}, p \right). \quad [9]$$

This distribution has previously been proposed to model neural spiking activity (31), and it retains the characteristic relationship between mean and variability in the scaled Poisson distribution: The Fano factor (the ratio of variance to mean) is constant, equal to the value of a single sample, $\text{Var}[\omega]/E[\omega] = \omega_1$. This distinguishes the stochastic models from the fixed sampling model, where the Fano factor is at or close to zero (mean ~ 0.25 of ω_1 based on ML parameters and typical set sizes) and varies in an idiosyncratic manner between set sizes, due to the varying combinatorial possibilities of allocating a fixed number of samples to a fixed number of items (Fig. 3D, purple).

The parameter p in the generalized stochastic model controls the discretization of the precision distribution: $p = 1$ corresponds to the Poisson model described above and illustrated in Fig. 4A (strictly, Eq. 8 is the limit of Eq. 9 as $p \rightarrow 1$), while $p < 1$ corresponds to a stochastic model with a greater mean number of samples, $\bar{n} = \gamma/p$, each with a lower individual precision, $\omega_1 p$. The mean and variance in precision ($E[\omega] = \omega_1 \gamma/N$ and $\text{Var}[\omega] = \omega_1^2 \gamma/N$), and thus also the Fano factor, are independent of the discretization p . Examples of precision distributions with different discretizations are shown in Fig. 4B and C.

As the discretization parameter becomes very small ($p \rightarrow 0$), the number of samples becomes very large, and the distribution of precision described by Eq. 9 approaches a continuous function (Fig. 4D and *SI Appendix*), specifically the Gamma distribution,

$$\omega \sim \text{Gamma} \left(\frac{\gamma}{N}, \omega_1 \right). \quad [10]$$

Two previous studies (8, 9) independently proposed that a continuous scale mixture of normal distributions with Gamma-distributed precision provided a good account of VWM data, but did not provide a theoretical motivation for this choice of distribution. In particular, ref. 9 proposed distributing precision as $\text{Gamma}(\bar{J}_1/N^\alpha, \tau)$, with \bar{J}_1 , τ , and α as free parameters. With $\alpha = 1$, this is identical to Eq. 10 (see *SI Appendix* for results regarding this parameter). We can now explain Gamma-distributed precision as a limit case of the stochastic sampling model with large numbers of low-precision samples.

Fig. 3E, *Top* shows the results of fitting the generalized stochastic model with different levels of discretization, p , to the single-report dataset. The best fit was obtained with a discretization roughly one-third that of the Poisson model, $p = 0.39$. However, varying discretization produced differences in fit an order of magnitude smaller than those between fixed and stochastic sampling (varying by ~ 1.5 versus ~ 15 log likelihood points). Fitting the same model with p as a free parameter that could

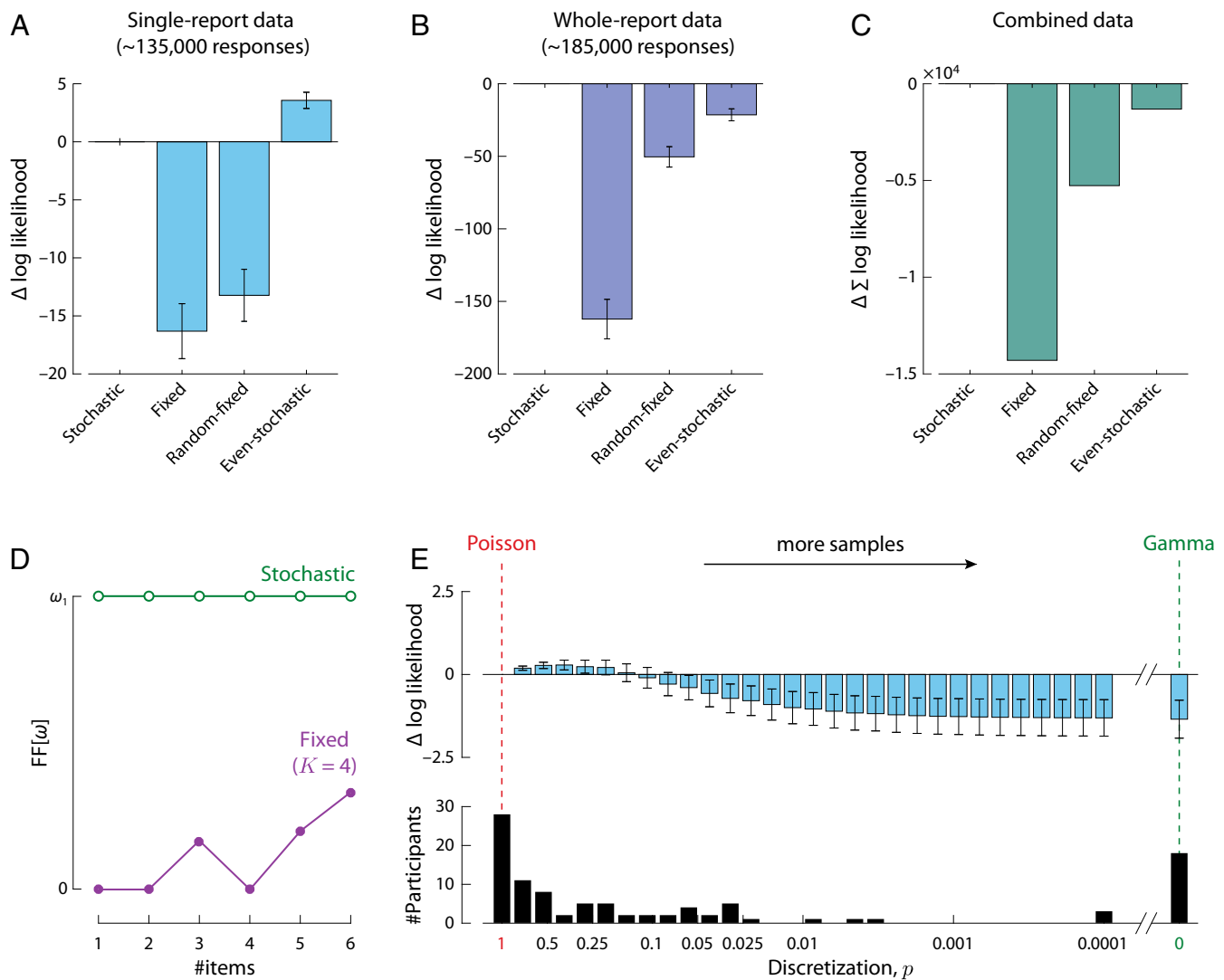


Fig. 3. Model comparison based on single- and whole-report data. (A) Mean difference in log likelihood of each model from the stochastic sampling model (with independence between items), for a benchmark dataset of single-report experiments. More positive values indicate better fits to data. Error bars indicate ± 1 SE across participants. (B) The same comparison for a set of whole-report experiments. (C) Total difference in log likelihood between models across single- and whole report experiments. (D) Fano factor (ratio of variance to mean) of precision distribution. A constant Fano factor is characteristic of the stochastic model and contrasts with the varying Fano factor (dependent on set size and number of samples) in fixed sampling. (E) Mean difference in log likelihood for differing levels of discretization in the generalized stochastic model (Top), and number of participants best fit with each discretization level (Bottom). Differences in log likelihood are plotted relative to the maximum discretization ($p = 1$; Left) corresponding to the standard stochastic model with Poisson-distributed precision. Lower discretization ($p < 1$) corresponds to more samples each of lower precision, converging to a continuous Gamma distribution over precision as p approaches zero (Right). All models have the same number of free parameters and include a fixed per-item probability of swap errors (SI Appendix).

vary between participants, we found that ML estimates of discretization were very broadly distributed (Fig. 3 E, Bottom), with a majority of participants (72%) best described by a sampling model with less discreteness than the Poisson, and a minority (18%) better captured by the continuous limit ($p \rightarrow 0$) than any discrete value of p we tested (as low as 0.0001, corresponding to $\sim 100,000$ samples). Formal model comparison was equivocal with respect to an advantage of including the discretization parameter in comparison to either the Poisson model (i.e., $p = 1$; difference in Akaike Information Criterion, $\Delta\text{AIC} = -0.61 \pm 0.49$; difference in Bayesian Information Criterion, $\Delta\text{BIC} = +4.2 \pm 0.46$; negative values favor the added parameter) or the continuous Gamma model (i.e., $p \rightarrow 0$; $\Delta\text{AIC} = -3.3 \pm 0.93$; $\Delta\text{BIC} = +1.5 \pm 0.89$). Overall, these results do not allow strong conclusions to be drawn regarding

the discreteness of sampling, which has relatively little effect on error distributions (Fig 4, Insets) or the quality of fits.

Probabilistic Item Limits

In the fixed sampling model, at higher set sizes, a meaningful proportion of estimates are random “guesses” based on no samples (Fig. 5 A and B). Specifically, if an estimate was generated for every item in the memory array, then, as set size N increased, the number of estimates based on at least one sample, $S_{\omega>0}$, would reach a maximum at the fixed total number of samples,

$$\lim_{N \rightarrow \infty} S_{\omega>0} = K. \quad [11]$$

This is a trivial consequence of sharing out a fixed number of samples evenly between items.

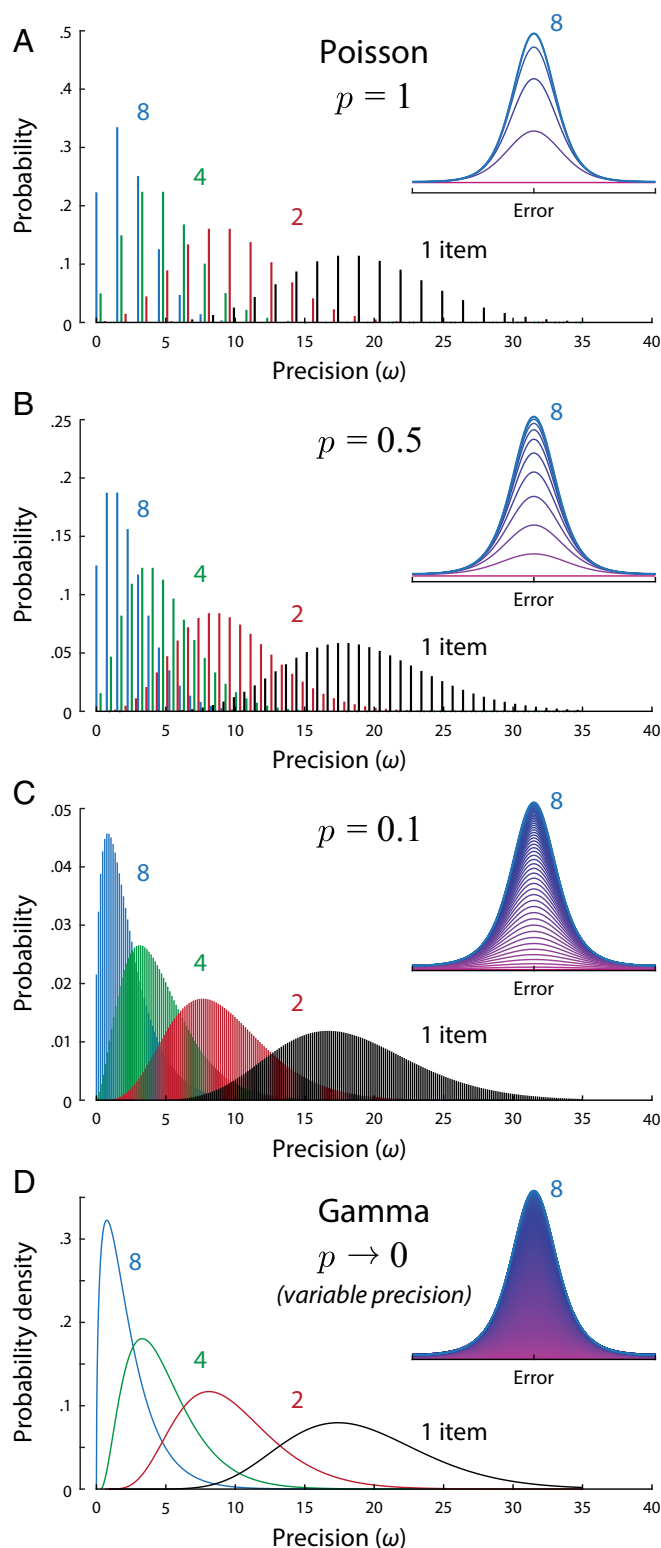


Fig. 4. Precision distributions in the generalized stochastic model, for different levels of discretization, p , and different set sizes. (Insets) Construction of the corresponding distributions of response error (for set size eight), with thin lines showing normal distributions with different precisions incrementally accumulated in ascending order (magenta to blue). (A) Example of discrete Poisson-distributed precision values ($p = 1$). For typical ML parameters, estimates are based on a small mean number of samples (here, $\gamma = 12$), each of moderate precision ($\omega_1 = 1.5$). (B and C) With decreasing discretization ($p < 1$), estimates are based on larger mean numbers of samples, and discrete precision values are more finely spaced. (D) In the limit as dis-

cretization falls to zero, the mean number of samples becomes infinite, and the distribution over precision approaches a continuous Gamma distribution. The ratio of variance to mean precision (Fano factor) is fixed (at $\omega_1 = 1.5$) across all set sizes and levels of discretization.

