Visual working memory models of delayed estimation do not generalize to whole-report tasks

Benjamin Cuthbert[1] ✉, **Dominic Standage**[1,2], **Martin Paré**[2,3], **Gunnar Blohm**[1,2,3,4,5]

[1]Centre for Neuroscience Studies, Queen's University, Kingston, Canada; [2]Department of Biomedical and Molecular Sciences, Queen's University, Kingston, Canada; [3]Department of Psychology, Queen's University, Kingston, Canada; [4]School of Computing, Queen's University, Kingston, Canada; [5]Department of Mathematics and Statistics, Queen's University, Canada

Abstract

Whole-report working memory tasks provide a measure of recall for all stimuli in a trial, and afford single-trial analyses that are not possible with single-report delayed estimation tasks. However, most whole-report studies assume that trial stimuli are encoded and reported independently, and do not consider the relationships between stimuli presented and reported within the same trial. Here, we present the results of two independently conducted whole-report experiments. The first dataset was recorded by Adam, Vogel, and Awh, 2017, and required participants to report color and orientation stimuli using a continuous response wheel. We recorded the second dataset, which required participants to report color stimuli using a set of discrete buttons. We find that participants often group their reports by color similarity, contradicting the assumption of independence implicit in most encoding models of working memory. Next, we show that this behavior is consistent across participants and experiments when reporting color but not orientation, two circular variables often assumed to be equivalent. Finally, we implement an alternative to independent encoding where stimuli are encoded as a hierarchical Bayesian ensemble, and show that this model predicts biases that are not present in either dataset. Our results suggest that assumptions made by both independent and hierarchical ensemble encoding models—which were developed in the context of single-report delayed estimation tasks—do not hold for the whole-report task. This failure to generalize highlights the need to consider variations in task structure when inferring fundamental principles of visual working memory.

✉ **For correspondence:** b.cuthbert@queensu.ca (Benjamin Cuthbert)

Present address: Centre for Neuroscience Studies, 18 Stuart Street, Kingston, ON, K7L 3N6, Canada

Data availability: Data and code required to reproduce this entire paper are available on GitHub.

Acknowledgements: We would like to sincerely thank K Adam, E Vogel, and E Awh for sharing their experimental data publicly and therefore for making this project possible.

Funding: Financial support for this project was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Foundation for Innovation (CFI).

Competing interests: The authors declare no competing interests.

Introduction

Many recent models of visual working memory (VWM) were developed in the context of the delayed estimation task (Wilken and Ma, 2004), where participants report the value of a presented stimulus—such as the color of a square or orientation of a bar—after a blank delay period. In this task, report error is measured by taking the difference between presented and reported values, and error distributions over many trials are assumed to reflect the structure of stimulus encodings. Models that reproduce delayed estimation error distributions have therefore been used to support and refute numerous claims about VWM encoding (Zhang and Luck, 2008; Bays and Husain, 2008;

T. F. Brady and Alvarez, 2011; van den Berg et al., 2012; Swan and Wyble, 2014; Oberauer, 2017; Nassar, 2018; Schurgin, Wixted, and T. F. Brady, 2020).

Recently, debate over the existence of a fixed storage limit in VWM led Adam and colleagues to introduce a novel delayed estimation variant: the *whole-report task* (Adam, Mance, et al., 2015; Adam, Vogel, and Awh, 2017). Unlike typical delayed estimation tasks—where only one stimulus is reported—each trial of the whole-report task requires participants to report all presented stimuli. This has the advantage of measuring recall of the entire stimulus display, and allowed Adam et al. to investigate item limits at the single-trial level.

The whole-report task is now seeing more widespread use (Peters, Rahm, Czoschke, et al., 2018; Adam and Vogel, 2018; Killebrew et al., 2018; Peters, Rahm, Kaiser, et al., 2019; deBettencourt et al., 2019; Robison and Unsworth, 2019; Utochkin and T. F. Brady, 2020; Hao et al., 2021), and the original publicly-available dataset has been used for model evaluation (Schneegans, Taylor, and Bays, 2020). Critically, these studies treat the whole-report task as an extension of single-report delayed estimation, focusing on error distributions accumulated across trials and reports. Only a few studies report within-trial effects like inter-item interference and bias (Utochkin and T. F. Brady, 2020; Hao et al., 2021; Udale et al., 2021), and none consider the joint distribution of reports made within the same trial.

