Stochastic sampling provides a unifying account of working memory limits

Sebastian Schneegans, Robert Taylor & Paul M Bays*

University of Cambridge, Department of Psychology, Cambridge, CB2 3EB, UK

Abstract. 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 new framework to understand working memory limits. Reconceptualizing existing models in these terms reveals strong commonalities between seemingly opposing accounts, and shows that random variability in sample counts, rather than discreteness, is the key to reproducing human behavioral performance. 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.

Elementary features of objects are represented within the human visual system in the form of population codes [1]. A simple model [2] of limits on representing multiple stimuli [3–5] assumes each stimulus is encoded in a separate pool of neurons with identical tuning curves, each centered on a different (preferred) feature value, such that the cells densely and uniformly cover a one-dimensional feature space (Fig. 1A). Each neuron's response to a stimulus consists of discrete spikes generated by a Poisson process at the rate determined by its tuning function. To make a connection with sampling [6–12], we associate each spike from a given

*Email for correspondence: pmb20@cam.ac.uk.

pool with the preferred feature value of the neuron that emitted it, obtaining a probability distribution $p(\varphi)$ that any randomly selected spike is associated with a preferred value φ (Fig. 1B). Each spike can be viewed as a sample drawn from this distribution, which has the same shape as the neural tuning function and is centered on the true stimulus value (see SI Sec. 4.1).

Retrieval of a stimulus feature value is modeled as maximum likelihood estimation based on the spikes generated within a fixed decoding window. For Gaussian tuning functions this is realized by simple averaging of the sample values. Due to the superposition property of Poisson processes, the number of spikes – or samples – generated by a pool within the decoding window is also a Poisson random variable. If the total spike rate is normalized at a population level γ and distributed evenly among N stimuli (as proposed in [2]), the mean number of samples available to recover each item is γ/N, but the actual number varies from one retrieval to the next according to a Poisson distribution. Decoding of population activity can therefore be interpreted as *stochastic* sampling of stimulus features.

This stochasticity stands in contrast with most previous sampling-based models in the attentional and memory literature, and with the influential *slots+averaging* model [13], which can also readily be interpreted in terms of sampling (Fig. 1E–G). Each slot is postulated to hold a representation of a single object with a fixed precision, and thus provides a noisy sample of the objects' feature values. Multiple slots, or samples, that correspond to the same object are averaged at retrieval to improve recall precision. Independent of whether working memory representations are feature- or object-based [14, 15], the critical difference from stochastic sampling is that the total number of samples available for all items is fixed at a value K.

We now consider the distribution of representational precision in these models. For any given set of samples, the information they provide about the stimulus is described by the likelihood function. The width of the likelihood function is a measure of uncertainty in the estimate that also reflects trial-to-trial variability (see SI Sec. 4.2 and Fig. S1), so we define the precision of an individual estimate in terms of this width. If samples are drawn from a Gaussian distribution, precision increases linearly with the number of samples.

In the stochastic sampling model, precision has a Poisson distribution scaled by the precision of a single sample (Fig. 1C). The distribution of decoding errors can be described as a scale mixture of normal distributions with precision proportional to the sample count (Fig. 1D; for