$$\Pr(\omega = 0) = \exp\left(-\frac{\gamma}{N}\right), \quad [12]$$

and the number of nonrandom retrievals in a set of N items has a binomial distribution,

$$S_{\omega>0} \sim \text{Bin}\left(N, 1 - \exp\left(-\frac{\gamma}{N}\right)\right). \quad [13]$$

As set size increases, the mean number of estimates based on at least one sample reaches a maximum at the expected total number of samples (Fig. 5 C and D). However, unlike the fixed sampling model, this limit is probabilistic, and (as illustrated in Fig. 5C) the actual number will vary from one set of memory items to the next, converging to a Poisson distribution for large N ,

$$\lim_{N \rightarrow \infty} S_{\omega>0} \sim \text{Poisson}(\gamma). \quad [14]$$

As we increase the expected number of samples by reducing the discretization, p , the probability of zero samples falls to zero: $\Pr(\omega = 0) = p^{\gamma/(N(1-p))}$. However, if we choose a precision threshold that is less than or equal to the base precision ω_1 , it can be shown that the mean number of items with above-threshold precision converges to a finite positive number at large set sizes (SI Appendix). This saturation is illustrated for different levels of discretization and various precision thresholds in Fig. 5 E–H.

Item limits or “magic numbers” (2, 24, 32) are usually considered synonymous with slot-based accounts, occurring when some items must go unrepresented because other items have filled the available capacity. The present results show that a probabilistic item limit, that is, an upper limit on the average number of items successfully retrieved that is not exceeded at any set size, can arise even when the probability of success for one item is independent of each other item. This holds true in the Poisson sampling model if we define success as obtaining one or more samples, but also more generally, even in a continuous model, if we define success as exceeding a threshold level of precision in estimation. Note, however, that the item limit does not, in general, have a simple relationship to the underlying number of samples. For example, the probabilistic limit at approximately five items in Fig. 5E is obtained from a model with a mean of 24 samples.

Toward Biophysically Realistic Models

The idealized description of population coding on which we based the stochastic sampling model overlooks a number of important considerations in order to reveal relationships between cognitive and neural-level accounts of VWM. For instance, the statistics of spike counts in the neural system often deviate from the Poisson distribution assumed in the original population coding model in that they are “overdispersed” (i.e., Fano factor of >1). Such an overdispersion in the sample count also occurs in the generalized stochastic model as the discretization p decreases (in order for the Fano factor of the precision distribution to remain constant, the Fano factor of the sample count has to increase). Spike counts in visual cortical neurons typically show a Fano factor in the range 1.5 to 3 (e.g., ref. 33),

cretization falls to zero, the mean number of samples becomes infinite, and the distribution over precision approaches a continuous Gamma distribution. The ratio of variance to mean precision (Fano factor) is fixed (at $\omega_1 = 1.5$) across all set sizes and levels of discretization.

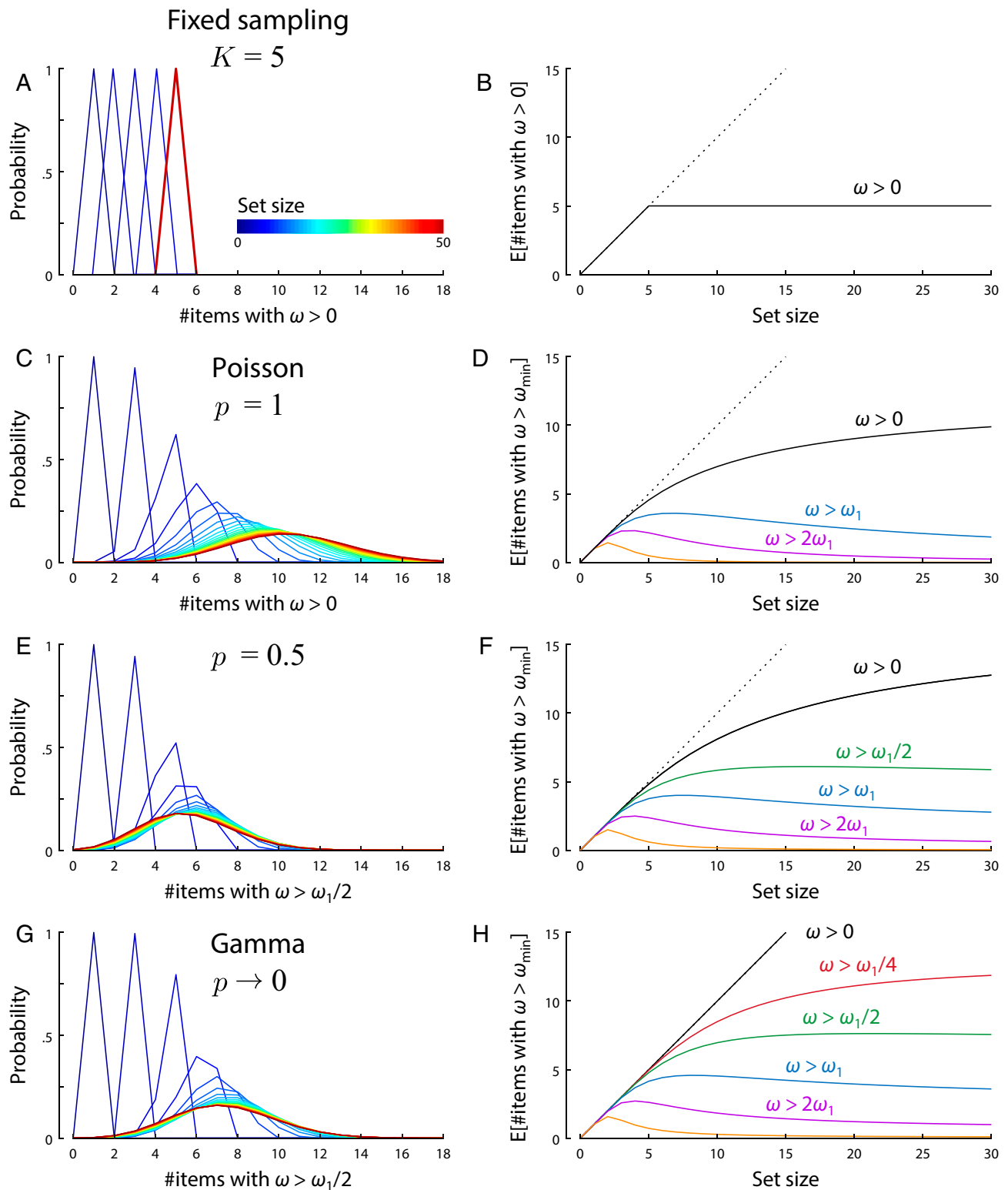


Fig. 5. Item limits in sampling models. For each model, A, C, E, and G show how the probability distribution of the number of items recovered with greater than zero precision (A and C; greater than a fixed threshold for E and G) changes with set size (color coded, increasing blue to red; discrete probability distributions are depicted as line plots for better visualization). B, D, F, and H plot the mean number of items with above-threshold precision as a function of set size for different threshold values. Thresholds are defined as a proportion of the base precision ω_1 . (A and B) In the fixed sampling model, the number of items with nonzero precision increases with set size, then plateaus when the number of items equals the number of samples. (C and D) The stochastic sampling model with Poisson variability also has a limit on the number of items with nonzero precision, although this limit is probabilistic and emerges asymptotically (converging to the distribution shown by the red curve in C for large set sizes, corresponding to the mean number of items plotted as black curve in D). (E and F) Stochastic models with lower discretization display similar probabilistic item limits for precision exceeding a fixed threshold, but with the expected number of items saturating at different values depending on threshold (different colors in F). (G and H) This property also extends to models with continuous precision distributions.

corresponding, in our model, to discretization p in the range 0.33 to 0.75.

In real neural populations, there is also considerable variability between individual cells' tuning curves (34). Due to this heterogeneity, neurons differ in the amount of information each spike provides about a stimulus. From a sampling perspective, this means that estimates are based on samples that vary in precision, and this has the effect of "smoothing out" the discrete distribution of precision values predicted by the stochastic model (*SI Appendix*, Fig. S6). This has similar consequences for estimation error as decreasing p in the generalized model. We fit the single report data with a variant of the population model with random variability in the neurons' tuning curves (affecting baseline activity, gain, and tuning curve width, as well as adding heterogeneity in the coverage of the feature space by neural tuning curves; *SI Appendix*), scaled by a global heterogeneity parameter ν . Incorporating biologically realistic heterogeneity into the population model improved fits to data ($\Delta\text{AIC} = 8.3 \pm 1.8$, $\Delta\text{BIC} = 3.4 \pm 1.7$ compared to the stochastic sampling model). The mean heterogeneity parameter in the ML fits was $\nu = 0.66 \pm 0.08$, where $\nu = 0$ means no heterogeneity, and $\nu = 1$ was approximately matched to heterogeneity of orientation-selective neurons in recordings from primary visual cortex (34).

Finally, spikes in real neural populations are not independent events as assumed by the sampling interpretation, but rather are correlated within and between neurons. This will tend to result in deviations from the simple additivity assumed by sampling. An implementation of short-range pair-wise correlations in the heterogeneous population model (see *SI Appendix* for details) greatly increased the numbers of decoded spikes required to reproduce behavioral data (on average, 168 times higher), without changing quality of fit ($\Delta\text{AIC/BIC} = 0.045 \pm 0.28$). We note, however, that the exact consequences of spike correlations for decoding depend on details of correlation structure that are difficult to measure experimentally (35–37), and suboptimal inference (in the form of a mismatched decoder) may play an important part (38).

Discussion

Taking, as a starting point, a mathematical idealization of the way neural populations encode information, we have shown that retrieval of a visual feature from working memory can be described as estimation based on a stochastically varying number of noisy samples. Two other influential models of VWM can be reconceptualized in the same framework: The slots+averaging model, because it modified the original slot model to allow multiple representations with independent noise, is directly equivalent to a sampling model with a fixed number of samples (13). And the variable precision model (8, 9) constitutes the continuous limit of a sampling model as samples are made less precise and more numerous, while maintaining the fixed proportionality between the variance and mean of precision in the decoded estimate.

Formulating all three models in the same mathematical framework (a formal "unification") allowed us to pinpoint specific differences between them. We determined the effect of these differences on the models' ability to account for human behavior by fitting multiple variants of the sampling model to a large database of delayed reproduction tasks. We found that stochasticity both in the total number of samples and in their distribution among items has a major impact on the quality of fit, with the best fits obtained if the number of samples is drawn randomly and independently for each item in each trial. Note that this form of stochasticity is poorly captured by the concept of memory "slots," because of the implication that a slot occupied by one item leaves fewer slots available for other items—this would predict dependencies between items in whole report that were not supported experimentally.

