Efficient Bayesian observer models: the effects of noise during working memory maintenance on perceptual bias

Word count: 3999 (wordcounter.net)

Summary

Biases are the signatures of cognition. To explain them, constraints are often evoked without rules for determining the conditions under which they apply. Recently, an integration of Bayesian observer models with efficient coding constraints (“efficient Bayesian observer models”) proposed a general expression for bias under any prior distribution and noise conditions, contributing towards a unifying theory of perception. Importantly, this explains “anti-Bayesian” biases repelling perception away from the Bayesian prior, like a repulsive bias to perceive away from cardinal orientations. It does so by postulating a critical dependence of bias on noise sources. However, little work has investigated rich types of noise in perceptual decisions.

Here, we explored sources of noise introduced while maintaining information working memory (WM) over delays between stimuli and response. This let us test how WM might be integrated into efficient Bayesian observer models. Specifically, we investigated how noise from memory delay and set size (number of memoranda) affected bias in orientation perception using an analogue report task, in which subjects reproduced memorized orientation stimuli after a variable delay. Depending on the type of noise that set size and delay constitute, and the representations they affect, this gave qualitatively different predictions on how repulsive bias changes over delay and set size.

Crucially, we found that repulsive bias did not change with delay or set size variations. Under efficient Bayesian model predictions, this suggested that noise from WM delays didn't affect sensory representations from early encoding, but affected percepts at stages after sensory decoding. This challenged the prevailing view that WM stores sensory representations until they are decoded right before a decision. Rather, it suggested an early inference model of WM and perception, in which sensory representations are decoded immediately, then stored in WM. Our results also suggested a reconceptualization of set size effects on WM as affecting the signal-to-noise ratio of only decoded percepts, but not earlier sensory representations. This implied that WM might use scalar value representations, not probabilistic distributions, more than contemporary views believed.

Introduction

Perceptual biases often provide insight into the nature of representation and computation in perception. According to the Bayesian hypothesis, biases are the products of optimally integrating information based on its uncertainty. Investigating biases have produced various disjoint explanatory heuristics. One is bias towards the peak of the Bayesian prior, e.g., the "light from above" prior to explain shape perception (Sun & Perona, 1998). Another is regression towards the mean, e.g., the central tendency bias to explain perceiving two colors as closer to their average (Olkkonen et al., 2014). The challenge is formulating a unifying theory that explains perceptual biases under general conditions. A promising direction is to integrate models of perception and working memory (WM), which overlap significantly in theoretical and empirical findings. Here, we explore this overlap by investigating noise during WM maintenance and its effects on perceptual bias.

A basis for a unifying theory of perception is the recent integration of Bayesian observer models with efficient encoding constraints (Wei & Stocker, 2015). Efficient Bayesian observer models give a general expression of bias direction and magnitude under any prior distribution and source of noise. Their structure connects an efficient encoder and a Bayesian decoder, respectively constrained by the principle of efficient coding and Bayesian inference. Efficient coding formulates a notion of an optimal encoder, which allocates coding resources according to the environmental statistics of stimuli to maximize information transmitted under limited resources (Barlow, 1960). Bayesian inference formulates a notion of an optimal decoder, which combines uncertain sensory input with prior expectations using Bayes' rule (Knill & Richards, 1996). In this encoding-decoding framework of perception, sensory measurement m of a stimulus θ is first encoded in neural activity, then decoded into an estimate $\hat{\theta}(m)$ of the stimulus (Fig. 1a).

Importantly, this explains biases *away* from prior expectation. In orientation perception, studies demonstrated a well-replicated "repulsive bias" away from cardinal orientations (0°, 90°) and towards oblique orientations (±45°) (de Gardelle et al., 2010; Tomassini et al., 2010), despite that cardinals appear more frequently in nature according to statistical image analyses (Girshick et al., 2011). Efficient Bayesian observer models decompose biases into two competing components, prior attraction and likelihood repulsion (Hahn & Wei, 2022; Wei & Stocker, 2015). The latter arises from efficient coding, which encodes sensory measurements with inhomogeneous fidelity, coding more frequent stimuli more precisely at the expense of less frequent stimuli. Formally, the encoder encodes sensory measurements, $m = F(\theta + \delta_s) + \delta_n$, where F is the cumulative distribution function of stimulus θ and generates an inhomogeneous sensory space containing m. This means small deviations from cardinals are more precisely discriminated than deviations from obliques. When decoded by a Bayesian decoder, this creates asymmetry in the likelihood function, with a long tail over low encoding fidelity regions (i.e., higher uncertainty) and short tail over high fidelity regions (i.e., lower uncertainty), the latter coinciding with prior peaks. Likelihood asymmetry generates a repulsion from prior peaks as deviations are more discriminable, which competes with prior attraction during decoding, sometimes resulting in net repulsion.