Within-trial joint distributions are a key affordance of the whole-report task because they allow us to test VWM model assumptions in ways that are not possible with single-report delayed estimation data. For example, many models assume that stimuli are encoded in VWM independently (Zhang and Luck, 2008; Bays and Husain, 2008; van den Berg et al., 2012; Fougnie, Suchow, and Alvarez, 2012; Schneegans, Taylor, and Bays, 2020). If this were the case, a given stimulus encoding would have no effect on other stimuli encoded within the same trial, and the joint distribution of reports would be uniform. Alternative models suggest that stimuli are encoded together in an "ensemble" (Alvarez and Oliva, 2009; T. F. Brady and Alvarez, 2011; Orhan and Jacobs, 2013; Nassar, 2018), which would result in dependencies in within-trial joint distributions.

Here, we investigate previously uncharacterized within-trial behavior in the whole-report task, and find that neither independent nor hierarchical ensemble encoding models can explain the results. We use two whole-report datasets: the original dataset recorded by Adam, Vogel, and Awh, 2017, and data from a whole-report task variant independently conducted in our lab (Cuthbert et al., 2018). In both datasets, we find strong evidence that color reports made in the same trial are not independent. In many cases, participants grouped consecutive within-trial reports by color similarity. In addition, reports made later in the trial—past the canonical capacity limit of 3-4 items—were consistently biased away from colors reported earlier in the trial.

We show that this effect is consistent across participants and set sizes when recalling color stimuli, but is either weaker or absent in task conditions with orientation stimuli. This is surprising, because color and orientation are typically treated as equivalent circular variables (van den Berg et al., 2012; Schneegans, Taylor, and Bays, 2020), and aggregate error distributions for color and orientation stimuli are qualitatively very similar.

Finally, we consider whether hierarchical ensemble encoding models might account for within-trial dependencies in color reports. In these models, dependencies arise due to Bayesian integration over a hierarchical prior (T. F. Brady and Alvarez, 2011; Orhan and Jacobs, 2013). We implement multiple Bayesian ensemble models, and find no evidence for any form of their predicted biases—even after restricting our analyses to specific set sizes or contexts where ensembles might confer a significant storage advantage.

Taken together, our results suggest that whole-report task data cannot be entirely accounted for by VWM models that assume independent encoding, equivalence between circular stimuli, or hierarchical ensemble encoding.