On the other hand, contrary to the assumptions of continuous resource models, we did find limited support for discreteness of memory representations (3). The fully continuous model with Gamma-distributed precision proposed in previous studies provided fits to data that were, overall, a little worse than the discrete Poisson model, in both single- and whole-report tasks (*SI Appendix*). When we attempted to fit discretization as a free parameter, however, we found that ML estimates varied widely between participants, and many were best fit by continuous or near-continuous versions of the generalized stochastic sampling model. So, while discreteness in memory representations is plausible—even inevitable if based on discrete spiking activity—recall errors do not provide strong evidence for any one particular level of discreteness, or, as a corollary, any particular mean number of samples.

Our findings further highlight the need to distinguish between two concepts that have previously been elided: discreteness in representation and discreteness in allocation. In the stochastic sampling model, the resource underlying capacity limits in VWM is equated with the mean number of samples (or the mean spike rate in the population coding interpretation), which can be distributed among items in a continuous fashion, even though the consequent number of samples obtained by each item is a discrete integer. This view on memory resources was strongly motivated by studies showing that prioritized items can be represented more precisely in VWM, at the cost of decreased precision for other items (7, 39–42). The stochastic sampling model can account for such findings through an uneven distribution of resources among memory items, corresponding to a higher mean number of samples for some items at the cost of a lower mean for other items. The actual number of samples available on an individual memory retrieval varies randomly about the item's mean. In the neural population model, this mechanism has previously been shown to successfully reproduce data from tasks in which one item is cued as the likely target (10).

While the stochastic sampling model is based on a highly idealized implementation of population coding, it nevertheless provides a link to a concrete neural mechanism that could form the basis of VWM performance. We have shown that adapting this model to achieve a higher degree of biophysical realism—by introducing heterogeneity in neural tuning curves and correlated spiking activity—improved the quality of fit to behavioral data. It has recently been shown that more neurally realistic population coding models preserve the key characteristics of the idealized model, and that signatures of neural tuning may even be visible in behavioral data (43).

Our results also provide a link between models of working memory used in the psychological literature and more biophysically detailed neural models such as continuous attractor networks (44–46), whose greater complexity typically precludes quantitative fits to behavioral data. These models are, likewise, based on principles of population coding and emphasize the role of neural noise in explaining variability in working memory performance. They are capable of producing probabilistic item limits similar to those described here, but it remains unresolved how these models could account for the graded variations in recall fidelity that we have found to be essential for capturing human behavioral performance.

In keeping with most previous work on VWM limits, we have not here attempted to reproduce the variations in bias and precision that are observed for different feature values, exemplified by the finding that cardinal orientations can be reproduced with greater precision than obliques. However, previous work has shown that these effects can be simply and elegantly captured within the population coding framework via the principle of efficient coding (47–49). The idea is that neural tuning functions are adapted to the stimulus statistics of the natural environment in such a way as to maximally convey information in that

environment (effectively by distributing neural resources preferentially to the most frequently occurring stimulus features). Although it should be possible to formulate this model as a modification of stochastic sampling, without reference to neural populations, it seems that the modifications required would not have a natural explanation within the sampling framework. These observations, and the results of incorporating heterogeneity described above, illustrate the value of connecting abstract cognitive models to neural theory.

We also did not address here the question of how individual features of a visual stimulus are bound together, which forms another point of contention in the debate on the format of VWM representations. In the model fits, we allowed failures of binding memory in the form of swap errors to occur with a fixed rate, although taking into account similarity of items with respect to the cue feature is likely to improve model fits (50, 51). While the discrete memory representations in slot models have traditionally been associated with a strongly object-based view (1), the sampling framework is agnostic as to whether objects or features are the units of VWM storage. Both views are compatible with the population coding interpretation, depending on whether the neurons in question are sensitive to a single feature (52) or a conjunction of features (51, 53).

A recent proposal that VWM errors can be explained in terms of a perceptual rescaling of stimulus space can also be expressed in terms of population coding, with some minor differences from the version presented here (see ref. 54 for details and discussion). In particular, the idea of retrieval based on normally distributed “memory-match” signals maps exactly onto an idealized population code with continuous-valued activity and constant Gaussian noise (51, 55). This predicts a continuous distribution over precision, not dissimilar to the Gamma distribution. Continuousness in representation does not appear a necessary component of this account, however, and it should be possible to reformulate it with arbitrary levels of discreteness, as in our generalized stochastic model.

There are other models of working memory that address capacity limits without explicitly postulating a limited memory resource (56). Some accounts stress the importance of memory decay over time, and active rehearsal to counteract this decay (57, 58). These theories do not have a clear analogue in the sampling framework, although effects of retention time have been incorporated into the neural population model (28). Other accounts have sought to explain capacity limits by interference between different memorized items (59). While the sampling framework does not explicitly address interference, the effect of normalization could be described as a form of nonspecific interference between items. A model of feature binding based on the neural population model shows some notable congruencies with

an interference account of VWM, and both models make similar predictions regarding swap errors (50, 51). Further research will be needed to determine the exact relationship between these models.

Taken together, our results reveal a surprising convergence between prominent models of VWM. Despite the fact that these competing models were independently motivated by different behavioral and neural findings, they can be expressed within the shared formal framework of sampling, which reveals specific distinguishing factors as well as shared general principles. This convergence gives cause for confidence that the stochastic sampling model captures key characteristics of VWM and will provide a solid foundation for future research.

Materials and Methods

We fit computational models of VWM to behavioral data from a large dataset of delayed estimation experiments. The dataset included 15 individual single-report experiments (*SI Appendix, Table S1*; see *SI Appendix, Supporting Information Text* for inclusion criteria), as well as four whole-report experiments (*SI Appendix, Table S2*). Each model defines a parameterized distribution of response probabilities given the true feature values of the target and nontarget items in each trial (*SI Appendix*). For fits to whole-report data, we determined the probabilities of obtaining the given combination of responses within a single trial, taking into account the correlations of recall precision between different items within a trial predicted by each model.

We obtained an ML fit of each subject's data for each model. The stochastic sampling, fixed sampling, and continuous sampling (Gamma) model, as well as the fixed-random and stochastic-even variants, each have three free parameters (including one parameter for the probability of swap errors). We fit these to both the single-report and whole-report data using the Nelder-Mead simplex algorithm (see *SI Appendix* for details). The generalized stochastic model and the neural population model with heterogeneous tuning curves have four free parameters each. For these models, we employed a grid search to obtain fits only of the single-report data (fitting them to whole-report data was not computationally feasible). We further evaluated model variants employing a more accurate method for ML decoding for circular feature spaces (rather than the Gaussian approximation used for fits reported in *Response Errors Discriminate between Models*), models without swap errors, models with an additional free parameter for the power law in set size effects, and models with a temporal decay of memory precision over varying response delays in the whole-report experiments (*SI Appendix*).

Data Availability. Data and code associated with this study are publicly available in Open Science Framework (60).

ACKNOWLEDGMENTS. We thank Ronald van den Berg, Wei Ji Ma, Máté Lengyel, and Masud Husain for helpful conversations, Zakhar Kabluchko and Martin Bays for statistical advice, and all of the researchers who publicly shared data that facilitated this study. We used resources provided by the Cambridge Service for Data Driven Discovery operated by the University of Cambridge Research Computing Service. This research was supported by the Wellcome Trust (Grant 106926).

1. S. J. Luck, E. K. Vogel, The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279–281 (1997).
2. N. Cowan, The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav. Brain Sci.* **24**, 87–114 (2001).
3. W. Zhang, S. J. Luck, Discrete fixed-resolution representations in visual working memory. *Nature* **453**, 233–235 (2008).
4. P. M. Bays, R. F. G. Catalao, M. Husain, The precision of visual working memory is set by allocation of a shared resource. *J. Vis.* **9**, 7–7 (2009).
5. W. J. Ma, M. Husain, P. M. Bays, Changing concepts of working memory. *Nat. Neurosci.* **17**, 347–356 (2014).
6. P. Wilken, W. J. Ma, A detection theory account of change detection. *J. Vis.* **4**, 11 (2004).
7. P. M. Bays, M. Husain, Dynamic shifts of limited working memory resources in human vision. *Science* **321**, 851–854 (2008).
8. D. Fougner, J. W. Suchow, G. A. Alvarez, Variability in the quality of visual working memory. *Nat. Commun.* **3**, 1229 (2012).
9. R. van den Berg, H. Shin, W.-C. Chou, R. George, W. J. Ma, Variability in encoding precision accounts for visual short-term memory limitations. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8780–8785 (2012).
10. P. M. Bays, Noise in neural populations accounts for errors in working memory. *J. Neurosci.* **34**, 3632–3645 (2014).
11. A. Pouget, P. Dayan, R. Zemel, Information processing with population codes. *Nat. Rev. Neurosci.* **1**, 125–132 (2000).
12. M. L. Shaw, Identifying attentional and decision-making components in information processing. *Atten. Perform.* **8**, 277–295 (1980).
13. J. Palmer, Attentional limits on the perception and memory of visual information. *J. Exp. Psychol. Hum. Percept. Perform.* **16**, 332–350 (1990).
14. D. K. Sewell, S. D. Lilburn, P. L. Smith, An information capacity limitation of visual short-term memory. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 2214–2242 (2014).
15. A.-M. Bonnel, J. Miller, Attentional effects on concurrent psychophysical discriminations: Investigations of a sample-size model. *Percept. Psychophys.* **55**, 162–179 (1994).
16. E. Vul, N. Goodman, T. L. Griffiths, J. B. Tenenbaum, One and done? Optimal decisions from very few samples. *Cognit. Sci.* **38**, 599–637 (2014).
17. G. Orbán, P. Berkes, J. Fiser, M. Lengyel, Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* **92**, 530–543 (2016).
18. M. N. Shadlen, D. Shohamy, Decision making and sequential sampling from memory. *Neuron* **90**, 927–939 (2016).
19. A. P. Georgopoulos, A. B. Schwartz, R. E. Kettner, Neuronal population coding of movement direction. *Science* **233**, 1416–1419 (1986).
20. M. Carandini, D. J. Heeger, Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).