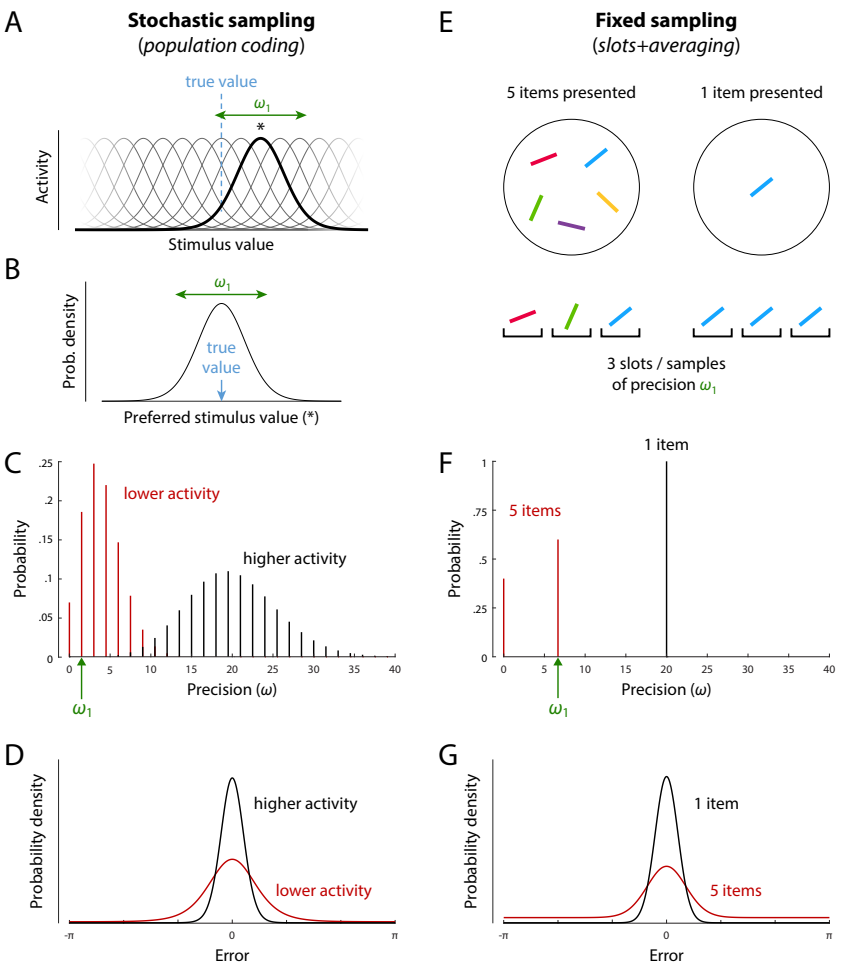

Figure 1: Working memory models. (A–D) Sampling interpretation of a population coding model. (A) Tuning curves of an idealized neural population. One neuron's tuning and preferred value [*] is highlighted. (B) Probability distribution over stimulus space obtained by associating each spike with the preferred stimulus of the neuron that generated it. This distribution has the same precision (ω_1) as the tuning functions. (C) Precision of ML estimates (defined in terms of the likelihood function, see Fig. S1) follows a Poisson distribution scaled by the tuning precision ω_1. (D) Errors in estimation (in a circular feature space) are described by a scale mixture of distributions with precision shown in (C). (E–G) Sampling interpretation of the *slots+averaging* model. (E) Allocation of a fixed number of samples to memory displays of different sizes. (F) Precision is discretely distributed as a product of the precision of one sample ω_1 and the number of samples allocated per item. (G) Corresponding distribution of estimation errors.

circular stimulus spaces typically used experimentally, this is a close approximation rather than exact, see SI Sec. 5.2). The dispersion of errors increases with decreasing activity (e.g. as a result of increasing set size; black curve vs red curve in Fig. 1D) and their distribution is leptokurtic, with long tails evident at lower activity levels (red curve).

In the fixed sampling model, making the common assumption that samples are distributed as evenly as possible among items [13, 16], we obtain a discrete distribution over at most two precision values (Fig. 1F), which are multiples of the sample precision. 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. 1G).

We fit the stochastic and fixed sampling models to a large dataset of single-report and whole-report tasks (see SI Sec. 1 and Fig. S2). In the latter, participants reported the feature values of all presented items, providing additional information about subjective confidence (which we equate with likelihood width) and error correlations between items, for which fixed and stochastic models make differing predictions. The stochastic model fit data substantially better than the fixed sampling model for both types of task (Fig. 2A and B, Fig. S3), indicating that stochasticity is critical for capturing behavioral performance. Contrary to previous interpretations [17], model comparison on whole-report data did not support a slot-like mechanism with a fixed item limit. Intermediate models in which a fixed number of samples were randomly allocated to items (*random–fixed* model) or a Poisson random number of samples was distributed as evenly as possible between items (*even–stochastic* model) produced intermediate qualities of fit overall (Fig. 2C), with the latter's advantage over the fully stochastic model in single-report data outweighed by its significantly worse fit to whole-report data.