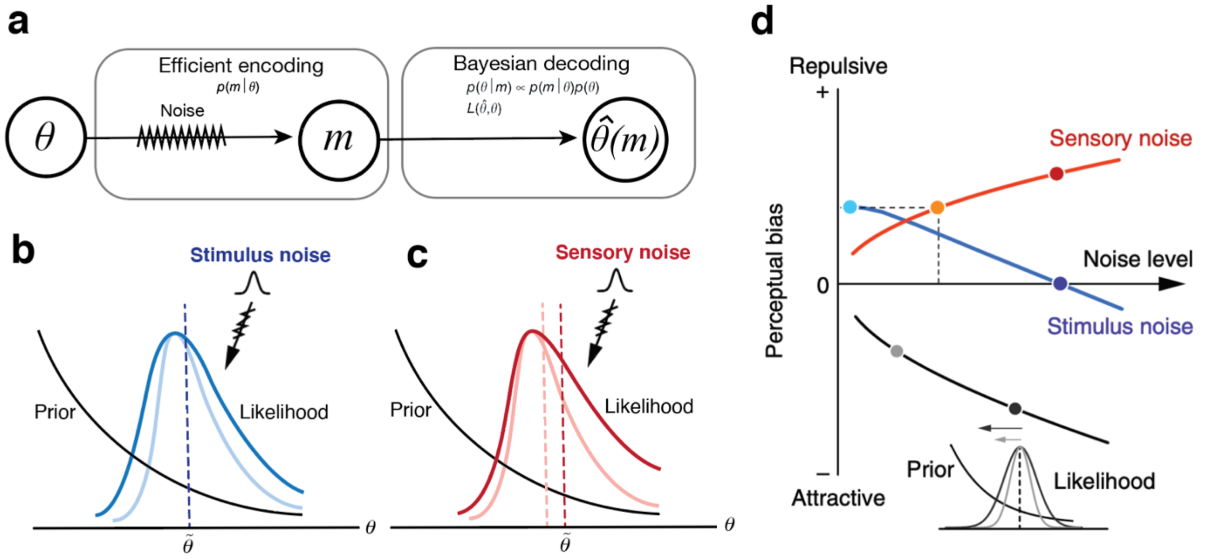

Fig. 1. Efficient Bayesian observer model | a. Perception is modelled as an encoding-decoding process. Encoding generates a sensory measurement m of stimulus θ under efficient coding constraint. This determines the likelihood function $p(m|\theta)$. Decoding is Bayesian and generates the percept $\hat{\theta}(m)$ from the posterior distribution and the loss function. **b.** Stimulus noise symmetrically increases likelihood width, which increases the relative prior attraction because a wider likelihood is weighted less in Bayesian decoding. Thus overall, bias becomes more attractive. **c.** Sensory noise increases likelihood width and asymmetry, thus increases likelihood repulsion from the prior peak, leading to an increase in repulsive bias. **d.** Summary of the effects of sensory and stimulus noise during encoding on the direction and magnitude of bias taken from Wei and Stocker (2015). This figure was produced by the candidate based on Wei and Stocker (2015).

A critical relationship is the dependence of bias on the source of noise. The inhomogeneous sensory space has different metrics than the stimulus space (analogous to how the motor homunculus is a distorted map of the body). This enriches noise composition: it differentiates noise introduced into the sensory space (sensory noise δ_n) and the stimulus space (stimulus noise δ_s). Sensory and stimulus noise differentially affect the likelihood, creating opposite changes in bias (Fig. 1d). Sensory noise increases likelihood asymmetry as noise accumulates heterogeneously in sensory space. This increases likelihood repulsion, hence increases repulsive bias (Fig. 1c). Stimulus noise symmetrically increases likelihood width. A broader likelihood is weighted less during decoding, increasing relative prior attraction, hence repulsive bias decreases and becomes more attractive (Fig. 1b). Predictions of noise-dependent bias variations received empirical support. Reducing stimulus presentation time (de Gardelle et al., 2010) and contrast (Girshick et al., 2011), which were thought to reduce sensory neural activity hence increase the effects of sensory noise, increased repulsive bias. Increasing variance of Gabor patch orientations, hypothesized to increase stimulus noise, increase attractive bias (Tomassini et al., 2010).

A limitation is that this encoding-decoding framework only considers noise during encoding. Noise can occur at many stages of the perceptual decision process, one of which is while maintaining information in WM. A delay between sensory input and behavioral decision requires WM to maintain sensory representations against accumulating neural noise. One successful model of WM maintenance is continuous attractor networks (Compte, 2006), which model WM representations as sustained neural population activities of recurrent neural connections. Noise affecting WM representations is modelled as a stochastic diffusion of the active region across the network, causing stochastic changes in the stored feature value, while keeping total population activity constant (Burak & Fiete, 2012; Compte et al., 2000). Diffusion model of WM noise better reproduced human recall errors and latencies than other models (Taylor & Bays, 2018). However, WM neural activity noise is not commensurable with noise in efficient Bayesian observer models. It is unknown how they correspond or how WM fits into efficient Bayesian observer models.

We investigated how noise during WM maintenance affects bias in orientation perception, using a modified analogue recall task (Wilken & Ma, 2004) with a variable delay between stimulus and response. Our working hypothesis was that WM could maintain either the sensory measurement m from the encoder (Fig. 2a) or the decoded percept $\hat{\theta}(m)$ from the decoder (Fig. 2b) over delay. Noise during delay would correspond either to encoding noise affecting m or an estimation noise affecting $\hat{\theta}(m)$, the latter not yet incorporated into Bayesian observer models. This gave neat qualitative predictions on bias. If WM stores m, delay noise could be (i) sensory noise, which predicts an increase in repulsive bias with longer delay, or (ii) stimulus noise, which predicts a decrease in repulsive bias with longer delay. If WM stores $\hat{\theta}(m)$, delay noise would be (iii) an estimation noise on $\hat{\theta}(m)$, which is a scalar point estimate, typically the mean or mode of the posterior distribution. During delay, this point would randomly fluctuate, predicting no net change in bias with delay. Secondly, we varied set size (number of memory items) as another noise manipulation. Studies on WM capacity have interpreted set size as affecting the relative weight of internal noise, but this is also unintegrated with efficient Bayesian observer models. We extended our hypotheses about delay also to set size.