21. G. A. Alvarez, P. Cavanagh, 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).
22. K. Oberauer, S. Eichenberger, Visual working memory declines when more features must be remembered for each object. *Mem. Cognit.* **41**, 1212–1227 (2013).
23. P. L. Smith, The Poisson shot noise model of visual short-term memory and choice response time: Normalized coding by neural population size. *J. Math. Psychol.* **66**, 41–52 (2015).
24. K. C. S. Adam, E. K. Vogel, E. Awh, Clear evidence for item limits in visual working memory. *Cognit. Psychol.* **97**, 79–97 (2017).
25. R. Van den Berg, A. H. Yoo, W. J. Ma, Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychol. Rev.* **124**, 197–214 (2017).
26. P. M. Bays, A signature of neural coding at human perceptual limits. *J. Vis.* **16**, 4 (2016).
27. R. van den Berg, E. Awh, W. J. Ma, Factorial comparison of working memory models. *Psychol. Rev.* **121**, 124–149 (2014).
28. S. Schneegans, P. M. Bays, Drift in neural population activity causes working memory to deteriorate over time. *J. Neurosci.* **38**, 4859–4869 (2018).
29. E. Zohary, M. N. Shadlen, W. T. Newsome, Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143 (1994).
30. K. H. Britten, M. N. Shadlen, W. T. Newsome, J. A. Movshon, The analysis of visual motion: A comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).
31. R. L. T. Goris, J. A. Movshon, E. P. Simoncelli, Partitioning neuronal variability. *Nat. Neurosci.* **17**, 858–865 (2014).
32. G. A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **101**, 343–352 (1994).
33. R. Vogels, W. Spilleers, G. A. Orban, The response variability of striate cortical neurons in the behaving monkey. *Exp. Brain Res.* **77**, 432–436 (1989).
34. A. S. Ecker *et al.*, Decorrelated neuronal firing in cortical microcircuits. *Science* **327**, 584–587 (2010).
35. B. B. Averbach, P. E. Latham, A. Pouget, Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
36. A. S. Ecker, P. Berens, A. S. Tolias, M. Bethge, The effect of noise correlations in populations of diversely tuned neurons. *J. Neurosci.* **31**, 14272–14283 (2011).
37. R. Moreno-Bote *et al.*, Information-limiting correlations. *Nat. Neurosci.* **17**, 1410–1417 (2014).
38. J. M. Beck, W. J. Ma, X. Pitkow, P. E. Latham, A. Pouget, Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).
39. P. M. Bays, N. Gorgoraptis, N. Wee, L. Marshall, M. Husain, Temporal dynamics of encoding, storage, and reallocation of visual working memory. *J. Vis.* **11**, 6–6 (2011).
40. A. H. Yoo, Z. Klyszejko, C. E. Curtis, W. J. Ma, Strategic allocation of working memory resource. *Sci. Rep.* **8**, 16162 (2018).
41. S. M. Emrich, H. A. Lockhart, N. Al-Aidroos, Attention mediates the flexible allocation of visual working memory resources. *J. Exp. Psychol. Hum. Percept. Perform.* **43**, 1454–1465 (2017).
42. Z. Klyszejko, M. Rahmati, C. E. Curtis, Attentional priority determines working memory precision. *Vis. Res.* **105**, 70–76 (2014).
43. R. Taylor, P. M. Bays, Theory of neural coding predicts an upper bound on estimates of memory variability. *Psychol. Rev.*, 10.1037/rev0000189 (2020).
44. J. S. Johnson, J. P. Spencer, S. J. Luck, G. Schöner, A dynamic neural field model of visual working memory and change detection. *Psychol. Sci.* **20**, 568–577 (2009).
45. Z. Wei, X.-J. Wang, D.-H. Wang, From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. *J. Neurosci.* **32**, 11228–11240 (2012).
46. D. Standage, M. Paré, Slot-like capacity and resource-like coding in a neural model of multiple-item working memory. *J. Neurophysiol.* **120**, 1945–1961 (2018).
47. R. Taylor, P. M. Bays, Efficient coding in visual working memory accounts for stimulus-specific variations in recall. *J. Neurosci.* **38**, 7132–7142 (2018).
48. X.-X. Wei, A. A. Stocker, A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nat. Neurosci.* **18**, 1509–1517 (2015).
49. D. Ganguli, E. P. Simoncelli, Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Comput.* **26**, 2103–2134 (2014).
50. K. Oberauer, H.-Y. Lin, An interference model of visual working memory. *Psychol. Rev.* **124**, 21–59 (2017).
51. S. Schneegans, P. M. Bays, Neural architecture for feature binding in visual working memory. *J. Neurosci.* **37**, 3913–3925 (2017).
52. F. Bouchacourt, T. J. Buschman, A flexible model of working memory. *Neuron* **103**, 147–160 (2019).
53. L. Matthey, P. M. Bays, P. Dayan, A probabilistic palimpsest model of visual short-term memory. *PLoS Comput. Biol.* **11**, e1004003 (2015).
54. P. M. Bays, Correspondence between population coding and psychophysical scaling models of working memory. *bioRxiv*:10.1101/699884 (11 July 2019).
55. H. Sompolinsky, H. Yoon, K. Kang, M. Shamir, Population coding in neuronal systems with correlated noise. *Phys. Rev.* **64**, 051904 (2001).
56. J. Jonides *et al.*, The mind and brain of short-term memory. *Annu. Rev. Psychol.* **59**, 193–224 (2008).
57. A. D. Baddeley, *Working Memory* (Oxford University Press, Oxford, United Kingdom, 1986).
58. P. Barrouillet, V. Camos, As time goes by: Temporal constraints in working memory. *Curr. Dir. Psychol. Sci.* **21**, 413–419 (2012).
59. K. Oberauer, S. Lewandowsky, S. Farrell, C. Jarrold, M. Greaves, Modeling working memory: An interference model of complex span. *Psychon. Bull. Rev.* **19**, 779–819 (2012).
60. S. Schneegans, R. Taylor, P. M. Bays, Stochastic sampling provides a unifying account of visual working memory limits. *Open Science Framework*. <https://osf.io/buxp9/>. Deposited 28 July 2020.



Supplementary Information for

Stochastic sampling provides a unifying account of visual working memory limits

Sebastian Schneegans, Robert Taylor, and Paul M. Bays

Corresponding Author: Paul M. Bays

E-mail: pmb20@cam.ac.uk

This PDF file includes:

Supplementary text
Figs. S1 to S7
Tables S1 to S5
SI References

Supporting Information Text

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 11 experiments with over 130,000 trials in total. We included available data from published delayed reproduction experiments (either single-report or whole-report tasks; Fig. S2) 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 are 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 (1), 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 (2). In the whole-report tasks, items were always cued or selected via their location.

Models

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 $\theta = (\theta_1, \dots, \theta_N)$. For classical continuous report tasks with a single report, we denote the reported feature value with $\hat{\theta}$. For whole-report tasks, the sequence of reported feature values is $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_N)$.

Each model defines a response probability distribution $p(\hat{\theta}|\theta)$ 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 [Excluding swap errors](#)). We assume that response distributions around the selected feature value are identical for target and non-target features, and that the probability of reporting a non-target feature increases linearly with set size, with each non-target item having an equal probability p_{NT} of being used as the basis for response generation (these assumptions are of course not tenable for very large N , but suffice for studies in our dataset, with $N \leq 8$). For the response $\hat{\theta}_i$ corresponding to a cued item θ_i , we then obtain the probability distribution

$$p(\hat{\theta}_i|\theta) = p_T p(\hat{\theta}_i|\theta_i) + p_{NT} \sum_{j \in \{1, \dots, N\}, j \neq i} p(\hat{\theta}_i|\theta_j), \quad [1]$$

where $p_T = 1 - (N - 1)p_{NT}$ 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.

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. It can be derived as an idealization of neural population coding. 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,

$$p(\hat{\theta}|\theta) = \sum_{k=0}^{\infty} \text{Pr}_{\text{Poisson}}\left(k; \frac{\gamma}{N}\right) \phi_{\circ}(\hat{\theta}; \theta, \kappa(k\omega_1)). \quad [2]$$

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

$$\text{Pr}_{\text{Poisson}}(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad [3]$$

and ϕ_{\circ} is the von Mises distribution,

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

The term $\kappa(\omega)$ is the concentration parameter that yields a von Mises distribution with precision ω . With precision expressed as Fisher Information (3), 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 consider sample counts k within a range that covers the cumulative probabilities of the Poisson distribution from 10^{-5} to $1 - 10^{-5}$.

We note that the linear scaling of precision with the number of samples that we have assumed here is only an approximation of the exact distribution of ML estimates in circular space. We implemented the exact method (4) as a variant for all sampling

models, and describe results of a comparison between the different variants in Section [Full maximum likelihood decoding for circular feature spaces](#).

In the whole-report task with random response 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

$$p(\hat{\theta}|\theta) = \prod_{i=1}^N p(\hat{\theta}_i|\theta), \quad [5]$$

where $p(\hat{\theta}_i|\theta)$ is the probability distribution including swap errors as defined in Eq. 1.

In the free response order condition, responses are ordered by precision, which is directly proportional to 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 $\lambda = (\lambda_1, \dots, \lambda_N)$, $\lambda_i \geq \lambda_j \forall i < j$. The probability that each such sequence of sample counts will be generated by the model is

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

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

$$p(\hat{\theta}|\theta) = \sum_{\lambda} \Pr(\lambda) \prod_{i=1}^N \left(p_{\text{T}} \phi_{\circ}(\hat{\theta}_i; \theta_i, \kappa(\lambda_i \omega_1)) + p_{\text{NT}} \sum_{j \in \{1, \dots, N\}, j \neq i} \phi_{\circ}(\hat{\theta}_i; \theta_j, \kappa(\lambda_j \omega_1)) \right). \quad [7]$$

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

Fixed sampling model. The fixed sampling model assumes that a fixed number K of samples, each with a fixed precision ω_1 , is distributed as evenly as possible among the memory items in each trial. This model is mathematically equivalent to the slots+averaging model (5). We note that the slot concept is commonly associated with an object- rather than feature-based view of working memory storage, but the sampling interpretation is agnostic with respect to this distinction.

The response probability distribution for the fixed sampling model is given by

$$p(\hat{\theta}|\theta) = \frac{K \bmod N}{N} \phi_{\circ}(\hat{\theta}; \theta, \kappa(\lceil \frac{K}{N} \rceil \omega_1)) + \left(1 - \frac{K \bmod N}{N}\right) \phi_{\circ}(\hat{\theta}; \theta, \kappa(\lfloor \frac{K}{N} \rfloor \omega_1)). \quad [8]$$

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

$$p(\hat{\theta}|\theta) = \prod_{i=1}^N p(\hat{\theta}_i|\theta) \quad [9]$$

with

$$p(\hat{\theta}_i|\theta) = \begin{cases} \phi_{\circ}(\hat{\theta}_i; \theta, \kappa(\lceil \frac{K}{N} \rceil \omega_1)), & \text{if } i \leq K \bmod N \\ \phi_{\circ}(\hat{\theta}_i; \theta, \kappa(\lfloor \frac{K}{N} \rfloor \omega_1)), & \text{otherwise.} \end{cases} \quad [10]$$

In the random response order condition, the fixed sampling model predicts negative correlations between response errors in a single trial (because having more samples for one item means fewer samples are available for others). We determine all possible sequences of sample counts $\lambda = (\lambda_1, \dots, \lambda_N)$, $(\lambda_i = \lfloor \frac{K}{N} \rfloor \vee \lambda_i = \lceil \frac{K}{N} \rceil) \forall i, \sum_{i=1}^N \lambda_i = K$. Each such sequence will occur with equal probability $\Pr(\lambda) = \left[\binom{N}{K \bmod N} \right]^{-1}$. The response probability $p(\hat{\theta}|\theta)$ can then be expressed in the same way as in Eq. 7.

We tested the fixed sampling model (as well as the variant with random allocation described below) for values of K in the range (0, 25), going substantially beyond the typical estimates of three or four memory slots to ensure that model comparison results were not biased by a limited parameter range.

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,

$$\Pr_{\text{Bin}}\left(k; K, \frac{1}{N}\right) = \binom{K}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{K-k}, \quad [11]$$

and the response probability distribution is

$$p(\hat{\theta}|\theta) = \sum_{k=0}^K \text{PrBin}\left(k; K, \frac{1}{N}\right) \phi_{\circ}(\hat{\theta}; \theta, \kappa(k\omega_1)). \quad [12]$$

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

$$\text{Pr}(\lambda) = \frac{1}{N^K} \prod_{i=1}^N \binom{K - \sum_{j=1}^{i-1} \lambda_j}{\lambda_i} \prod_{k=0}^{\max(\lambda)} \binom{N - \sum_{j=0}^{k-1} n_j(\lambda)}{n_k(\lambda)}. \quad [13]$$

For the random response order condition, the set of possible sample count sequences is $\{\lambda | \sum_{i=1}^N \lambda_i = K\}$. The probability of each sequence is given by

$$\text{Pr}(\lambda) = \frac{1}{N^K} \prod_{i=1}^N \binom{K - \sum_{j=1}^{i-1} \lambda_j}{\lambda_i}. \quad [14]$$

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

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 a weighted sum of probabilities from the fixed sampling model with different numbers of samples k ,

$$p(\hat{\theta}|\theta) = \sum_{k=0}^{\infty} \text{PrPoisson}(k; \gamma) p_{\text{fs}}(\hat{\theta}|\theta; k). \quad [15]$$

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.

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

$$p(\hat{\theta}|\theta) = \sum_{k=0}^{\infty} \text{PrNegBin}\left(k; \frac{\gamma}{(1-p)N}, p\right) \phi_{\circ}(\hat{\theta}; \theta, \kappa(k\omega_1 p)) \quad [16]$$

with

$$\text{PrNegBin}(k; r, p) = \frac{\Gamma(k+r)}{k! \Gamma(r)} p^r (1-p)^k \quad [17]$$

for $0 < p < 1$. For fitting the model to data, we only compute the sum over the most likely sample counts k up to a cumulative probability of $1 - 10^{-4}$. 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.

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 $p \rightarrow 0$ (see Section [Gamma distribution as limiting case](#)), and has independently been proposed in two previous studies (3, 6). In the formulation of van den Berg et al., the precision follows a Gamma distribution with mean \bar{J}_1/N^α and scale parameter τ , which is identical to the model described here for $\bar{J}_1 = \gamma\omega_1$, $\tau = \omega_1$, and $\alpha = 1$ (see Section [Power law for set size effects](#) for model comparisons with α as free parameter).