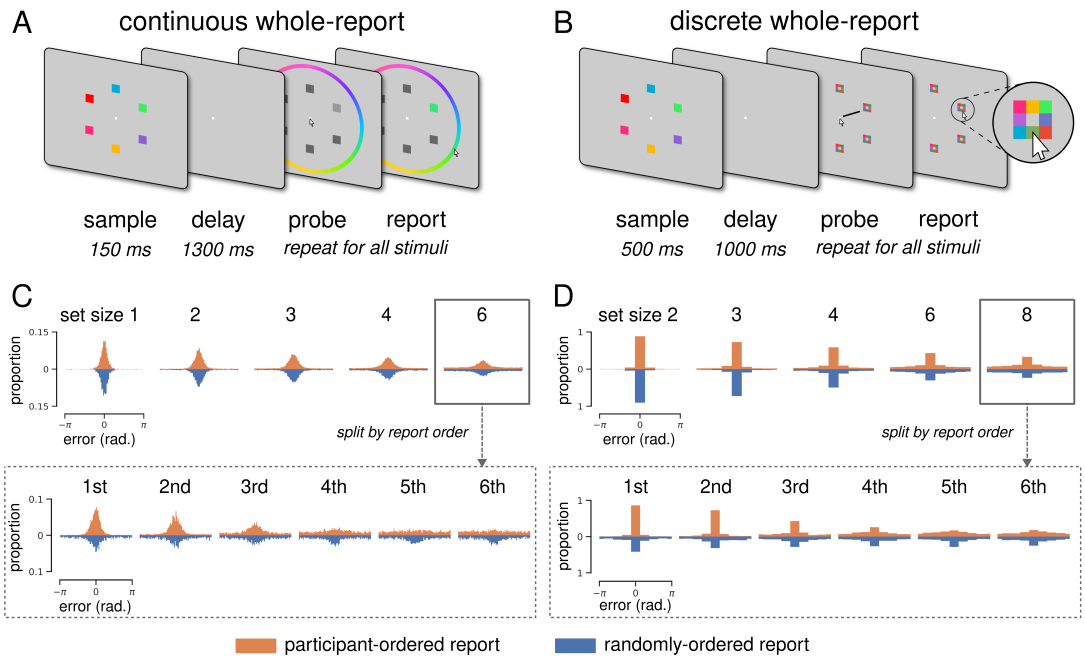

Figure 1. Whole-report task comparison. **A.** Continuous whole-report task. Colored squares were briefly presented, and after a blank delay period the participant reported all colors by clicking a color wheel. Report order was either participant-selected or randomly-generated. Stimuli were sampled from 360 "continuous" color values. **B.** Discrete whole-report task. Same as A, but participants reported by clicking a square in the response array (*inset*). Stimuli were sampled from 8 "discrete" color values. **C.** Continuous error distributions. 360 possible error values divided into 90 bins for visualization. Orange and blue histograms show errors from participant-selected and randomly-generated report conditions, respectively. *Upper*: All report errors, separated by set size. *Lower*: Set size 6 report errors, separated by report order. **D.** Discrete error distributions. Same as C, but bins correspond to all possible error values.

Methods

We analyzed two whole-report datasets, which we refer to as the *discrete whole-report* and *continuous whole-report* datasets.

Discrete whole-report dataset

The discrete whole-report dataset was recorded at Queen's University, and has been previously presented in conference proceedings (Cuthbert et al., 2018).

Task and participants

The discrete whole-report task was adapted from Adam, Mance, et al., 2015. 16 healthy adult participants (18-40 years of age; 7 female) completed two experimental sessions each, with 12 blocks of 30 trials completed per session. Participants were required to memorize an array of colored squares ("stimuli"), then report the color of each stimulus after a retention period.

Each trial began with the simultaneous presentation of 2, 3, 4, 6, or 8 stimuli for 500 ms. This was followed by a blank grey screen, presented for 1000 ms, during which participants were required to maintain fixation on the central point. Subjects were then presented with "response matrices," each comprised of 8 selectable buttons (one for each possible color; see Fig 1). All response matrices were identical within each trial, but button colors were randomized between trials. Response matrices appeared at all locations previously occupied by stimuli, and participants were required to report the color of each stimulus.

There were two report conditions (one session each): in the *participant-ordered* report condition, participants were instructed to report stimulus colors in order of confidence; in the *randomly-*

ordered report condition report order was random, and cued with a black radial line. Color reporting was unspeeded, and each trial concluded once all stimuli had been reported.

Experimental setup

Stimuli were generated with MATLAB R2016a (MATLAB, 2016), and presented on a VIEWPixx 3D LCD monitor (VPixx Technologies, Saint-Bruno, QC) using the Psychophysics Toolbox (Kleiner et al., 2007). Compliance during stimulus presentation and the retention period of the task was verified using the EyeLink 1000 Tower Mount (SR Research, Ottawa, ON). The sampling rate was set to 1000 Hz, and eye movement traces of position and velocity were examined for exclusion criteria. Trials were excluded if subjects did not maintain fixation during stimulus presentation (blinks not permitted) and the retention period (blinks permitted). In total, 771 trials were excluded (13.4%).

Stimulus generation

Colours were selected from 8 equidistant points on a circle in CIE $L^*a^*b^*$ colour space centered at (L^* = 70, a^* = 20, b^* = 38) with a radius of 60. Luminance and chromaticity was calibrated prior to each experimental session using a ColorCAL MKII colorimeter (Cambridge Research Systems Ltd., Rochester, UK), and the coordinates above were selected to ensure that all colors fell within the LCD monitor's gamut. Stimuli each measured 1.5 x 1.5 ° (visual angle), and were spaced equidistantly around a central fixation point with a 7.5 ° radius.

Continuous whole-report dataset

The continuous whole-report dataset was recorded by Kirsten Adam and colleagues at the Universities of Chicago and Oregon. The following is a brief summary of their methods, with particular emphasis on the differences between their task and the discrete whole-report task above. For complete experimental details, see Adam, Vogel, and Awh, 2017.