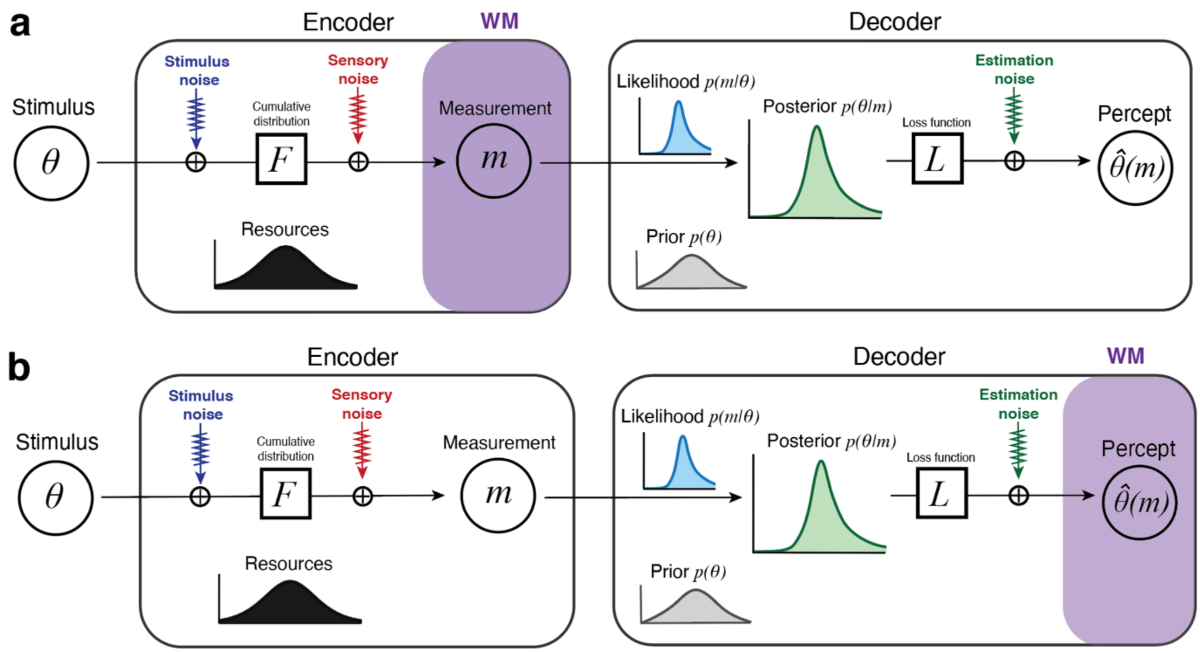

Fig. 2. Two hypotheses on how WM integrates into efficient Bayesian observer models. | a. The conventional view that WM occurs after sensory encoding to maintain the sensory measurement m, which is then decoded before making a response. **b.** An early inference model of WM and perception, in which decoding occurs immediately after sensory encoding and the decoded percept $\hat{\theta}(m)$ is maintained in WM over delay. This figure was produced by the candidate.

Methods

Subjects

10 subjects (aged 18-44) with normal or corrected-to-normal vision participated after giving informed consent in accordance with the Declaration of Helsinki.

Stimuli and apparatus

Orientation stimuli θ were sampled from a bimodal distribution peaking at cardinal orientations (0°, 90°)

$$p(\theta) = 2 - |\sin\theta|,$$

which well-approximated orientation statistics in nature based on Fourier spectrum analysis of natural images (Girshick et al, 2011). Subjects were seated 60 cm before an LCD monitor with their head supported by head rest. Eye movements were monitored by an infrared eye-tracker (Eyelink 1000, SR Research). Each stimulus was positioned randomly at one of eight equidistant positions from center. Orientations were evenly binned into 25 bins across [–90°, 90°) for analysis.

Design

The experiment used a 2 (set size: 1 or 3 items) x 2 (delay: 1 or 7 s) design. Each subject completed 700 trials, divided into 14 blocks of 50 trials. Set size and delay were randomly varied across trials and balanced within each block. The experiment contained two tasks, an analogue recall task and a reward confidence decision (Honig et al., 2020), which generated two datasets, response error and confidence.

Psychophysical procedure

In each trial, subjects fixated on a central annulus. Once fixated, a memory array of 1 or 3 oriented bars appeared for 500 ms, then disappeared for a delay period of 1 or 7 s, during which participants must remember all orientations. Afterwards, one bar was randomly cued by an annulus circling its previous location. Subjects reproduced the orientation of the cued bar from memory using the mouse.

Next, subjects drew a “confidence arc” symmetrically around their orientation estimate. If the true orientation fell within the arc, subjects were rewarded with points (0–100), which inversely correlated with arc size A (0–180°) (points $= \frac{180-A}{180} \times 100$); otherwise, they received zero points. Subjects were thus incentivized to draw the smallest arc that they believed contains the true orientation. Afterwards, subjects were shown the true orientation, points earned that trial, and total points across trials, which offered feedback on their response accuracy.

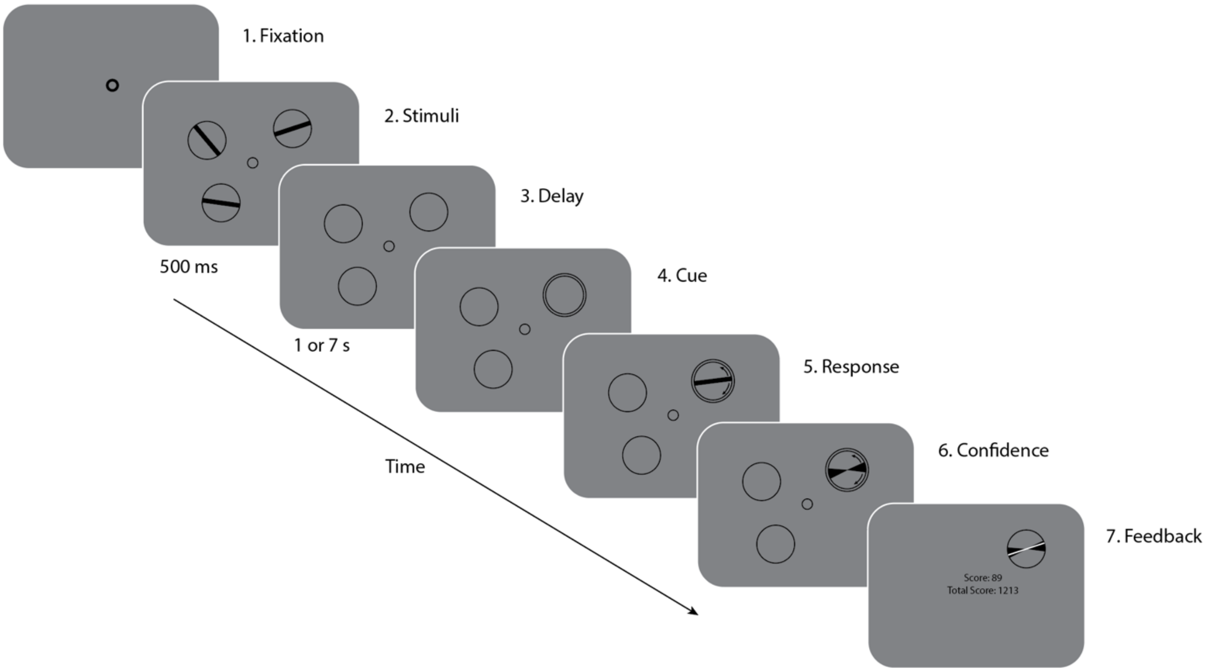

Fig. 3. Task procedure | Analogue orientation report task containing 7 stages. This figure was produced by the candidate.