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

$$p(\hat{\theta}|\theta) = \int_{\omega=0}^{\infty} p_{\text{Gamma}}\left(\omega; \frac{\gamma}{N}, \omega_1\right) \phi_{\circ}(\hat{\theta}; \theta, \kappa(\omega)) d\omega, \quad [18]$$

with

$$p_{\text{Gamma}}(\omega, k, \theta) = \frac{1}{\Gamma(k)\theta^k} \omega^{k-1} e^{-\frac{\omega}{\theta}}, \quad [19]$$

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

$$p(\hat{\theta}|\theta) = \prod_{i=1}^N p(\hat{\theta}_i|\theta). \quad [20]$$

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

$$p(\hat{\theta}|\theta) = \int_{\omega_1=0}^{\infty} \cdots \int_{\omega_N=0}^{\infty} p(\omega) \prod_{i=1}^N \left(p_T \phi_o(\hat{\theta}_i; \theta_i, \kappa(\omega_i)) + p_{NT} \sum_{j \in \{1, \dots, N\}, j \neq i} \phi_o(\hat{\theta}_i; \theta_j, \kappa(\omega_i)) \right) d\omega_1 \dots d\omega_N, \quad [21]$$

where $\omega = (\omega_1, \dots, \omega_N)$ is the sequence of ordered precision values for the responses, and $p(\omega)$ 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 bin indices $\mathbf{b} = (b_1, \dots, b_N)$ as

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

Here, $n_k(\mathbf{b})$ denotes the number of entries in \mathbf{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 a weighted sum over all possible sequences \mathbf{b} ,

$$p(\hat{\theta}|\theta) \approx \sum_{\mathbf{b}} \Pr(\mathbf{b}) \prod_{i=1}^N \int_{\omega=\omega_{\text{low}}(b_i)}^{\omega_{\text{high}}(b_i)} p_{\text{Gamma}}\left(\omega; \frac{\gamma}{N}, \omega_1\right) \left(p_T \phi_o(\hat{\theta}_i; \theta_i, \kappa(\omega)) + p_{NT} \sum_{j \in \{1, \dots, N\}, j \neq i} \phi_o(\hat{\theta}_i; \theta_j, \kappa(\omega)) \right) d\omega. \quad [23]$$

Here, $\omega_{\text{low}}(b)$ and $\omega_{\text{high}}(b)$ 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. We note that with this binning approach, we still obtain precise estimates of the response error distribution at each ordinal position; only the correlations between them are affected by the approximation.

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 (for the origin of this model and a more complete investigation of the topic, see 7). 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, neurons' preferred values are selected at random from a uniform distribution. As in the original implementation of the neural population model (8), 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,

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

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 (8), 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 (9). 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:

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

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

$$\log \alpha_i \sim \mathcal{N}(\log(0.04\nu\beta_i), 2) \quad [27]$$

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

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:

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

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

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

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

$$\gamma = \bar{\mathcal{I}} / \check{\mathcal{I}}. \quad [31]$$

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,

$$r_i \sim \text{Poisson} \left(\frac{\gamma}{N} \check{f}_i(\theta) \right). \quad [32]$$

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

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

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 [Fitting procedure](#)), we randomly drew 100 sets of single-neuron parameters for $M = 1000$ neurons* from the distributions specified in Eqs. 25 – 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 ML estimates $\hat{\theta}$ as described above. The response error distribution is then approximated by a histogram over the decoding errors, $\hat{\theta} \ominus \theta$, averaged over all sets of single-neuron parameters (\ominus indicates subtraction on the circle). Mean ML parameters obtained for the single-report dataset were $\tilde{\kappa} = 1.53 \pm 0.15$, $\bar{\mathcal{I}} = 18.6 \pm 1.0$, and $\nu = 0.66 \pm 0.08$.

Unlike the discrete distribution predicted by the (homogeneous) stochastic sampling model (Fig. S6E), 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. S6D).

Example precision distributions from the Gamma (variable precision; 3, 6) model are shown in Fig. S6F. 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.

We also fit a further variant of the heterogeneous model incorporating short-range correlations of the form

$$c_{ij} = c_0 \exp(-|\varphi_i \ominus \varphi_j|), \quad [34]$$

with c_0 set to 0.2. We assumed that the decoder did not have knowledge of the correlation structure. Due to the significant computational challenge of simulating correlated Poisson activity, we used a Gaussian approximation to Poisson (see 7 for details), and compared model fits to an otherwise identical model without correlations (i.e. with c_0 set to 0).

*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.

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 $p_{NT} = 0$ 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 were determined by Monte Carlo 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. Results shown in Fig. 3E of the main manuscript are based on the grid search fits to allow fair comparison between the stochastic sampling (Poisson) model, generalized stochastic sampling model, and Gamma model.

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, $E[\omega]$, 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. The values for the discretization parameter, p , in the generalized stochastic sampling model was spaced logarithmically in the range $[10^{-4}, 1]$. 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,

$$\text{AIC} = 2k - 2\log(L), \quad [35]$$

and Bayesian information criterion (BIC) scores,

$$\text{BIC} = \log(n)k - 2\log(L). \quad [36]$$

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. S3A and E for single-report and whole-report data, respectively. ML fit values of free parameters are reported in Tables S3 and S4.

Additional model comparisons

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 (8) and (4). 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,

$$p(\hat{\theta}|\theta, m) = \int p(r|m, \kappa_1) \phi_\circ(\hat{\theta}; \theta, r\kappa_1) dr, \quad [37]$$

with

$$p(r|m, \kappa) = \frac{I_0(\kappa r)}{(I_0(\kappa))^m} r \psi_m(r). \quad [38]$$

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 $p(\hat{\theta}|\theta, m)$, 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. S3B and F), with only minimal

changes for the stochastic sampling model, and largest improvements 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.

Mean parameter values for the models with exact ML decoding differ substantially from those obtained using the approximation method. This is driven by a subset of participants for which the best model fits using the exact method are characterized by a large number of low-precision samples. Within this region of the parameter space, the exact method and the approximation deviate more strongly from each other (Fig. S7), with the exact method producing slightly broader distributions of response errors that can provide a better fit to some experimental result. We note that for the fixed sampling model, the variant with exact ML decoding produces fits with $K > 10$ for a substantial proportion of participants (40% in the single-report dataset, 18% in the whole-report dataset), which is in conflict with typical estimates of 3-4 memory slots.

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. S3C and G). This is consistent with previous findings that inclusion of swap errors improves model fit (1, 10).

Power law for set size effects. The variable precision model (3) 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. S3D and E), with the stochastic sampling model still providing better quality of fit for both single-report and whole-report datasets.

Random drift over memory delays. The most striking effect in the whole-report data (2) is the salient decrease in recall precision for successive responses when response order is freely chosen. A weaker decrease in precision, however, can also be observed in the random response order condition (Fig. S4B), where it cannot be explained by a strategy of reporting the best-remembered items first. All models described so far lack any mechanism that could capture the effects of increasing memory delay for later responses, or interference from intervening reports. Thus, they inevitably fail to fit this aspect of the behavioral data (the model fits for successive responses at each set size are identical in Fig. S4B).

We tested a mechanism for random drift of memorized feature values during memory delays that was previously described for the population coding model (11). The effect of this drift is described by convolving the response probability distribution of each model with a wrapped normal distribution centered on zero,

$$\tilde{p}(\hat{\theta}) = (p * f_{WM}(0, \sigma))(\hat{\theta}). \quad [39]$$

We assume that the parameter σ increases linearly from the first to the last response in each trial of the whole-report task. Moreover, we assume that the rate of random drift also scales linearly with set size, which has been found to provide the best fits to delayed reproduction data with varying memory delays (11). This yields

$$\sigma_i = iN\eta \quad [40]$$

for the i th response in a whole-report trial, where η is a new free parameter specifying the base drift rate.

When fitting a model with random drift to whole-report data, there is the concern that a high drift rate provides an alternative mechanism to produce the salient decrease in precision observed in the free response order condition. The much weaker decrease in precision in the random response order should constrain the drift rate parameter, but the two conditions were performed by separate groups of participants. To see whether a single drift rate parameter can account for response patterns across response order conditions, we pooled participant data to create a single super-participant each for the color report (Experiments 1a & 2a) and orientation report conditions (Experiments 1b & 2b). We then fit this data with one set of parameters each, varying only whether responses were ordered by memory precision or not to capture the different response order conditions (Table S5).

Pooling the data did not qualitatively change the pattern of quality-of-fit measures for the whole-report task (Fig. S3I), and the inclusion of a drift rate improved quality of fits in all models (Fig. S3J). The stochastic sampling model still provides the best fit for the data, and it closely reproduces the overall pattern of response distributions across response order conditions (Fig. S5), including the nearly uniform response distributions for the last reports at set size 6. We note that the slots+averaging model may be at a disadvantage in fitting this pooled data since it is forced to commit to a fixed number of slots, whereas the data may come from participants with different numbers of slots. For a more fair and thorough model comparison, it would be desirable to obtain data from individual participants performing both free and random response order conditions of the whole-report task.

Limiting cases of the negative binomial distribution

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

$$\frac{\omega}{\omega_1 p} \sim \text{NegBin}\left(\frac{\gamma}{N(1-p)}, p\right), \quad [41]$$

with parameters $\gamma > 0$, $\omega_1 > 0$, and $0 < p < 1$. The probability of a precision value ω that is an integer multiple of $\omega_1 p$ is defined as

$$\Pr(\omega) = \frac{\Gamma(\frac{\omega}{\omega_1 p} + \frac{\gamma}{N(1-p)})}{(\frac{\omega}{\omega_1 p})! \Gamma(\frac{\gamma}{N(1-p)})} p^{\frac{\gamma}{N(1-p)}} (1-p)^{\frac{\omega}{\omega_1 p}} \quad \text{for } \frac{\omega}{\omega_1 p} \in \mathbb{Z}^{\geq 0}. \quad [42]$$

All other precision values have zero probability.

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

Convergence result for large N. For a set size of N , the number of items with precision exceeding a threshold, defined as a multiple $m > 0$ of the base precision ω_1 , follows a binomial distribution,

$$S_{\omega > m\omega_1} \sim \text{Bin}(N, \Pr(\omega > m\omega_1)), \quad [43]$$

with mean

$$E[S_{\omega > m\omega_1}] = N \Pr(\omega > m\omega_1). \quad [44]$$

For the generalized stochastic sampling model, with precision following a scaled negative binomial distribution, we have

$$\begin{aligned} N \Pr(\omega > m\omega_1) &= N \sum_{k=\lfloor m/p+1 \rfloor}^{\infty} \frac{\Gamma(k + \frac{\gamma}{N(1-p)})}{k! \Gamma(\frac{\gamma}{N(1-p)})} p^{\frac{\gamma}{N(1-p)}} (1-p)^k \\ &= N \frac{p^{\frac{\gamma}{N(1-p)}}}{\Gamma(\frac{\gamma}{N(1-p)})} \sum_{k=\lfloor m/p+1 \rfloor}^{\infty} \frac{\Gamma(k + \frac{\gamma}{N(1-p)})}{k!} (1-p)^k. \end{aligned} \quad [45]$$

In the limit $N \rightarrow \infty$, we have

$$p^{\frac{\gamma}{N(1-p)}} \rightarrow 1, \quad [46]$$

$$\frac{\Gamma(k + \frac{\gamma}{N(1-p)})}{k!} \rightarrow \frac{\Gamma(k)}{k!} = \frac{1}{k} \quad \text{for } k \in \mathbb{Z}^{\geq 0}, \quad [47]$$

and

$$\frac{N}{\Gamma(\frac{\gamma}{N(1-p)})} \rightarrow \frac{\gamma}{1-p}. \quad [48]$$

We obtain

$$\lim_{N \rightarrow \infty} E[S_{\omega > m\omega_1}] = \gamma \sum_{k=\lfloor m/p+1 \rfloor}^{\infty} \frac{(1-p)^{k-1}}{k}. \quad [49]$$

Note that all terms of the sum are strictly positive. Convergence of the sum can be demonstrated by a ratio test. The ratio of successive terms is

$$R = \frac{(1-p)^k k}{(1-p)^{k-1} (k+1)} = (1-p) \frac{k}{k+1}, \quad [50]$$

so

$$\lim_{k \rightarrow \infty} |R| = 1-p < 1, \quad [51]$$

as $p > 0$. So the mean number of items with above-threshold precision converges to a finite positive number at large set sizes, for all model parameters, implementing a probabilistic form of “item limit”.