Task and participants

Here, we consider data from four task conditions recorded by Adam, Vogel, and Awh, 2017: participant-ordered report with color stimuli (22 participants), randomly-ordered report with color stimuli (17 participants), participant- ordered report with orientation stimuli (23 participants), and randomly-ordered report with orientation stimuli (21 participants). The continuous task used set sizes 1, 2, 3, 4, and 6 (vs 2, 3, 4, 6, and 8 for the discrete task). There were also timing differences; continuous task stimuli were presented for 150 ms (vs 500 ms), and the delay period lasted 1300 ms (vs 1000 ms).

Discrete vs continuous stimuli

While color stimuli for both tasks were sampled from equidistant points around circles in CIE $L^*a^*b^*$ color space, the discrete task only used 8 distinct colors while the continuous task used 360. The area of color space used was also different—the continuous task used a circle centered at (L^* = 54, a^* = 18, b^* = -8)—and monitors were not calibrated before the continuous experiment, so the displayed colors cannot be considered perceptually equivalent to discrete task stimuli. The continuous task also included conditions with orientation stimuli, which were sampled from 360 equally-spaced angles.

Discrete vs continuous report method

As illustrated in Figure 1, the two tasks employed different report methods. In the discrete task, 8 possible report colors were presented in the form of clickable buttons at the stimulus location being probed. In the continuous task, 360 possible report colors were presented in the form of a clickable color wheel surrounding the entire display. Similarly, the continuous task with orientation stimuli required participants to click on a surrounding (uncolored) wheel to report remembered orientation.

Analyses

Kullback-Leibler (KL) divergence estimation

To quantify within-trial dependence, we estimated the KL divergence between within-trial relative distance distributions and a circular uniform distribution. KL divergence is a measure of the difference beteen two distributions that involves computing their relative entropy (in bits). A low KL divergence between a discrete distribution P and a nominal uniform distribution Q is evidence that P has high entropy—and is close to uniformly distributed. A high KL divergence provides evidence that P is not uniformly distributed. This approach is useful because unlike correlation, it does not assume a linear relationship between variables, and can therefore capture higher-order dependencies such as those evident in Fig. 2.

For each participant, report condition, and pair of within-trial reports, we performed 1000 bootstrapped KL divergence estimates between the empirical distribution P and a nominal circular uniform distribution Q:

$$D_{KL}(P||Q) = \sum_{x=1}^{n} P(x) \log_2(\frac{P(x)}{Q(x)})$$

where n is the number of equal bins used to define the probability space ($n = 36$ for the continuous task and $n = 8$ for the discrete task).

To generate a baseline estimate, we repeated this process but replaced each empirical distribution P with a size-matched sample from a circular uniform distribution. Median estimates for all participants are shown alongside their respective baseline median estimates in Fig. 2C and D.

Hierarchical encoding models

Two-level Hierarchical Bayesian Model (HBM)

This model is adapted from the simplest model introduced by T. F. Brady and Alvarez, 2011. In the HBM implemented here, stimuli are probabilistically represented in working memory at two levels: (1) individual stimulus color estimates and (2) a joint "ensemble" of all stimuli.

Here, both color and orientation are considered circular stimulus spaces, so we substitute the original Gaussian distributions with the von Mises (circular normal) distribution:

$$VM(x|\mu, \kappa) = \frac{e^{\kappa cos(x-\mu)}}{2\pi I_0(\kappa)}$$

where $I_0(\cdot)$ is the modified Bessel function of order 0, μ is the mean circular direction, and κ is the concentration parameter (analagous to $1/\sigma^2$).

The HBM assumes that on each trial , N stimuli $\theta_{i=1}^{N}$ are sampled from the same "ensemble" von Mises distribution:

$$\theta_i|\mu_e, \kappa_e \sim VM(\mu_e, \kappa_e) \quad i = 1, ..., N$$

where μ_e and κ_e are the ensemble mean and precision with the following priors:

$$\mu_e \sim \mathcal{U}(-\pi, \pi)$$
$$\kappa_e \sim \mathcal{U}(0, 100)$$

The HBM observer only has access to noisy von Mises observations of $\theta_{i=1}^{N}$:

$$x_i|\theta_i \sim VM(\theta_i, \kappa_{obs}) \quad i = 1, ..., N$$

where κ_{obs} is the precision of noisy observations $x_{i=1}^{N}$ centered on $\theta_{i=1}^{N}$.