Data analysis

We recorded recall response error as the absolute deviation between reported and true orientation, $\Delta\hat{\theta}_j = |\hat{\theta} - \theta_i|$, and calculated the mean absolute deviation (MAD) for each condition. We then calculated two measures: response bias as the circular mean of response error, and response variability as the circular SD of response error. We calculated confidence as (180° – arc size) normalized to [0, 1].

Bayesian hypothesis testing. We used two-way repeated-measure Bayesian ANOVAs in JASP to test delay and set size effects on behavioral measurements. Bayesian statistics quantify evidence for and against different models. When reporting Bayesian ANOVAs, we reported the Bayes inclusion factor (BF_{incl}) as overall evidence for an effect across all candidate models. The prior probability of a model is the proportion of models that include an effect (e.g., if 3 out of 5 models contain effect A, prior probability is 0.6). BF_{incl} quantifies the change from the prior to the posterior probability of all models that include an effect (e.g., set size or delay) once data is observed. For example, BF_{incl} = 4 means that models containing the effect are four times more likely given data than without data. We interpreted BF using Lee & Wagenmakers (2014)'s criteria:

Bayes factor (BF)	Interpretation
BF ≤ 1	No evidence
1 < BF < 3	Anecdotal evidence
3 ≤ BF < 10	Moderate evidence
10 ≤ BF < 30	Strong evidence
30 ≤ BF < 100	Very strong evidence
BF ≥ 100	Extreme evidence

Table 1. Lee & Wagenmakers (2014)'s criteria for interpreting Bayes factors (BF).

Parametric contamination model. We fitted a parametric contamination model (Kennedy et al., 2017) to response errors to obtain parametric estimates of bias and variability, which were used for Bayesian ANOVAs. The model assumes that data contain a mixture of noise and target responses, using two distributions to model each component. It assumes that subject responses follow a von Mises distribution (a circular analogue to the normal distribution) and noise follows a uniform distribution. This was not intended to model meaningful psychological processes (e.g., inattention or neural spiking stochasticity), but only for robust parameter estimation against noise. We fitted response error $\varDelta\hat{\theta}_j$ on every trial j to the distribution modified from Taylor & Bays (2018),

$$p\left(\varDelta\hat{\theta}_j \middle| \theta_i, \eta, \lambda, \alpha, \beta\right) = \ \lambda \times \ VM\left(\hat{\theta}; \mu(\theta_i), \kappa(\theta_i)\right) + \frac{1-\lambda}{2\pi},$$

where VM is the von Mises density function, λ is the mixture weight. We let bias vary as the mean of the von Mises distribution using parameter $\mu(\theta_i) = \eta\sin(2\theta_i)$, which lets bias vary sinusoidally against orientation. The bias parameter η returns bias magnitude and direction, with positive η expressing repulsive bias from cardinals and negative η expressing attractive bias. We let variability vary as the SD of the von Mises distribution using parameter $\kappa(\theta_i)$, calculated as a function of $\beta exp(-\alpha\cos(2\theta_i))$. This lets variability vary cosinusoidally against orientation with amplitude α and baseline β in the positive value domain. We fitted the model to obtain maximum log-likelihood parameter values.

Confidence was estimated as values fitted to $cos(2\theta)$ that minimized mean squared error, because contamination model assumes that target responses report the correct value and there is no "correct" confidence.

Kernel smoothing for visualization. To better visualize data across orientations, we estimated bias, variability and confidence using a smoothing kernel for 50 evenly spaced orientation bins. At each of the 50 points (bin centers), we calculated the weighted circular mean and SD of response error data, and weighted mean of confidence data, with weights given by a von Mises kernel. These estimates recovered a nonparametric smooth function from raw data. Smoothed data were used in Fig. 5b, 6b, 7b.

Results

We measured people's ability to reproduce orientation stimuli over delays using an analogue report task and their ability to assess their uncertainty using a rewarded confidence decision. We analyzed error distributions, bias, variability and confidence across orientations, delay and set size conditions.

Error distributions showed positive kurtosis that decreased with greater delay and set size. Bayesian ANOVAs found extreme evidence for mean response error (MAD) to increase with set size (BF_{incl} = 14,163) and delay (BF_{incl} = 362), with strong evidence for an interaction (BF_{incl} = 8.19). The distribution of response errors across trials showed positive kurtosis (i.e. leptokurtic), with sharper peaks and longer tails than normal distribution as previously documented (Bays, 2014) (Fig. 4). This became broader and flatter with increase in both set size and delay, reflecting both greater mean and variance of errors.

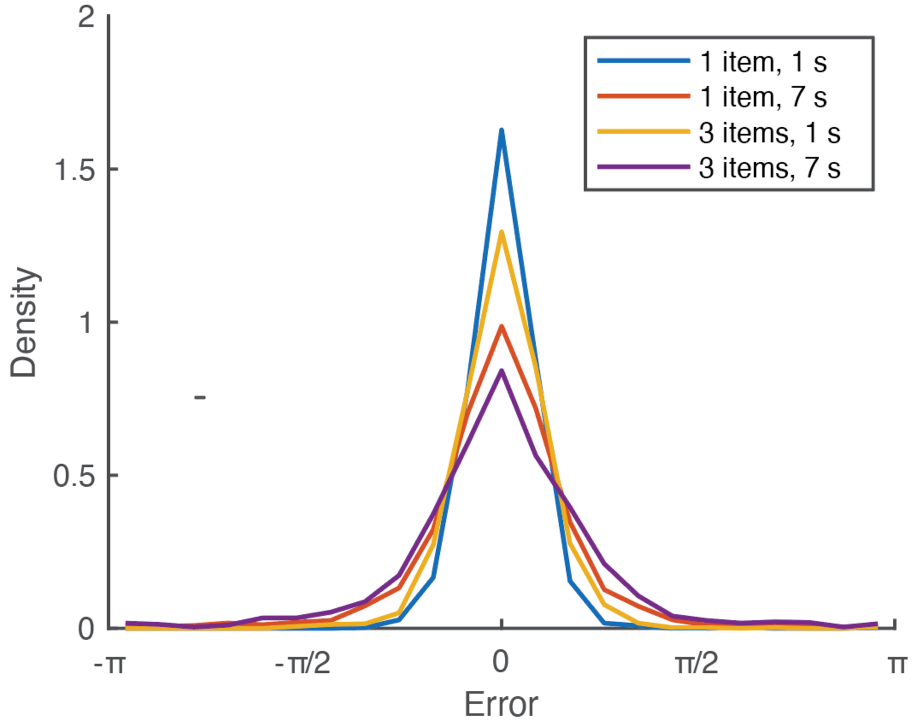

Fig. 4. Response error distributions across set size and delay | Response error distributions showed positive kurtosis that became flatter and broader with greater set size and delay. This figure was produced by the candidate.