Poisson distribution as limiting case. To determine the limiting cases of the negative binomial distribution for $p \rightarrow 1$ and $p \rightarrow 0$, we use its characteristic function. For $X \sim \text{NegBin}(r, p)$, this is given by

$$\varphi_\omega(t) = E[e^{itX}] = \left(\frac{p}{1 - (1-p)e^{it}} \right)^r. \quad [52]$$

For the scaled negative binomial distribution as used in the generalized stochastic model (Eq. 41), the characteristic function is

$$\varphi_\omega(t) = \left(\frac{p}{1 - (1-p)e^{it\omega_1 p}} \right)^{\frac{\gamma}{N(1-p)}}. \quad [53]$$

We can write this as

$$\begin{aligned} \varphi_\omega(t) &= \exp \left[\log \left[\left(\frac{p}{1 - (1-p)e^{it\omega_1 p}} \right)^{\frac{\gamma}{N(1-p)}} \right] \right] \\ &= \exp \left[\frac{\gamma}{N} \left(\frac{\log(p)}{1-p} - \frac{\log(1 - (1-p)e^{it\omega_1 p})}{1-p} \right) \right] \end{aligned} \quad [54]$$

In the limit $p \rightarrow 1$, we have

$$\frac{\log(p)}{1-p} \rightarrow -1 \quad [55]$$

and for $q = 1 - p$ (and thus $q \rightarrow 0$)

$$\frac{\log(1 + aq)}{q} \rightarrow a. \quad [56]$$

The limiting case of the characteristic function for $p \rightarrow 1$ is therefore

$$\varphi_\omega(t) \rightarrow \exp\left[\frac{\gamma}{N}(e^{it\omega_1} - 1)\right], \quad [57]$$

which is the characteristic function of the scaled Poisson distribution, $\frac{\omega}{\omega_1} \sim \text{Poisson}(\frac{\gamma}{N})$.

Gamma distribution as limiting case. We again start with the characteristic function of the scaled negative binomial distribution (Eq. 53). We use the Taylor series for the exponential function,

$$e^x = 1 + x + \mathcal{O}(x^2) \quad \text{as } x \rightarrow 0, \quad [58]$$

where $\mathcal{O}(x^2)$ designates terms that are of order x^2 or greater.

We can then write the characteristic function as

$$\begin{aligned} \varphi_\omega(t) &= \left(\frac{p}{1 - (1-p)(1 + it\omega_1 p + \mathcal{O}(p^2))} \right)^{\frac{\gamma}{N(1-p)}} \\ &= \left(\frac{p}{p(1 - it\omega_1) + \mathcal{O}(p^2)} \right)^{\frac{\gamma}{N(1-p)}} \\ &= \left(\frac{1}{1 - it\omega_1 + \mathcal{O}(p)} \right)^{\frac{\gamma}{N(1-p)}}. \end{aligned} \quad [59]$$

In the limit $p \rightarrow 0$, this yields

$$\varphi_\omega(t) \rightarrow \left(\frac{1}{1 - it\omega_1} \right)^{\frac{\gamma}{N}}. \quad [60]$$

This is the characteristic function of the Gamma distribution with $\omega \sim \text{Gamma}(\frac{\gamma}{N}, \omega_1)$. We are grateful to Zakhar Kabluchko for this derivation.

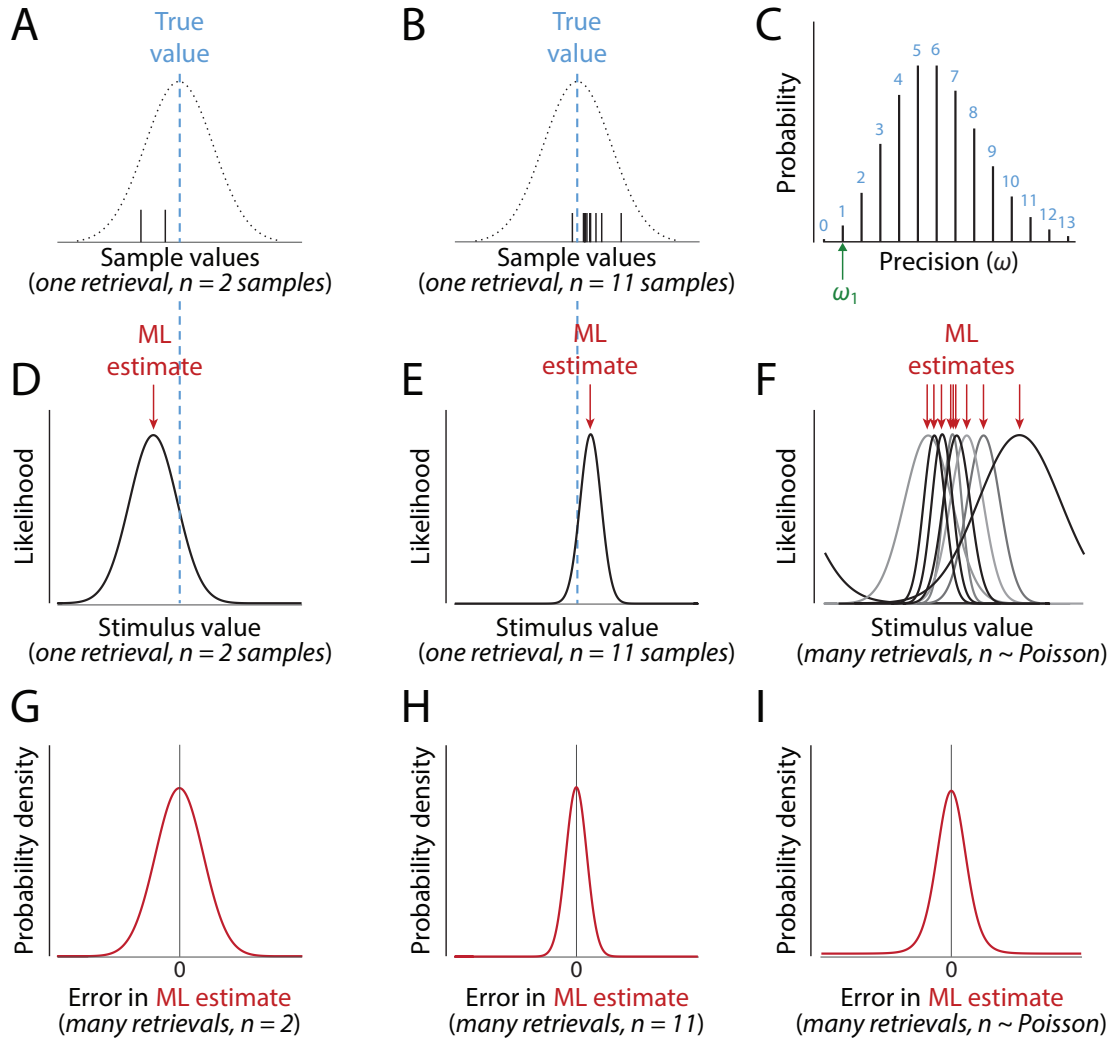


Fig. S1. Likelihood and variability in the stochastic sampling model. (A, B) Generation of samples. Retrieval of a stimulus feature corresponds to maximum likelihood (ML) estimation based on a set of n samples, each drawn from a normal distribution centered on the true feature value with a fixed precision (inverse of variability), ω_1 . The number of samples is drawn from a Poisson distribution. Examples shown are of estimation for $n = 2$ in (A) and panels below; $n = 11$ in (B) and panels below. (C) From retrieval to retrieval, precision (measured with respect to the likelihood) varies with a scaled Poisson distribution, taking on discrete values corresponding to the number of samples (blue) multiplied by the sample precision ω_1 . (D, E) The likelihood function on any particular retrieval measures the compatibility of different stimulus values with the obtained samples. For normally distributed samples, the likelihood is also normal with peak (the ML estimate) at the mean of the sample values. The likelihood width (corresponding to precision $n\omega_1$) indicates the reliability of the estimate and predicts subjective confidence. (F) Both the location and width of the likelihood vary between retrievals. (G, H) Considering only retrievals based on n samples, the error in the ML estimate varies according to a normal distribution with precision $n\omega_1$, equal to the precision of the corresponding likelihood functions. (I) Over all retrievals, the estimation error has a non-normal (leptokurtic) distribution, because it comprises a mixture of normal distributions with different precisions (as in (C)).

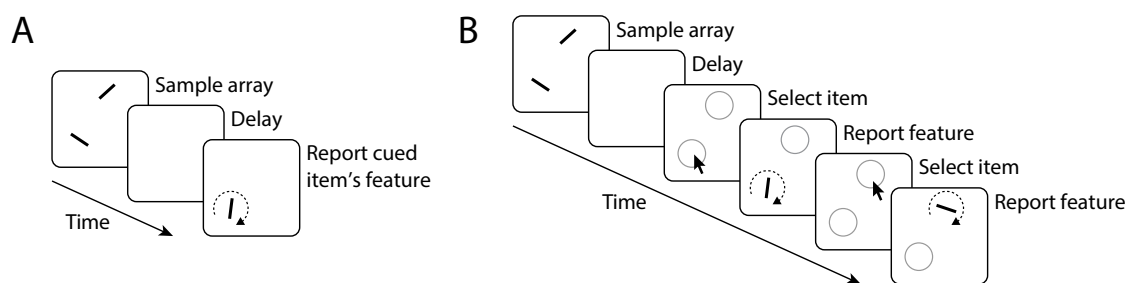


Fig. S2. Delayed reproduction tasks. (A) Structure of an illustrative single-report task with orientation report cued by object location. (B) Structure of an illustrative whole-report task. The participant sequentially reports the features of all items in the sample array, either in a freely chosen order (shown here) or a prescribed random order.