κ_{obs} is therefore the only free parameter in this model. We estimated plausible values of κ_{obs} by fitting a von Mises function—via maximum likelihood estimation—to error distributions from set size 1 of the continuous whole-report task (see Results; Fig. 3B). To generate a range of model predictions, all simulations were conducted in triplicate using $\kappa_{obs} = 5, 10, 20$.

Bayesian Finite Mixture Model (BFMM)

We also implemented a more complex generalization of the HBM known as a Bayesian finite mixture model (Orhan and Jacobs, 2013). In addition to individual- and ensemble-level representations, the BFMM also includes an intermediate level where stimuli are generated from a set of K components, such that similar stimulus values are assumed to be generated from the same component. Performing inference with the BFMM can be thought of as finding a posterior distribution over all possible assignments of N stimuli to K clusters.

Formally, each stimulus θ_i is assumed to be generated by one of K weighted von Mises components. The components are assumed to be generated by the following process:

$$G \sim DP(G_0, \alpha)$$

$$G_0(\mu_k, \kappa_k) = \mathcal{U}(-\pi, \pi)\mathcal{G}(1, \beta_\sigma)$$

where G is a discrete distribution over the parameters of the K components. G_0 is the prior distribution over the two-dimensional parameter space for μ_k and κ_k, the mean and precision of component k. The β_σ variable for the gamma prior is given a $\mathcal{G}(1, 1)$ hyperprior.

G can be expressed as a weighted sum of K "atoms" or points p in the two-dimensional parameter space G_0:

$$G(p) = \sum_{k=1}^{K} \pi_k \delta(p = p_k)$$

where π_k is the weighting of component k and δ is the Dirac delta function. The component weights π are drawn from a symmetric Dirichlet prior with concentration parameters α/K:

$$\pi \sim \text{Dirichlet}(\alpha/K, ..., \alpha/K)$$

The assumed stimulus-generating and observation processes are then:

$$\theta_i | \mu_k, \kappa_k \sim VM(\mu_k, \kappa_k) \quad i = 1, ..., N$$

$$x_i | \theta_i \sim VM(\theta_i, \kappa_{obs}) \quad i = 1, ..., N$$

where $x_{i=1}^N$ are noisy von Mises observations centered on true stimulus values $\theta_{i=1}^N$ with precision κ_{obs}.

The BFMM has two free parameters: κ_{obs} indicating the observation precision, and the Dirichlet clustering parameter α. As above, simulations were performed in triplicate with $\sigma_{obs} = 5, 10, 20$. The clustering parameter α was fixed at 1 for all simulations, roughly corresponding to a uniform distribution over component weights (Orhan and Jacobs, 2013).

Model simulations

Model simulations were performed in the same way for both the HBM and the BFMM. To simulate a whole-report trial at set size N, a noisy estimate was first generated for each stimulus by sampling from a von Mises distribution with precision κ_{obs}. The models do not have access to the true values of θ, but instead perform inference based on these noisy observations.

Inference and model specification were implemented using the probabilistic Python package PyMC3 (Salvatier, Wiecki, and Fonnesbeck, 2016). Model posteriors were sampled using a Hamiltonian Monte Carlo algorithm known as the No-U-Turn Sampler (Hoffman and Gelman, 2014). For each trial, 5000 samples were drawn (after 2000 tuning steps).

To simulate reporting, the mode of each stimulus posterior was used as a point estimate (taking the mean did not change our results), and point estimates were rounded to one of either 360 or 8 possible values for the continuous and discrete task, respectively.

Data and code availability

All Python code required to reproduce data processing, analysis, modeling, and visualization is available at this project's GitHub repository. The repository also contains both datasets in their original raw `.mat` format, as well as pickled binary versions (stored in pandas Dataframes) suitable for analysis with Python. Installation and usage instructions are provided, and all code is thoroughly documented.

Results

We analyzed behavioural data from two independent whole-report experiments to characterize within-trial report behavior and evaluate the applicability of encoding models developed in the context of single-report tasks.

These whole-report experiments used two subtly different tasks (Fig. 1A-B; Methods), so we began by replicating the continuous whole-report results of Adam, Vogel, and Awh, 2017 with discrete whole-report data. We found that cumulative error distributions were qualitatively similar despite differences in experimental setup, stimuli, timing, and report method.

We then analyzed within-trial behavior, and found that participants tended to group consecutive reports by color similarity—reports made later in a trial were often biased away from early-trial reports. This pattern of dependence was consistent across datasets, participants, and set sizes when reporting color, but not when reporting orientation.