Repulsive bias did not change with delay and set size variations.

Our crucial finding was that there was no evidence for delay or set size to affect bias magnitude (delay: BF_{incl} = 0.281; set size: BF_{incl} = 0.398), based on Bayesian ANOVAs on fitted bias parameters from the contamination model. Fig. 5a confirmed lack of change in bias graphically. Across orientations, we found a repulsive bias away from cardinals robustly across all experimental conditions (Fig. 5b). This manifested as a characteristic sinusoidal function, showing that orientations slightly larger than cardinals are perceived with a positive bias in the same direction, repelling perception further from cardinals, vice versa in the other direction.

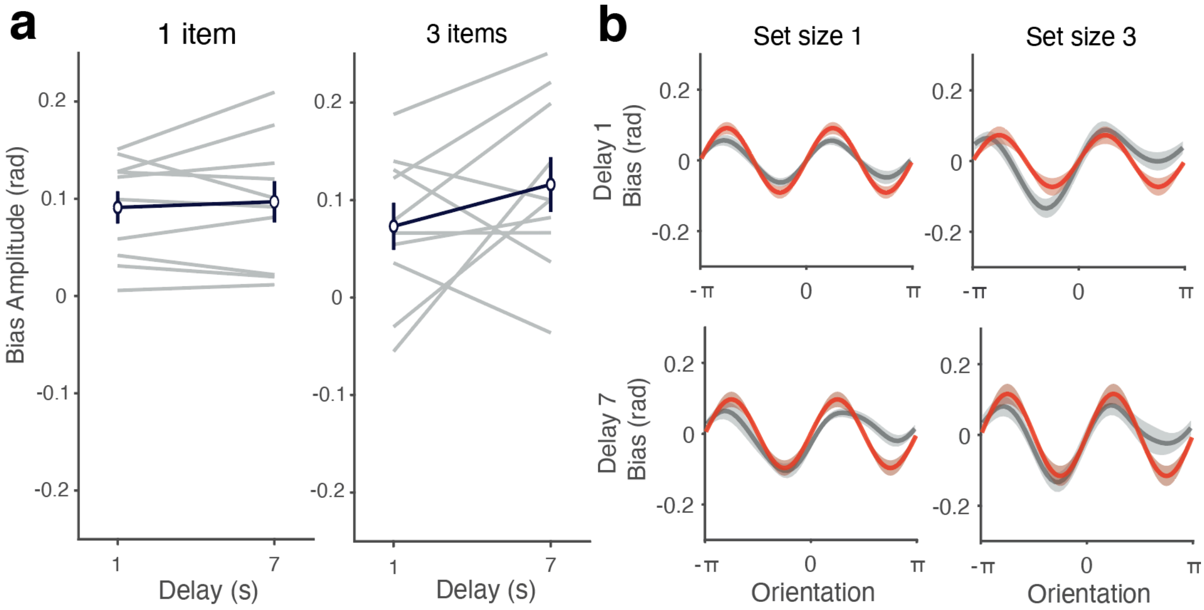

Fig. 5. Bias variations across set size and delay | a. Bias amplitude fitted from the contamination model. Grey lines connected within-subject data. Blue lines were predictions of the best-fitting Bayesian ANOVA model with 95% confidence interval (vertical bars). **b.** Orientation-specific variations in bias. Red lines were fitted bias amplitude parameter from the contamination model. Grey lines were the mean observed bias (circular mean response error). This was produced by the candidate and the supervisor.

Variability increased with delay and set size but showed no orientation-specific variations. We found extreme evidence for an increase in observed variability (circular SD) of response errors with greater set size (BF_{incl} = 30,588) and delay (BF_{incl} = 451), with moderate evidence for an interaction (BF_{incl} = 7.54). Fig. 6a displayed this variability increase across conditions graphically. We checked this against fitted variability baseline β and amplitude α parameters from the contamination model. Baseline variability reflected overall variability similar to circular SD, while variability amplitude reflected the degree of fluctuations about baseline across orientations. We found extreme evidence for increase in baseline variability with set size (BF_{incl} = 1417) and delay (BF_{incl} = 94.6), although weak evidence for interaction (BF_{incl} = 1.81). In contrast, there was strong evidence for decrease in variability amplitude with set size (BF_{incl} = 11.3) but not delay (BF_{incl} = 0.607).

Across orientations, we observed higher baseline variability for larger set size (left to right panels) and delay (top to bottom panels) confirming baseline variability ANOVA, although no visible decrease in orientation-specific fluctuations for bigger set size (Fig. 6b). Variability stochastically varied across orientations with no recognizable pattern. We did not find an oblique effect (Girshick et al., 2011), in which memory precision peaks at the cardinals and decreases towards the obliques. As variability can be considered as inverse precision, an analogous effect predicts variability to trough at cardinals and increase towards obliques.