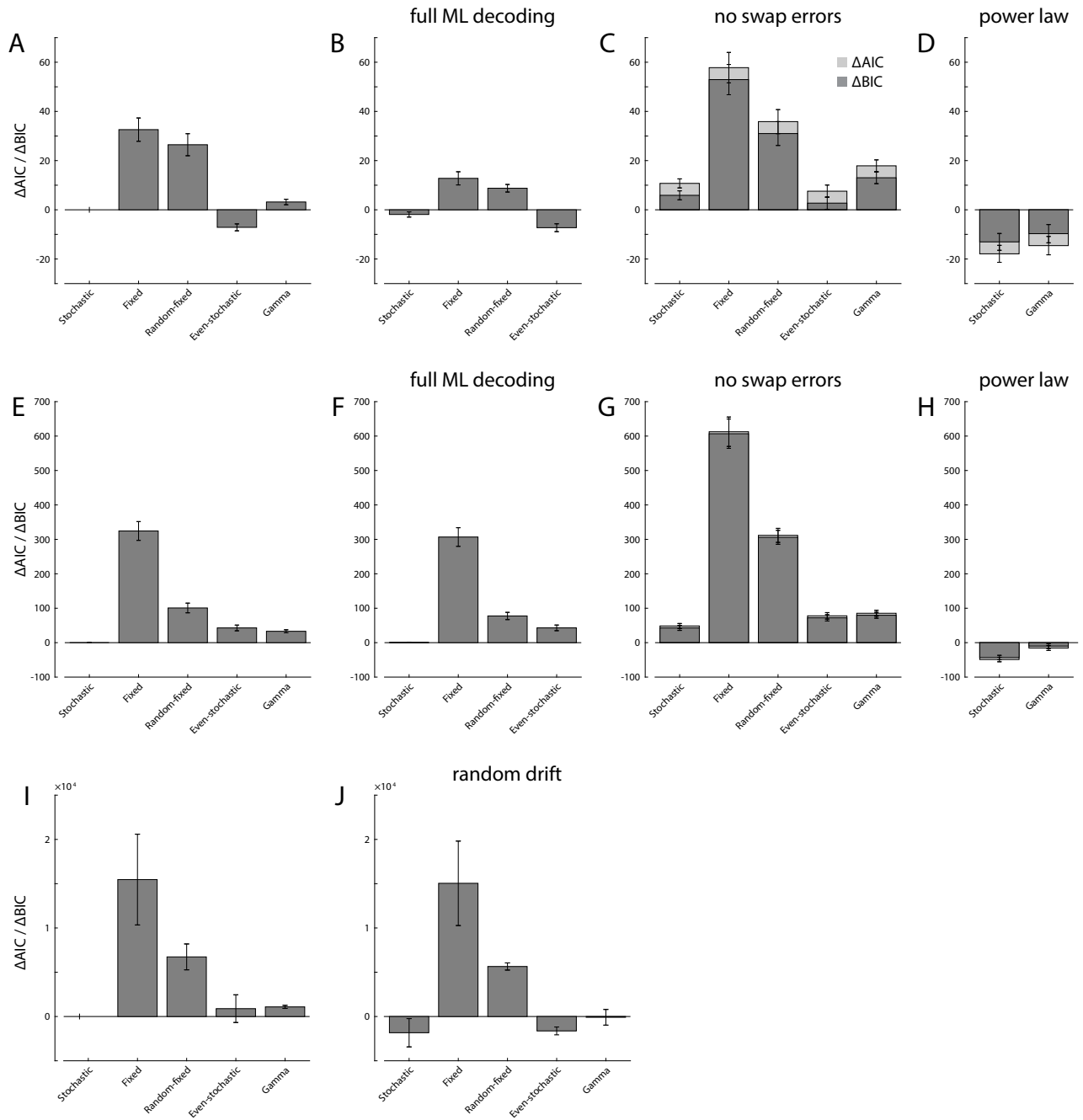


Fig. S3. Model comparison for additional model variants. Mean differences in AIC (light gray) and BIC values (dark gray) relative to the stochastic sampling model with swap errors are shown for single-report data (A-D), whole-report data (E-H), and whole-report data pooled over participants (I-J). Better models have lower values. Error bars indicate ± 1 SE. (A, E, I) 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 ML 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. (J) Model variants with random drift of memorized features over memory delay.

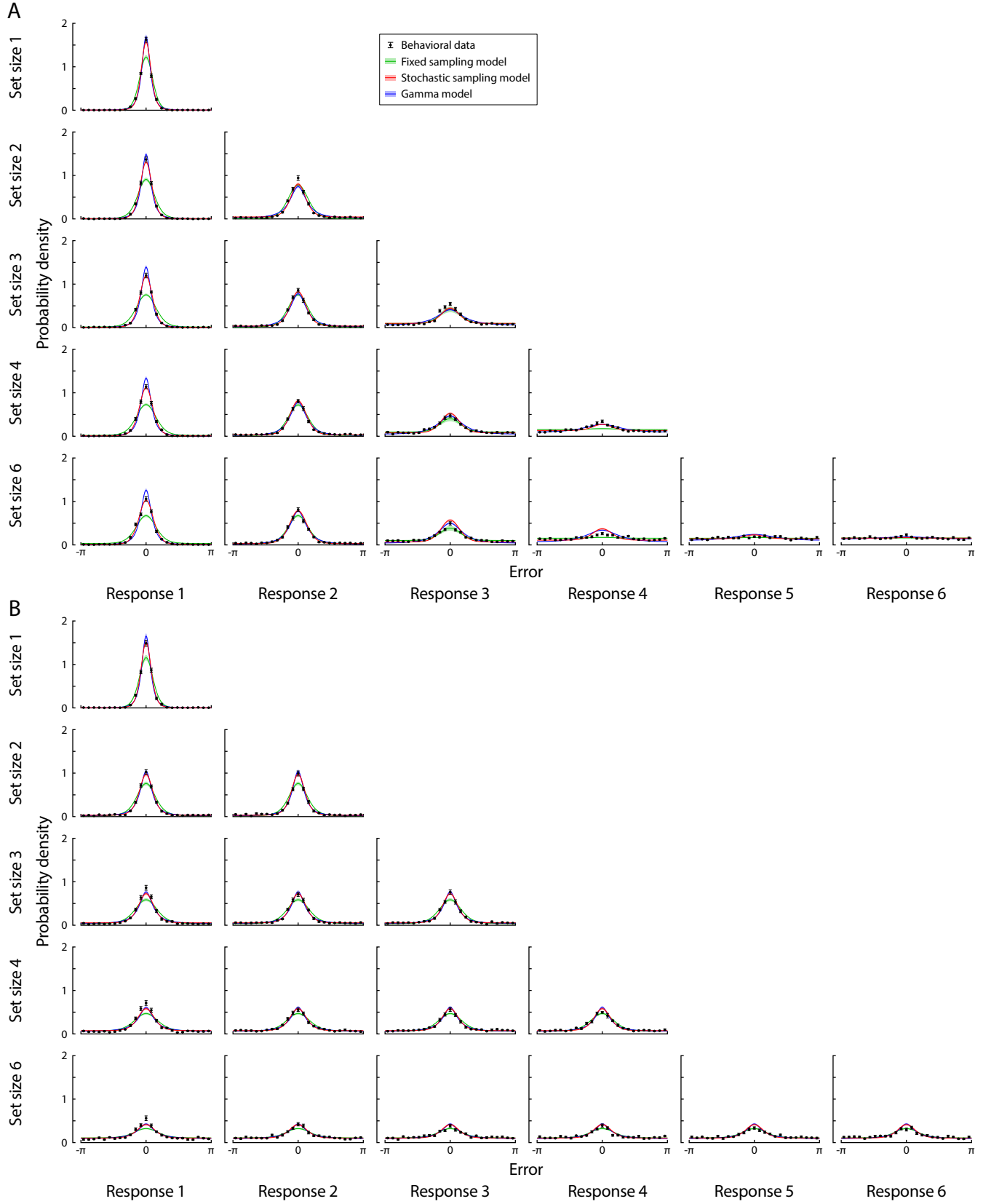


Fig. S4. Behavioral data and model fits in the whole-report task (2) for color. (A) Free response order condition (Experiment 1a). (B) Random response order condition (Experiment 2a). Solid lines show the mean across participants, and error bars and shaded areas indicate ± 1 SE.

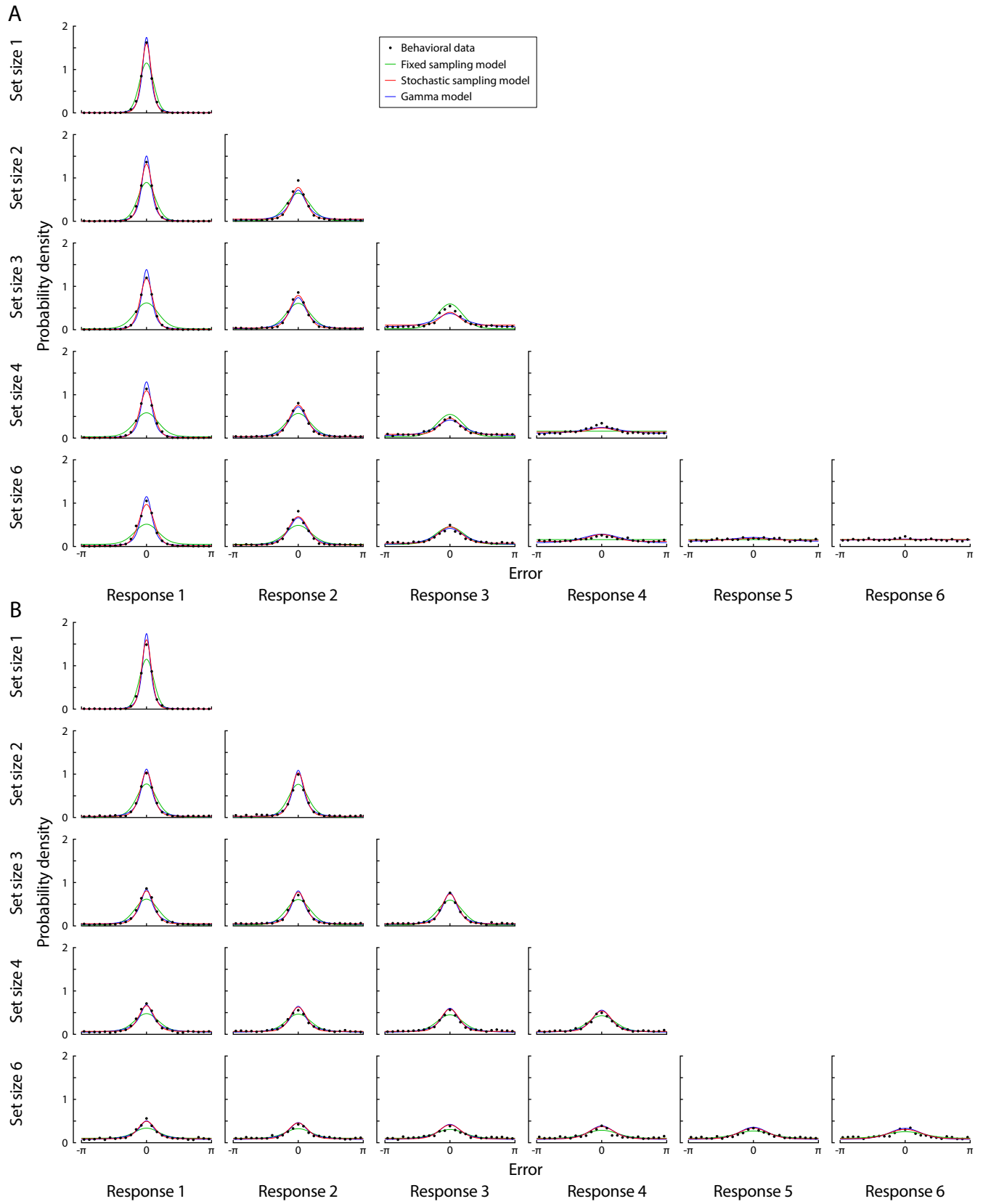


Fig. S5. Behavioral data pooled across participants and model fits with random drift of memorized feature values. Panels as in Fig. S4. For each model, a single set of parameters was obtained by ML fit across the two response order conditions, varying only in whether responses are ordered by precision of memory representations.