Finally, we implemented two hierarchical Bayesian ensemble encoding models to determine whether color report dependecies could be explained by the joint encoding of within-trial stimuli. We simulated both whole-report tasks using both models, and found no empirical evidence for the biases that either model predicted.

Continuous and discrete whole-report error distributions are qualitatively similar

In order to verify that differences in task timing, stimuli, and report method between the continuous and discrete tasks did not impact participant behavior, we replicated the aggregate results of Adam, Vogel, and Awh, 2017 with the discrete whole-report dataset. When comparing data between tasks, we considered both report order conditions (participant- and randomly-ordered), and restricted continuous task data to conditions using color stimuli (the discrete experiment did not include orientation stimuli).

Error distributions from both tasks—collapsed across trials and participants—are shown in Fig. 1. A classic delayed estimation result is that recall performance decreases as a function of set size, and Adam, Vogel, and Awh, 2017 found that continuous whole-report error distributions follow the same trend. Error distributions are less precise at higher set sizes (Fig. 1C; upper) in both the participant- and randomly-ordered report conditions.

Discrete whole-report error distributions are qualitatively similar (Fig. 1D; upper). Following Adam, Vogel, and Awh, 2017, we quantified this effect using mean resultant vector length (MRVL), a measure of circular dispersion. MRVL decreases as a function of set size in both report conditions, and repeated-measures analyses of variance (ANOVAs) show a significant effect of set size on MRVL in both the participant-ordered condition ($F(4,52) = 339.66$, $p = 1.40 \times 10^{-36}$) and randomly-ordered condition ($F(4,52) = 268.22$, $p = 5.00 \times 10^{-34}$).

Whole-report error distributions can be further separated by the order in which stimuli were reported within each trial. This is illustrated for set size 6 in the bottom panels of Fig. 1C and D (see Supplementary Fig. 8 for all set sizes). Adam, Vogel, and Awh, 2017 found that when report order is freely selected by participants, MRVL decreases as a function of report number (reproduced in Supplementary Fig. 9A).

Discrete whole-report error distributions remain similar to continuous distributions when separated by report order. In the participant-ordered condition, MRVL decreases as a function of