For the models examined above, typical fitted parameters indicate that estimates are based on relatively small numbers of samples. To investigate whether these low sample counts are important for reproducing human performance, we implemented a generalization of the stochastic model (based on a scaling of the negative binomial distribution, see SI Sec. 2.6 and 6) which maintains its key characteristics (mean and variance of precision scale inversely with set size, resulting in a fixed Fano factor [FF], Fig. 2D) while the number of samples and the sample precision are controlled in inverse relation by a discretization parameter p (Fig. 3).

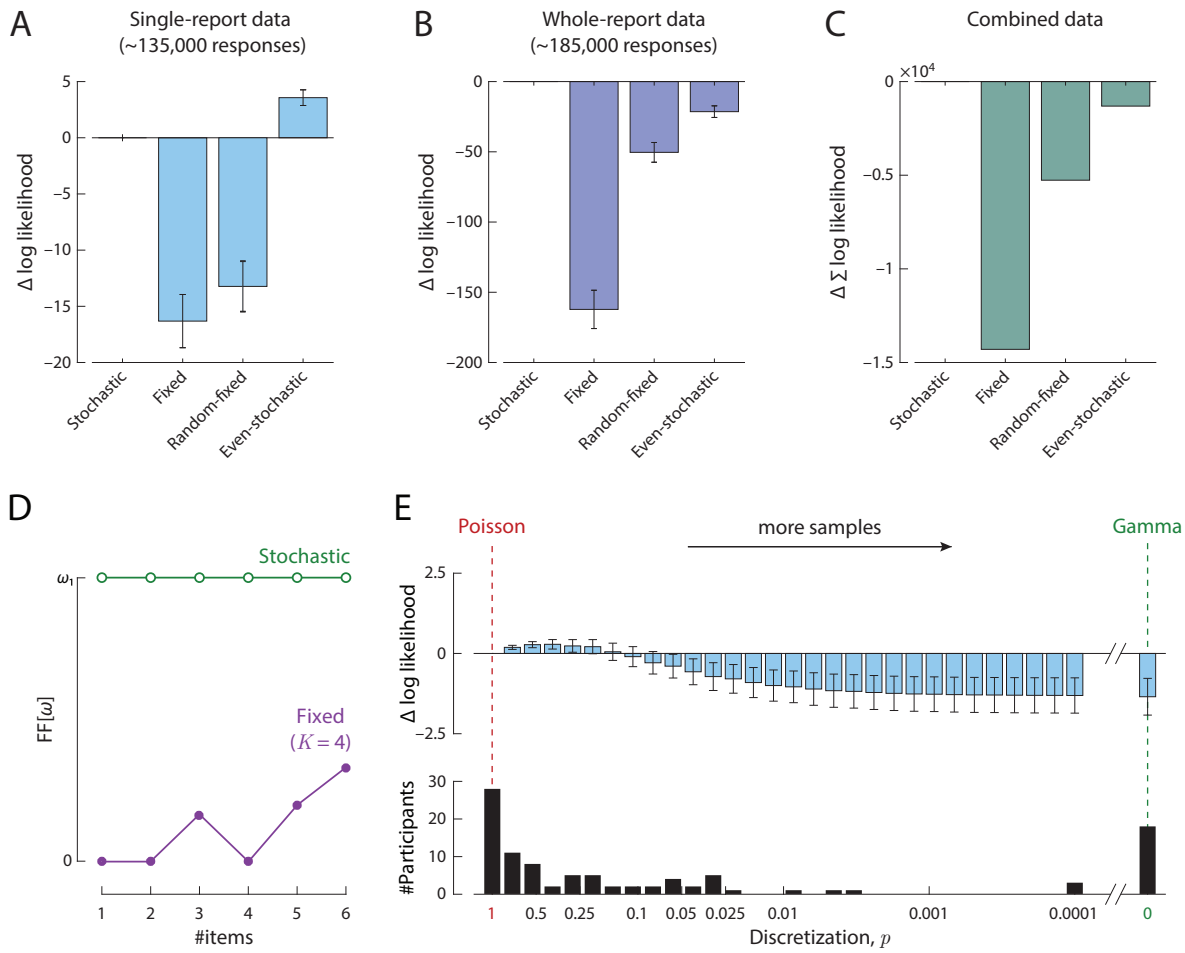

Figure 2: Model comparison based on single- and whole-report data. (A) Mean difference in log-likelihood for each model from the stochastic sampling model, for a benchmark data set of single-report experiments. Errorbars indicate ± 1 SE across participants. (B) The same comparison for a set of whole-report experiments. (C) Total difference in log-likelihood between models across single- and whole-report experiments. (D) Ratio of variance to mean precision (Fano Factor) is independent of set size only in the stochastic model. (E) Mean difference in log-likelihood for differing