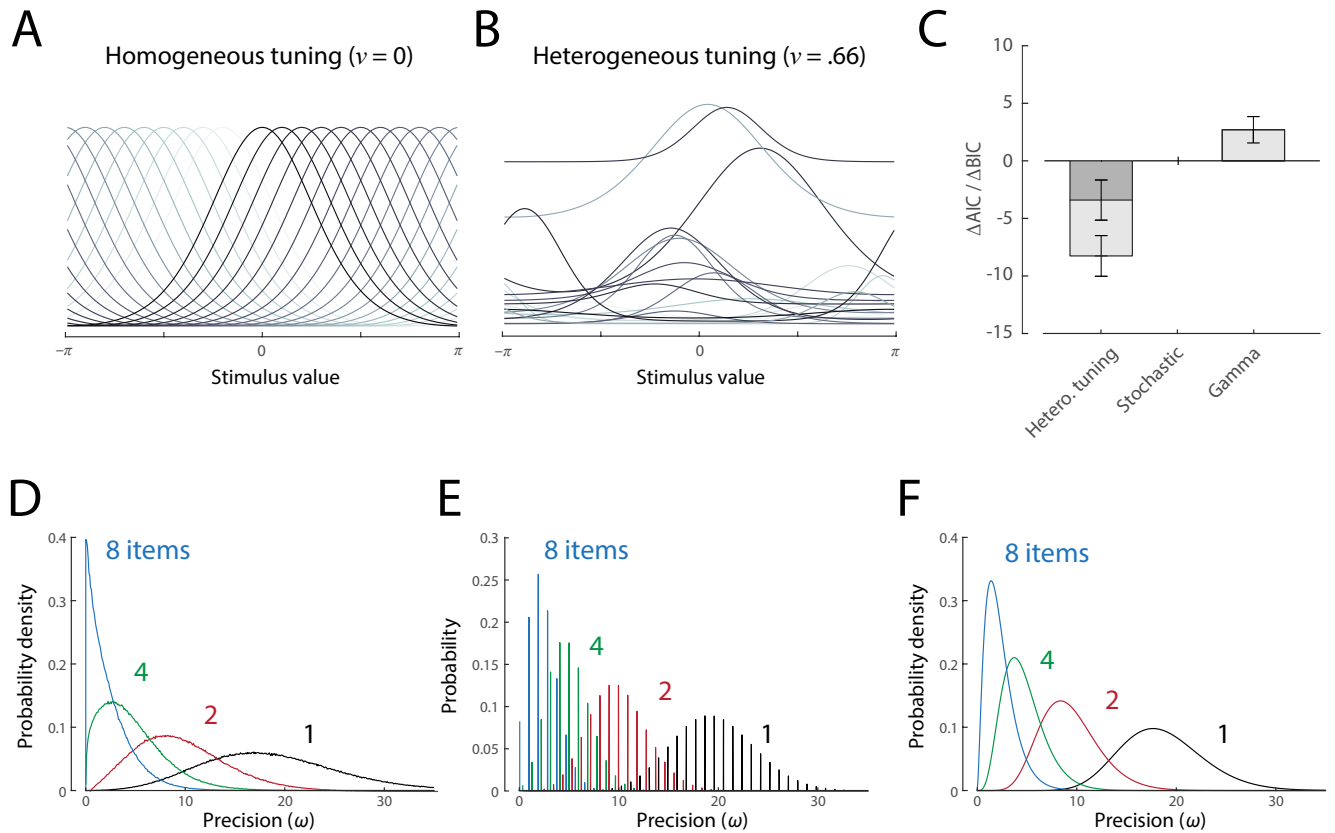


Fig. S6. Population coding model with heterogeneous tuning. (A) Examples of homogeneous tuning functions in the idealized neural population underlying the stochastic sampling model. (B) Examples of heterogeneous tuning functions corresponding to mean parameters of the fitted heterogeneous model. (C) 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.

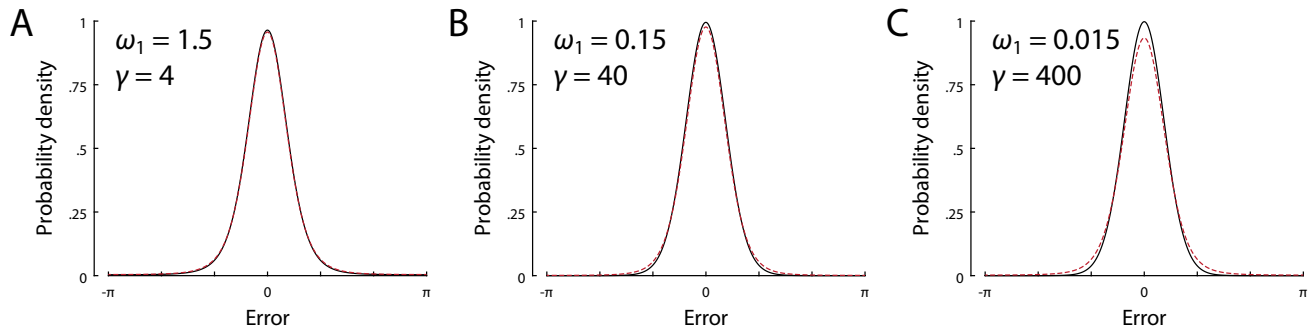


Fig. S7. Distributions of response errors in the stochastic sampling model, using either exact ML decoding in circular space (red dashed lines) or the approximation assuming a linear increase in precision (expressed as Fisher information) with the number of samples (black lines). The approximation is very close for typical parameter values (A; parameters here are close to median values of the stochastic sampling model fits scaled for set size 4), but deviations increase when moving to parameterizations with a higher number of low-precision samples (B, C).

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	Study	Feature	Set Sizes	Participants	Trials
1	Zhang & Luck, 2008 (5)	Color	1, 2, 3, 6	8	125
2	Bays et al., 2009 (10)	Color	1, 2, 4, 6	12	200
3	van den Berg et al., 2012 (3)	Color	1 – 8	13	216
4	van den Berg et al., 2012 (3)	Orientation	1 – 8	6	320
5	Rademakers et al., 2012 (12)	Orientation	3, 6	6	800
6	Bays, 2014 (8), Exp 1	Orientation	1, 2, 4, 8	8	230
7	Bays et al., 2011 (13), Exp 1	Orientation	1, 2, 4, 6	8	800
8	Bays, Wu & Husain, 2011 (14)	Orientation	1, 6	10	50, 250
9	Bays, Wu & Husain, 2011 (14)	Color	1, 6	10	50, 250
10	Gorgoraptis et al., 2011 (15)	Orientation	1 – 5	8	100
11	Pratte et al., 2017 (16)	Orientation	1, 2, 3, 6	12	640

Table S2. Whole-report experiments used for model comparison. All experiments are taken from (2).

No	Feature	response 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 S3. Parameter values of ML fits for single-report data (mean \pm 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) in order to provide more representative parameter values.

Model	γ or K	ω_1	p_{NT}	α
Stochastic	13.2 \pm 1.7	1.84 \pm 0.11	0.0245 \pm 0.0023	
Fixed	4.90 \pm 0.21	2.80 \pm 0.14	0.0265 \pm 0.0023	
Random-fixed	11.4 \pm 0.7	1.55 \pm 0.10	0.0238 \pm 0.0023	
Even-stochastic	7.31 \pm 1.47	3.04 \pm 0.15	0.0287 \pm 0.0024	
Gamma	8.63 \pm 1.94	5.00 \pm 0.36	0.0281 \pm 0.0024	
<i>full ML decoding for circular feature spaces</i>				
Stochastic	35.0 \pm 6.3	1.68 \pm 0.12	0.0261 \pm 0.0024	
Fixed	11.7 \pm 0.9	2.10 \pm 0.17	0.0305 \pm 0.0027	
Random-fixed	13.8 \pm 0.8	1.46 \pm 0.11	0.0262 \pm 0.0024	
Even-stochastic	15.9 \pm 3.3	2.93 \pm 0.17	0.0289 \pm 0.0024	
<i>excluding swap errors</i>				
Stochastic	8.35 \pm 0.81	2.28 \pm 0.13		
Fixed	3.67 \pm 0.13	3.25 \pm 0.17		
Random-fixed	6.91 \pm 0.34	2.08 \pm 0.12		
Even-stochastic	4.49 \pm 0.14	3.55 \pm 0.16		
Gamma	3.95 \pm 0.73	8.16 \pm 0.62		
<i>power law for set size effects</i>				
Stochastic	8.46 \pm 0.51	1.96 \pm 0.11	0.0310 \pm 0.0026	0.809 \pm 0.031
Gamma	4.35 \pm 0.37	5.60 \pm 0.40	0.0339 \pm 0.0026	0.809 \pm 0.032

Table S4. Parameter values of ML fits for whole-report data (mean \pm 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).

Model	γ or K	ω_1	p_{NT}	α
Stochastic	4.30 \pm 0.14	5.18 \pm 0.25	0.0367 \pm 0.0041	
Fixed	2.72 \pm 0.10	4.85 \pm 0.31	0.0529 \pm 0.0036	
Random-fixed	4.21 \pm 0.16	4.35 \pm 0.28	0.0481 \pm 0.0039	
Even-stochastic	2.87 \pm 0.07	7.20 \pm 0.32	0.0292 \pm 0.0039	
Gamma	1.40 \pm 0.06	23.2 \pm 1.6	0.0384 \pm 0.0043	
<i>full ML decoding for circular feature spaces</i>				
Stochastic	4.31 \pm 0.14	5.19 \pm 0.25	0.0367 \pm 0.0041	
Fixed	6.08 \pm 0.88	4.54 \pm 0.35	0.0551 \pm 0.0037	
Random-fixed	4.19 \pm 0.17	4.40 \pm 0.28	0.0491 \pm 0.0040	
Even-stochastic	2.87 \pm 0.07	7.23 \pm 0.32	0.0292 \pm 0.0039	
<i>excluding swap errors</i>				
Stochastic	3.57 \pm 0.11	5.66 \pm 0.24		
Fixed	2.42 \pm 0.06	2.97 \pm 0.21		
Random-fixed	3.61 \pm 0.12	3.32 \pm 0.23		
Even-stochastic	2.56 \pm 0.06	7.43 \pm 0.32		
Gamma	1.07 \pm 0.04	31.7 \pm 2.1		
<i>power law for set size effects</i>				
Stochastic	7.06 \pm 0.39	4.64 \pm 0.21	0.0239 \pm 0.0032	1.37 \pm 0.03
Gamma	2.33 \pm 0.12	19.5 \pm 1.4	0.0258 \pm 0.0034	1.38 \pm 0.03

Table S5. Parameter values of ML fits for pooled whole-report data (with a single set of parameters for Experiments 1a & 2a, and for Experiments 1b & 2b).

Model	γ or K	ω_1	p_{NT}	η
<i>Experiments 1a & 2a</i>				
Stochastic	3.66	4.25	0.0204	
Fixed	3.00	2.54	0.0631	
Random-fixed	4.00	2.72	0.0478	
Even-stochastic	2.54	5.74	0.00576	
Gamma	1.21	17.3	0.0251	
<i>Experiments 1b & 2b</i>				
Stochastic	3.80	6.81	0.0219	
Fixed	2.00	6.77	0.0309	
Random-fixed	4.00	4.79	0.0428	
Even-stochastic	2.64	9.20	0.0169	
Gamma	1.09	37.3	0.0261	
<i>Experiments 1a & 2a, with random drift</i>				
Stochastic	3.70	4.67	0.0181	0.0181
Fixed	3.00	2.69	0.0614	0.0194
Random-fixed	4.00	2.78	0.0464	0.0114
Even-stochastic	2.61	6.72	0.00434	0.0232
Gamma	1.18	19.6	0.0236	0.0151
<i>Experiments 1b & 2b, with random drift</i>				
Stochastic	3.99	9.62	0.0136	0.0309
Fixed	2.00	7.60	0.0265	0.0207
Random-fixed	4.00	5.34	0.0314	0.0265
Even-stochastic	2.81	14.16	0.0100	0.0341
Gamma	0.995	67.3	0.0196	0.0292

References

1. van den Berg R, Awh E, Ma WJ (2014) Factorial comparison of working memory models. *Psychological Review* 121(1):124.
2. Adam KC, Vogel EK, Awh E (2017) Clear evidence for item limits in visual working memory. *Cognitive Psychology* 97:79–97.
3. van den Berg R, Shin H, Chou WC, George R, Ma WJ (2012) Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences* 109(22):8780–8785.
4. Bays PM (2016) A signature of neural coding at human perceptual limits. *Journal of Vision* 16(11):4–4.
5. Zhang W, Luck SJ (2008) Discrete fixed-resolution representations in visual working memory. *Nature* 453(7192):233–235.
6. Fougner D, Suchow JW, Alvarez GA (2012) Variability in the quality of visual working memory. *Nature Communications* 3:1229.
7. Taylor R, Bays PM (2020) Theory of neural coding predicts an upper bound on estimates of memory variability. *Psychological Review* Advance Online Publication. <http://dx.doi.org/10.1037/rev0000189>.
8. Bays PM (2014) Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience* 34(10):3632–3645.
9. Ecker AS, et al. (2010) Decorrelated neuronal firing in cortical microcircuits. *Science* 327(5965):584–587.
10. Bays PM, Catalao RF, Husain M (2009) The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision* 9(10):7–7.
11. Schneegans S, Bays PM (2018) Drift in neural population activity causes working memory to deteriorate over time. *Journal of Neuroscience* pp. 3440–3417.
12. Rademaker RL, Tredway CH, Tong F (2012) Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision* 12(13):21–21.
13. Bays PM, Gorgoraptis N, Wee N, Marshall L, Husain M (2011) Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision* 11(10):6–6.
14. Bays PM, Wu EY, Husain M (2011) Storage and binding of object features in visual working memory. *Neuropsychologia* 49(6):1622–1631.
15. Gorgoraptis N, Catalao RF, Bays PM, Husain M (2011) Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience* 31(23):8502–8511.
16. Pratte MS, Park YE, Rademaker RL, 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.