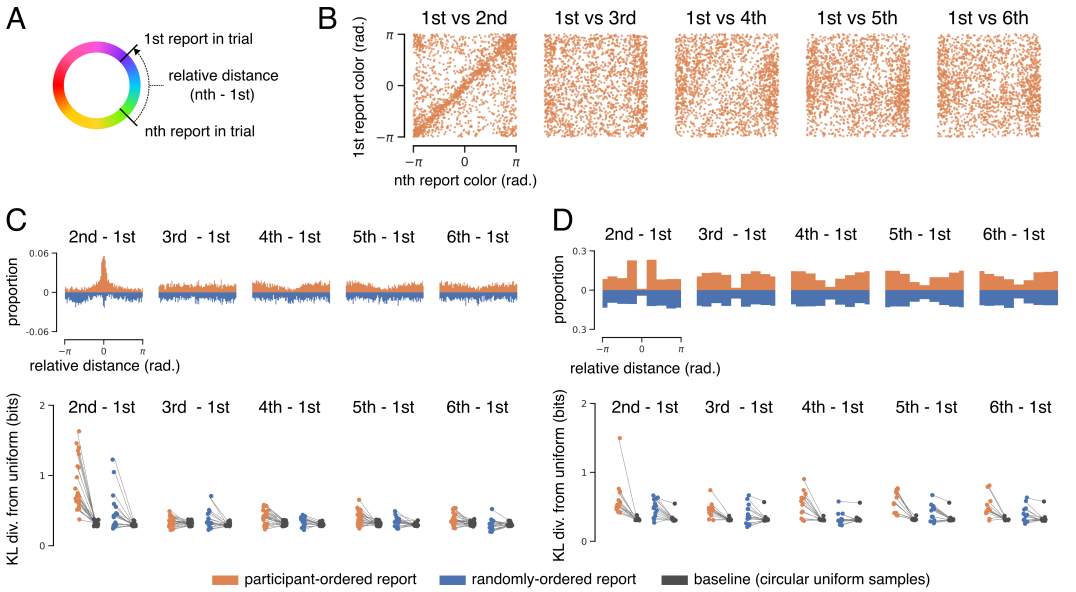

Figure 2. Dependence of within-trial color reports. **A.** Relative distance calculation, which was identical for both tasks. **B.** Example joint distribution of reported stimulus values. Data from the participant-ordered continuous whole-report task (set size 6; aggregated across all participants). Each panel plots the 1st color reported on each trial against a later report in the same trial. **C.** *Upper:* Distribution of relative distances for the continuous task (set size 6). *Lower:* Median bootstrapped estimates of the KL divergence between each participant's relative distance distribution and size-matched samples from a uniform distribution. **D.** Same as C, but for the discrete whole-report task (set size 6).

report number (Supplementary Fig. 9B), and repeated-measures ANOVA showed a significant effect of report order for all set sizes. Notably, in this condition discrete whole-report participants were instructed to report stimuli "in order of confidence," and participants in the continuous task appear to have done the same despite receiving no instruction.

Within-trial color reports are not independent

As discussed above, the whole-report task allows us to analyze the joint distribution of stimulus reports within a trial. Several models of delayed estimation make the assumption that stimuli are encoded independently—if this is the case, the joint distribution of within-trial reports should be uniform.

Fig. 2B shows example within-trial joint distributions for set size 6 of the participant-ordered continuous whole-report task. Each panel corresponds to the joint distribution between the 1st and the nth color reported within the same trial. [1] Two dependencies are apparent: participants tended to report similar colors for the 1st and 2nd reports in a trial, and they tended to avoid reporting colors similar to the 1st with later reports. This is not unique to the 1st report—consecutive reports are similar regardless of when in the trial they occurred.

[1] Note that the joint distribution of two circular variables lies on the surface of a torus. This means that a point in the upper-left or bottom-right of panels in Fig. 2B represents two values very near each other in circular space.

Another way to visualize this is to compute the relative angular distance between two colors reported in the same trial (Fig. 2A). The upper panel of Figure 2C shows relative distance distributions computed for the continuous whole-report task at set size 6. The pattern of dependencies described above is evident for the participant-ordered condition. Deviations from uniformity are also evident for the randomly-ordered report condition, despite the fact that the corresponding distributions of *presented* values were uniform. Critically, this effect was not driven by a few participants making idiosyncratic reports, but was consistent across all participants (Supplementary Fig. 10) and set sizes (Supplementary Fig. 11).

Deviations from uniformity are evident for the discrete whole-report task, and patterns of dependence were qualitatively similar to those evident in the continuous task data, with one key dif-

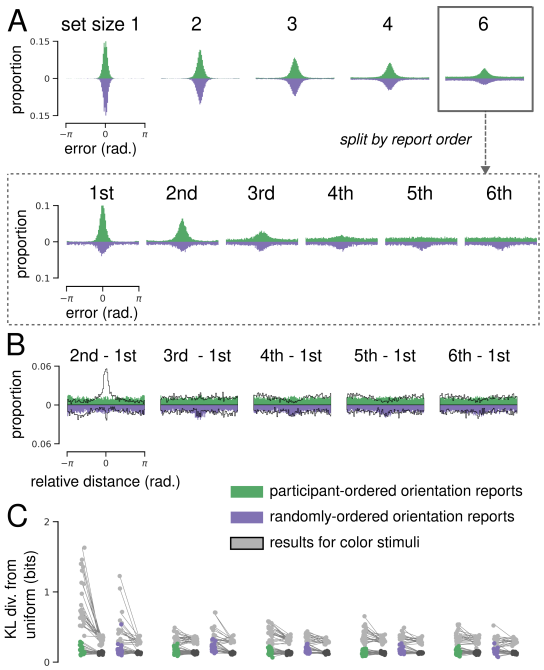

Figure 3. Continuous whole-report results for orientation stimuli. **A.** Orientation error distributions. 360 possible error values divided into 90 bins for visualization. Green and purple histograms show errors from participant-selected and randomly-generated report conditions, respectively. *Upper*: All report errors, separated by set size. *Lower*: Set size 6 report errors, separated by report order. **B.** Distribution of relative distances for orientation reports (set size 6). Color results overlayed in black for comparison. **C.** Median bootstrapped estimates of the KL divergence between each participant's relative distance distribution and size-matched samples from a uniform distribution. Color results overlayed in grey for comparison.

ference—in both conditions of the discrete task, participants rarely repeated their 1st report later in the trial. This suggests that participants were not simply reporting the same color repeatedly, but explicitly avoiding previous colors while consecutively reporting similar colors.