levels of discretization in the generalized stochastic model (top), plotted relative to the maximum discretization ($p = 1$), and (bottom) number of participants best fit by each level of discretization. All models have the same number of free parameters and include a fixed per-item probability of swap errors (see SI Sec. 2.1).

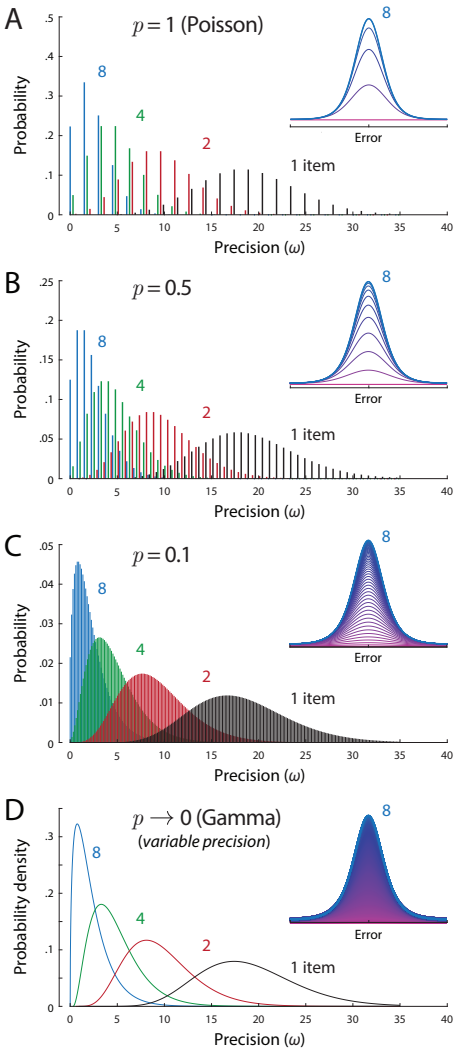

Figure 3: Generalized stochastic model. Main panels: Precision distributions in the generalized model with typical fitted parameters (γ = 12, ω_1 = 1.5) for different levels of discretization p and different set sizes. Insets: Construction of the corresponding distributions of response error (for set size 8), with thin lines showing normal distributions with different precisions incrementally accumulated in ascending order (magenta to blue). (A) Poisson-distributed precision values (p = 1). (B & 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 $p \rightarrow 0$, the mean number of samples becomes infinite and the distribution over precision approaches a continuous Gamma distribution (as in [16, 18]). The Fano factor is fixed at ω_1 across all set sizes and levels of discretization. Note that discrete precision values for different set sizes are slightly shifted for visibility.