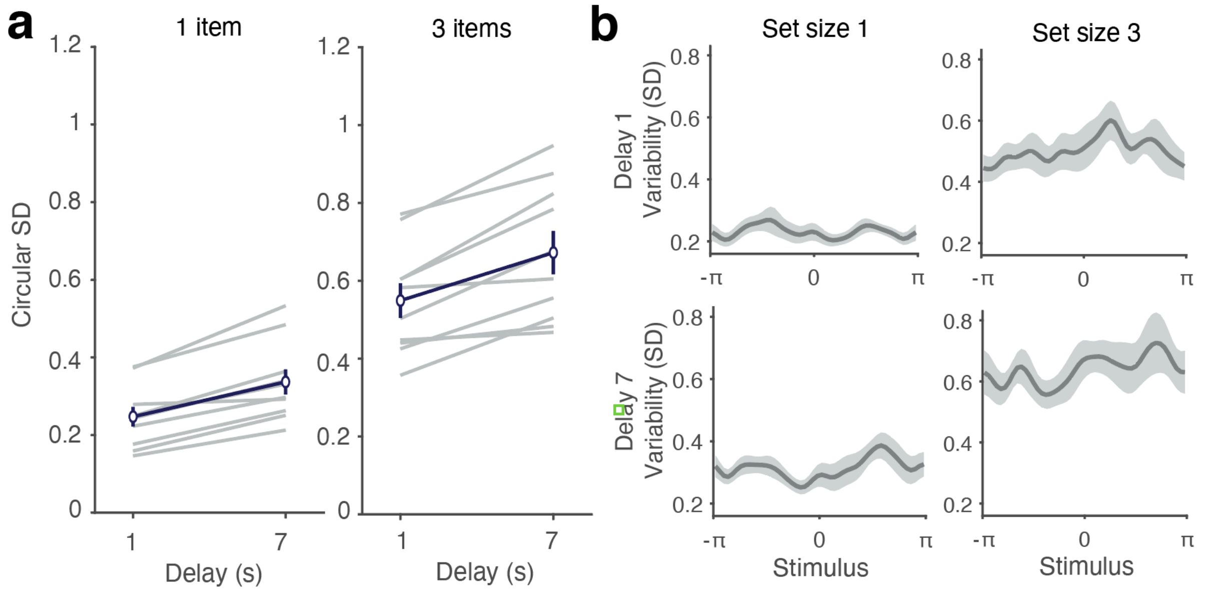

Fig. 6. Variability variations across set size and delay | a. Observed variability (circular SD) in response errors. Grey lines connected within-subject data. Blue lines were predictions of the best-fitting Bayesian ANOVA model with 95% confidence interval (vertical bars). **b.** Orientation-specific variations in variability. Grey lines were the mean variability (circular SD) in response errors. This figure was produced by the candidate and the supervisor.

Confidence increased with delay and set size, showed oblique effect for low set size, and reflected with trial-by-trial memory accuracy.

Confidence (normalized (180–arc size)) provided a measure of uncertainty in single-trial memory. We found extreme evidence for lower confidence with greater set size (BF_{incl} = 2087) and delay (BF_{incl} = 78.7), consistent with increased variability, but weak evidence for interaction (BF_{incl} = 1.60). Despite being a subjective measure, this increase was observed with high consistency across individual observers (Fig. 7a, grey lines rarely crossed). Across orientations, although no orientation-specific patterns were seen for set size 3, we observed a small oblique effect for set size 1. Confidence showed a modest peak at cardinals and dipped at obliques (Fig. 7b, left panels). Smaller oblique effect with larger set size was also consistent with literature (Girshick et al., 2011; Taylor & Bays, 2018).

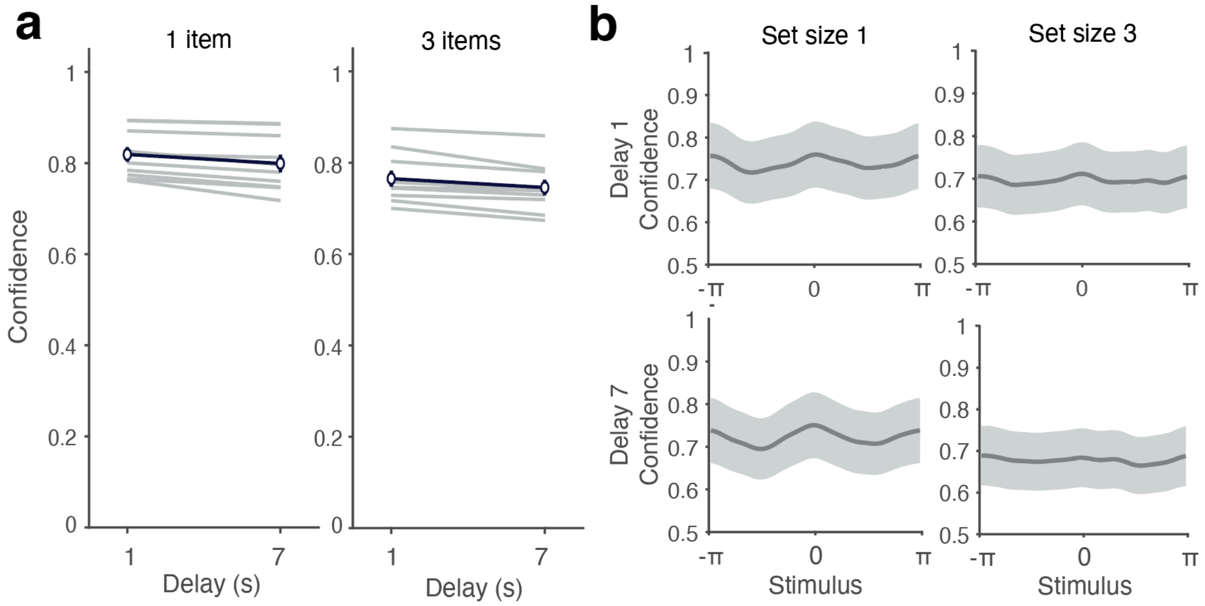

Fig. 7. Confidence variations across set size and delay | a. Confidence (normalized 180–arc size). Grey lines connected within-subject data. Blue lines were predictions of the best-fitting Bayesian ANOVA model with 95% confidence interval (vertical bars). **b.** Orientation-specific variations in confidence. Grey lines were the mean variability (circular SD) in response errors. This figure was produced by the candidate and the supervisor.

To determine whether confidence reflected real memory and performance quality, we analyzed the correlation between confidence arc size and absolute response error for every trial across subjects. We found a linear error-arc size relationship on a single-trial level (Fig. 8), indicating that people were more confident about their estimate when they were actually being more accurate. This suggested that people had single-trial knowledge of memory uncertainty.