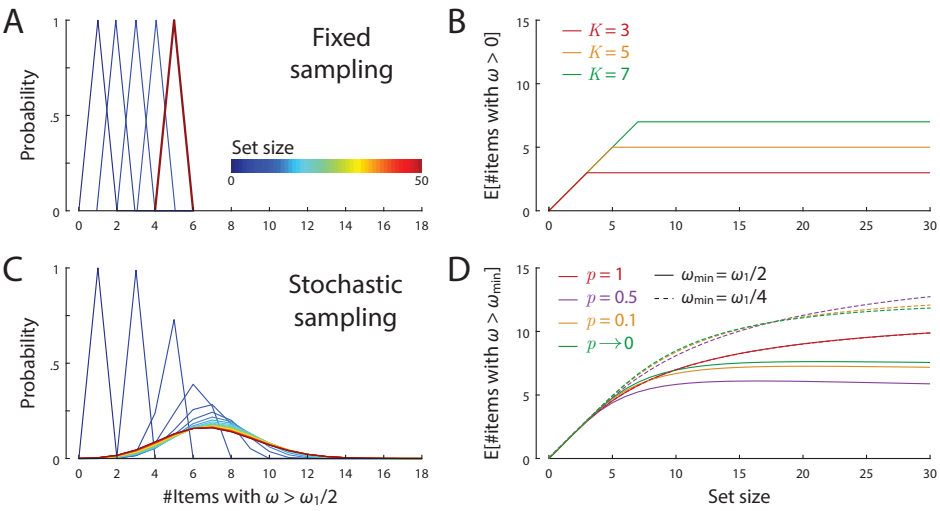

Figure 4: Item limits in sampling models. (A) Example probability distribution of the number of items recovered with greater than zero precision for different set sizes (color coded, increasing blue to red; discrete probability distributions are depicted as line plots for better visualization) in the fixed sampling model with K = 5 total samples. (B) Mean number of items with above-zero precision as a function of set size for different numbers of samples K. (C) Example probability distribution of the number of items recovered with greater than a fixed threshold precision (as proportion of base precision ω_1) in the generalized stochastic sampling model, with γ = 12, ω_1 = 1.5, p = 0.1. Note the existence of a probabilistic item limit at $\sim$7 items, which does not directly correspond to the number of samples in the model (mean 120). (D) Mean number of items with above-threshold precision for the generalized stochastic model with different levels of discretization (colors) and threshold precision (solid versus dashed lines). For finite thresholds smaller than the base precision ω_1, the number of above-threshold items saturates with increasing set size for all model parameters.

As the precision of each individual sample decreases, we find that the precision distribution approaches a continuous Gamma distribution (Fig. 3D; see SI Sec. 6.3). Two previous studies [16, 18] independently proposed a continuous scale mixture of normal distributions with Gamma-distributed precision to account for behavioral data, but could not motivate this choice theoretically. We can now account for these *variable precision* models as a limiting case of stochastic sampling with a very large number of very low-precision samples. We found that the Gamma model ($p \to 0$) fit single-report data more poorly than the Poisson model ($p = 1$), and the best fits were obtained at intermediate levels of discretization (maximum likelihood at p = 0.39; Fig. 2E, top). However, individuals varied considerably in their estimated discretization parameter (Fig. 2E, bottom), and differences in fit (measured in log likelihood) were on average an order of magnitude smaller than those between fixed and stochastic sampling (compare with Fig. 2A), reflecting the limited effect of sampling discreteness on predicted error distributions (insets in Fig. 3).

One prediction of working memory models with a fixed number of samples [13, 17] is the appearance of random "guesses" once the number of items exceeds that limit (Fig. 4A-B). In the stochastic sampling account, the number of samples available for each item varies probabilistically and independently of every other item. Nonetheless, if a small positive precision value is chosen as a threshold, the expected number of items that exceed that precision threshold will saturate as set size increases, irrespective of the level of discretization (Fig. 4C-D; SI Sec. 6.1). The asymptote depends on both the threshold precision and the level of discretization, and does not correspond in any direct way to the number of samples in the model. Thus, a probabilistic item limit is an emergent property of stochastic sampling that does not require an explicit mechanism nor imply a particular number of samples.