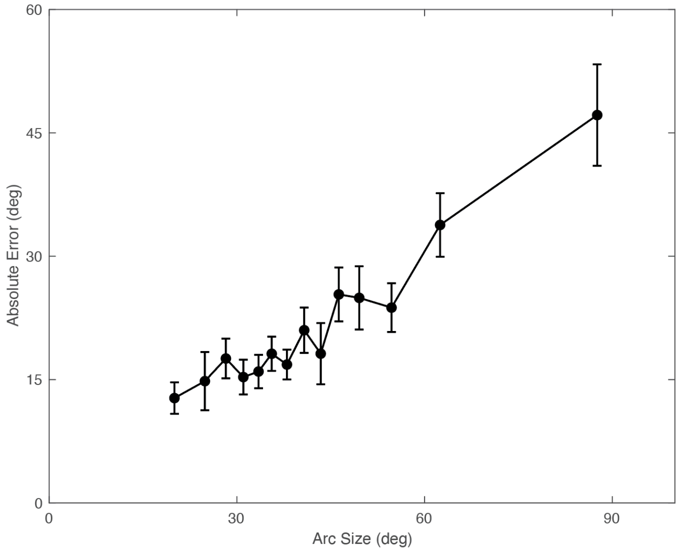

Fig. 8. Confidence arc size and absolute error correlation | Arc size and absolute error showed linear correlation. For each subject, data across 700 trials were separated into 14 quantiles along arc size, and mean arc size and absolute error were calculated per quantile. The mean (±SEM) across subjects for each quantile was plotted. This figure was produced by the candidate.

Discussion

The present work investigated how noise from set size and WM delay affected bias in orientation perception. We found that perception showed leptokurtic error distributions and repulsive bias from cardinals, consistent with literature. Bias across delay had important implications on how WM interfaces with efficient Bayesian observer models. Here, we conjectured model integrations that potentially explain these patterns. We compared two measures of memory uncertainty–response error variability and confidence–for assessing Bayesian predictions. Finally, we considered how present work informs open questions about WM representations.

Bias gives evidence for an "early inference" model of perception and WM

Our key finding was that repulsive bias did not change with delay variations ($\mathrm{BF_{incl}}$ = 0.281). This indicated that noise during delay was not equivalent to sensory or stimulus noise in efficient Bayesian observer models, as this predicted an increase or decrease in repulsive bias respectively. Because sensory and stimulus noise affected the sensory measurement m, this also implied that WM did not maintain a representation of m over delay. Instead, our result was consistent with delay introducing an estimation noise after decoding that affected only the decoded percept $\hat{\theta}(m)$, which are scalar point estimates of the stimulus. Assumed that WM storage of any scalar has equal fidelity, stored scalars exist in a homogenous stimulus space and estimation noise amounts to random diffusion of the neural activities representing scalars. This predicts no net change in bias with delay duration, consistent with our findings. Our findings were supported by unpublished results from Tomić et al (2022), who also found constant repulsive bias across delay in a similar task design. However, because they sampled stimulus from a uniform distribution, a confounding explanation was that the prior used by the Bayesian decoder rapidly adapted to this distribution (Adams et al., 2004), creating a flat prior. In this case, if delay increased stimulus noise, which would reduce the weight of likelihood repulsion by symmetrically widening the likelihood, a flat prior would also eliminate prior attraction, thus potentially cause no net change in bias. By sampling from a naturalistic bimodal distribution, we eliminated this possibility as rapid adaptation to a bimodal distribution would still preserve prior attraction.

This challenges the predominant view of how WM integrated into perception, which proceeded in the of order of sensory encoding, WM maintenance of m, then decoding m into $\hat{\theta}(m)$ (Ma et al., 2014). In contrast, our bias result suggested an early inference model of perception and WM in the order of sensory encoding, rapid decoding m into $\hat{\theta}(m)$, and WM maintenance of $\hat{\theta}(m)$ (Fig. 2b). However, this is conjectural as strong evidence for this model can only be obtained by quantitative model fits.

Surprisingly, we found that repulsive bias did not change with set size variations ($\mathrm{BF_{incl}}$ = 0.398). This implied set size, like delay, did not affected m as an encoding noise, but affected $\hat{\theta}(m)$

as an estimation noise. Notably, our findings contradicted results from Taylor and Bays (2018), who reported greater repulsive bias with set size in an analogue recall task without delay. They modelled set size as increasing the relative effects of sensory noise by dividing the population activity *before* decoding, thereby lowering signal-to-noise ratio (SNR). However, their bias result was a qualitative read-off of model fits, which showed only modest bias increase (see their Figure 3) and was assessed holistically not for bias specifically. Nevertheless, this is an important inconsistency. It shows that set size effects on perceptual bias are not well-understood by current theory, and previous models at best capture set size effects under limited conditions (e.g., without memory delays). Based on our results, we conjecture that set size effects might be misunderstood by not asking the *content* they affect. Different psychological and neural descriptions of set size effects on WM can be summarized as reducing SNR because they all revolve around the idea of *dividing* WM "resources" or population activities, i.e., weakening the signal. However, signal division is a general operation that can apply to any signal regardless of content. The efficient Bayesian framework stipulates specific, distinct representations and their connecting functions, in which noise of the same structure (e.g., Gaussian) can have differential effects by acting on different representations or occurring inside/outside a function transformation. Therefore, set size could perform signal division at a specific stage of processing on specific representational content, like the decoded percept.

A possible mechanism for set size effects within a diffusion model of WM maintenance is that set size could scale diffusion rate. Faster diffusion would cause larger response variability but no change in bias as observed here. Several attractor models have used set-size dependent diffusion rates in the past (Koyluoglu et al., 2017). Because we observed an interaction between set size and delay in response error magnitude ($BF_{incl} = 8.19$), diffusion speed could further include an acceleration term as a function of delay duration to model this effect.

Confidence could be a more sensitive measure of uncertainty and stimulus-specific effects

We tested efficient coding predictions on memory uncertainty with two measures, response error variability and confidence. We found that variability increased and confidence decreased with set size and delay, reflecting higher memory uncertainty caused by increased noise.[1] Across orientations, we tested for an oblique effect (Girshick et al., 2011), where memory precision is better at cardinals than obliques. This is predicted by efficiently encoding cardinals with higher fidelity/precision, measured as the size of just-noticeable-differences (JND) between stimuli in discrimination tasks. During decoding, this produces narrower likelihoods, whose width determines memory uncertainty. Variability varied stochastically without an oblique effect, whereas confidence showed a modest oblique effect for low set size, peaking at cardinals and troughing at obliques. This suggested that confidence might be

[1] Uncertainty and noise are not equivalent. Noise is information subtracted from the signal (or negative entropy), whereas uncertainty is a signal encoding probability. However, they are related in that if noise degrades the signal in the encoder, the decoder can assign a higher uncertainty to the sensory input.

a more sensitive measure of memory uncertainty, fit to test predictions on stimulus-specific effects. There are several reasons that variability might be less sensitive. First, uncertainty exists for individual memories *every trial*, whereas variability measures of spread of response errors *across trials*. This assumes that if single-trial memory uncertainty is high, responses over trials will also be more variable. In contrast, confidence is a direct trial-by-trial measure of uncertainty. Second, variability measures variance, which is a second statistical moment. This requires significantly more data for reliable values, hence is particularly vulnerable to trial number. Ours (N = 700) was significantly lower than others investigating variability (N = 7,200 in Girschirk et al., 2015; N = 4,315 in Tomassini et al., 2011). Hence, confidence might be a better proxy for JND in analogue report tasks than variability, although testing this would require side-by-side comparisons of confidence and precision results.

Implications on WM representations

One open question is where WM representations are localized anatomically. A traditional view is that the sensory neural population encoding the stimulus also maintains WM through sustained activity (Pasternak & Greenlee, 2005). However, if WM does not store sensory measurements but products of decoding, which require additional processing, WM is likely be underpinned by separate neurons than sensory neurons, although they might coexist in the same population. This reinforces the idea that WM is more complex than the passive persistence of sensory neuron activities (Bays et al., 2022).

Another open question is whether WM represent uncertainty as full probability distributions or point estimates. The confidence arc size-error correlation we found indicated that people had trial-to-trial representation of memory uncertainty and that this uncertainty was used for perceptual decisions. However, a striking conclusion from our bias findings was that WM did not store the likelihood function or the posterior distribution. Instead, WM could store scalar summary statistics of the likelihood or posterior, like width, to store trial-to-trial uncertainty. This opposes the prominent idea of probabilistic population coding (PPC) (Pouget et al., 2000), proposing that population activity naturally encode full probability distributions with complete information about uncertainty. While our evidence on WM representation was indirect, it warns against simple generalization of PPC and suggests the possibility of multiple representation schemes of uncertainty used by different systems or for different tasks.

Acknowledgements

I’d like to thank my lab partner; Dr. Ivan Tomić for his fantastic supervision and MATLAB codes for analysis; and Prof. Paul Bays for giving many clarifications.

References

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the ‘light-from-above’ prior. *Nature Neuroscience*, *7*(10), Article 10. https://doi.org/10.1038/nn1312

Barlow, H. B. (1960). Possible Principles Underlying the Transformations of Sensory Messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 216–234). The MIT Press. https://doi.org/10.7551/mitpress/9780262518420.003.0013

Bays, P. M. (2014). Noise in Neural Populations Accounts for Errors in Working Memory. *Journal of Neuroscience*, *34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P., Schneegans, S., Ma, W. J., & Brady, T. (2022). *Representation and computation in working memory*. PsyArXiv. https://doi.org/10.31234/osf.io/kubr9

Burak, Y., & Fiete, I. R. (2012). Fundamental limits on persistent activity in networks of noisy neurons. *Proceedings of the National Academy of Sciences*, *109*(43), 17645–17650. https://doi.org/10.1073/pnas.1117386109

Compte, A. (2006). Computational and in vitro studies of persistent activity: Edging towards cellular and synaptic mechanisms of working memory. *Neuroscience*, *139*(1), 135–151. https://doi.org/10.1016/j.neuroscience.2005.06.011

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X.-J. (2000). Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cerebral Cortex*, *10*(9), 910–923. https://doi.org/10.1093/cercor/10.9.910

de Gardelle, V., Kouider, S., & Sackur, J. (2010). An oblique illusion modulated by visibility: Non-monotonic sensory integration in orientation processing. *Journal of Vision*, *10*(10), 6. https://doi.org/10.1167/10.10.6

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

Hahn, M., & Wei, X.-X. (2022). *A unifying theory explains seemingly contradicting biases in perceptual estimation* (p. 2022.12.12.519538). bioRxiv. https://doi.org/10.1101/2022.12.12.519538

Honig, M., Ma, W. J., & Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. *Proceedings of the National Academy of Sciences*, *117*(15), 8391–8397. https://doi.org/10.1073/pnas.1918143117

Kennedy, L. A., Navarro, D. J., Perfors, A., & Briggs, N. (2017). Not every credible interval is credible: Evaluating robustness in the presence of contamination in Bayesian data analysis. *Behavior Research Methods*, *49*(6), 2219–2234. https://doi.org/10.3758/s13428-017-0854-1

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian Inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511984037

Koyluoglu, O. O., Pertzov, Y., Manohar, S., Husain, M., & Fiete, I. R. (2017). Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity. *ELife*, *6*, e22225. https://doi.org/10.7554/eLife.22225

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), Article 3. https://doi.org/10.1038/nn.3655

Olkkonen, M., McCarthy, P. F., & Allred, S. R. (2014). The central tendency bias in color perception: Effects of internal and external noise. *Journal of Vision*, *14*(11), 5. https://doi.org/10.1167/14.11.5

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), Article 2. https://doi.org/10.1038/nrn1603

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, *1*(2), Article 2. https://doi.org/10.1038/35039062

Sun, J., & Perona, P. (1998). Where is the sun? *Nature Neuroscience*, *1*(3), Article 3. https://doi.org/10.1038/630

Taylor, R., & Bays, P. M. (2018). Efficient Coding in Visual Working Memory Accounts for Stimulus-Specific Variations in Recall. *The Journal of Neuroscience*, *38*(32), 7132–7142. https://doi.org/10.1523/JNEUROSCI.1018-18.2018

Tomassini, A., Morgan, M. J., & Solomon, J. A. (2010). Orientation uncertainty reduces perceived obliquity. *Vision Research*, *50*(5), 541–547. https://doi.org/10.1016/j.visres.2009.12.005

Tomić, I., Gironés, Z., & Bays, P. M. (2022). Unpublished data.

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain ‘anti-Bayesian’ percepts. *Nature Neuroscience*, *18*(10), 1509–1517. https://doi.org/10.1038/nn.4105

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11. https://doi.org/10.1167/4.12.11