In order for the Fano factor relating the variance and mean of precision to be held constant as sample counts increase, the sample counts themselves must become "overdispersed" compared to Poisson variability (i.e. FF > 1). Overdispersion of spike counts is a common observation in visual cortical neurons, typically with FF in the range 1.5–3 (e.g. [19]), corresponding in our model to discretization p in the range 0.33–0.75. Additionally, several other factors present in real neural populations could have effects similar to decreasing discretization in the generalized model. Heterogeneity in tuning functions [20] leads to variation in the information carried by each spike, with the effect of smoothing out the discrete distributions over precision predicted by a homogeneous Poisson model. This has similar consequences for estimation error to decreasing p in the generalized model. Consistent with this idea, incorporating biologically realistic heterogeneity into the population model improved fits to data (see SI Sec. 5.1 and Fig. S4).

Spikes in real neural populations are not independent events as assumed by the sampling interpretation, but rather correlated within and between neurons. This will tend to result in deviations from the simple additivity assumed by sampling. An implementation of short-range correlations in the population model greatly increased the numbers of decoded spikes required to reproduce behavioral data, without changing quality of fit (see SI Sec. 5.1). We note however that the exact consequences of spike correlations for decoding depend on details of correlation structure that are difficult to measure experimentally [21–23], and suboptimal inference (in the form of a mismatched decoder) could play a part [24].

The degree to which working memory samples are discretized versus continuous has only very weak effects on predicted retrieval errors under the stochastic sampling model (e.g. insets of Fig. 3). Importantly, discrete representations are compatible with an underlying continuous memory resource that can be distributed according to behavioral goals [25–27]: indeed in the stochastic model the integer number of samples available for each item at retrieval is unpredictable, and so cannot be the basis of prioritization. Instead, the resource distributed between items corresponds to the mean or expected total number of samples, which is constant and continuous-valued – in the neural model [2] this is equated with the instantaneous firing rate or membrane potential, while decoding is based on the expression of this rate in discrete spikes.

The stochastic sampling model can be understood at multiple levels: in purely descriptive terms as a form of mixture model (like the *normal+uniform* model, [13]); at a cognitive level in terms of averaging samples; and at a neurocomputational level via its implementation in population coding. The neural interpretation provides the link to another recent proposal for understanding recall errors – psychophysical scaling [28] – which has an alternative expression as a Gaussian-noise population model [29].

Acknowledgments

We thank M. Lengyel, M. Bays, W. J. Ma & M. Husain for helpful discussion, Z. Kabluchko for statistical advice, and the researchers who publicly shared data that facilitated this study. We used resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service.

Funding

Supported by Wellcome Trust grant 106926 (P.M.B.)

Data Availability

Data and code associated with this article will be made publicly available at https://osf.io/buxp9/.

References

[1] A. Pouget, P. Dayan, and R. S. Zemel. “Inference and computation with population codes”. *Annual review of neuroscience* 26.1 (2003), pp. 381–410.

[2] P. M. Bays. “Noise in neural populations accounts for errors in working memory”. *Journal of Neuroscience* 34.10 (2014), pp. 3632–3645.

[3] S. M. Emrich et al. “Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory”. *Journal of Neuroscience* 33.15 (2013), pp. 6516–6523.

[4] T. C. Sprague, E. F. Ester, and J. T. Serences. “Reconstructions of information in visual spatial working memory degrade with memory load”. *Current Biology* 24.18 (2014), pp. 2174–2180.

[5] D. W. Sutterer et al. “Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory”. *PLoS biology* 17.4 (2019), e3000239.

[6] M. L. Shaw. “Identifying attentional and decision-making components in information processing”. *Attention and performance VIII* 8 (1980), pp. 277–295.

[7] J. Palmer. “Attentional limits on the perception and memory of visual information”. *Journal of Experimental Psychology: Human Perception and Performance* 16.2 (1990), p. 332.

[8] D. K. Sewell, S. D. Lilburn, and P. L. Smith. “An information capacity limitation of visual short-term memory”. *Journal of Experimental Psychology: Human Perception and Performance* 40.6 (2014), p. 2214.

[9] A.-M. Bonnel and J. Miller. “Attentional effects on concurrent psychophysical discriminations: Investigations of a sample-size model”. *Perception & Psychophysics* 55.2 (1994), pp. 162–179.

[10] E. Vul et al. “One and Done? Optimal Decisions From Very Few Samples”. *Cognitive Science* 38.4 (2014), pp. 599–637.

[11] G. Orbán et al. “Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex”. *Neuron* 92.2 (2016), pp. 530–543.

[12] M. N. Shadlen and D. Shohamy. “Decision Making and Sequential Sampling from Memory”. *Neuron* 90.5 (2016), pp. 927–939.

[13] W. Zhang and S. J. Luck. “Discrete fixed-resolution representations in visual working memory”. *Nature* 453.7192 (2008), pp. 233–235.

[14] G. A. Alvarez and P. Cavanagh. “The capacity of visual short-term memory is set both by visual information load and by number of objects”. *Psychological science* 15.2 (2004), pp. 106–111.

[15] K. Oberauer and S. Eichenberger. “Visual working memory declines when more features must be remembered for each object”. *Memory & Cognition* 41.8 (2013), pp. 1212–1227.

[16] R. van den Berg et al. “Variability in encoding precision accounts for visual short-term memory limitations”. *Proceedings of the National Academy of Sciences* 109.22 (2012), pp. 8780–8785.

[17] K. C. Adam, E. K. Vogel, and E. Awh. “Clear evidence for item limits in visual working memory”. *Cognitive psychology* 97 (2017), pp. 79–97.

[18] D. Fougnie, J. W. Suchow, and G. A. Alvarez. “Variability in the quality of visual working memory”. *Nature communications* 3 (2012), p. 1229.

[19] R. Vogels, W. Spileers, and G. A. Orban. “The Response Variability of Striate Cortical Neurons in the Behaving Monkey”. *Experimental brain research* 77.2 (1989), pp. 432–436.

[20] A. S. Ecker et al. “Decorrelated neuronal firing in cortical microcircuits”. *Science* 327.5965 (2010), pp. 584–587.

[21] B. B. Averbeck, P. E. Latham, and A. Pouget. “Neural correlations, population coding and computation”. *Nature reviews neuroscience* 7.5 (2006), p. 358.

[22] A. S. Ecker et al. “The effect of noise correlations in populations of diversely tuned neurons”. *Journal of Neuroscience* 31.40 (2011), pp. 14272–14283.

[23] R. Moreno-Bote et al. “Information-limiting correlations”. *Nature neuroscience* 17.10 (2014), p. 1410.

[24] J. M. Beck et al. “Not noisy, just wrong: the role of suboptimal inference in behavioral variability”. *Neuron* 74.1 (2012), pp. 30–39.

[25] A. H. Yoo et al. “Strategic allocation of working memory resource”. *Scientific reports* 8.1 (2018), p. 16162.

[26] S. M. Emrich, H. A. Lockhart, and N. Al-Aidroos. “Attention mediates the flexible allocation of visual working memory resources.” *Journal of Experimental Psychology: Human Perception and Performance* 43.7 (2017), p. 1454.

[27] Z. Klyszejko, M. Rahmati, and C. E. Curtis. “Attentional priority determines working memory precision”. *Vision research* 105 (2014), pp. 70–76.

[28] M. W. Schurgin, J. T. Wixted, and T. F. Brady. “Psychophysical scaling reveals a unified theory of visual memory strength”. *BioRxiv* (2018), p. 325472.

[29] P. M. Bays. “Correspondence between population coding and psychophysical scaling models of working memory”. *BioRxiv* (2019), p. 699884.