We quantified dependence between within-trial reports using Kullback-Leibler (KL) divergence, a measure of the statistical distance between two distributions. For each participant, report condition, and pair of reports, we estimated the KL divergence between the relative distance distribution and a circular uniform distribution. We compared these KL divergence estimates to a baseline divergence obtained by repeating this process using a size-matched circular uniform sample. This process was bootstrapped to obtain an estimate of uncertainty (see Methods for details), and median KL divergence estimates for each participant are shown alongside corresponding baseline estimates in the lower panels of Fig. 2C and D. This analysis confirmed that distributions of within-trial relative distance often deviated from uniformity for all participants of both tasks. Divergences tended to be higher in the participant-ordered condition than in the randomly-ordered condition, and were highest for consecutive reports.

Repeated-measures ANOVAs using estimate type (ie. data vs baseline) and report pair as within-subject factors confirmed a significant effect of estimate type on KL divergence in both the particpant-ordered (F(1,21) = 70.49, $p = 3.76 \times 10^{-8}$) and randomly-ordered (F(1,16) = 10.55, $p = 0.005$) conditions of the continuous task. The same analysis confirmed a significant effect in both the particpant-ordered (F(1,13) = 55.13, $p = 5.0 \times 10^{-6}$) and randomly-ordered (F(1,13) = 32.74, $p = 7.0 \times 10^{-5}$) conditions of the discrete task. These analyses also showed significant effects of report order, as well as an interaction between report order and estimate type.

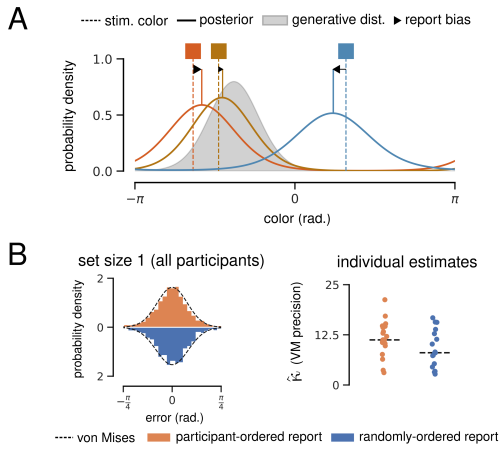

Figure 4. Hierarchical Bayesian Model (HBM). **A.** Illustration of the HBM encoding 3 color stimuli (vertical dashed lines). The model infers that all colors are drawn from the same ensemble distribution (grey). Solid lines show posterior estimates for each stimulus value, and black arrows illustrate the resulting bias toward the ensemble mean. **B.** Justification for free parameter κ_{obs}, the precision of noisy observations. *Left:* Error distributions for color reports at set size 1. 360 error values divided into 90 bins for visualization. Maximum-likelihood von Mises distribution fit overlayed. *Right:* Median bootstrapped estimate of the precision of a von Mises fit to set size 1 report errors for each participant.

Within-trial orientation reports do not show clear dependence

Many models of delayed-estimation tacitly assume that different types of "circular" stimuli—such as colored squares and orientated barscare equivalently encoded into working memory. This is supported by qualitative similarities in error distributions in single-report delayed estimation tasks (Zhang and Luck, 2008; van den Berg et al., 2012; Bays, Catalao, and Husain, 2009). Adam, Vogel, and Awh, 2017 found that aggregate error distributions for orientation and color stimuli were qualitatively similar in the whole-report task. Error distributions for reported orientations—collapsed across reports and participants—are less precise at higher set sizes (Fig. 3A; upper) in both the participant- and randomly-ordered report conditions. They also found that when report order was freely selected by participants, MRVL decreased as a function of report number—again mirroring results for color stimuli.

Despite this similarity at the level of aggregate errors, we found marked differences between color and orientation when analyzing within-trial reports (Fig. 3B and C). Relative distance distributions for orientation do not exhibit clear dependence between consecutive reports in the participant-ordered condition, nor did participants avoid repeating early reports later in the trial. As above, we quantified dependence between within-trial orientation reports using Kullback-Leibler (KL) divergence, a measure of the difference between two distributions. Median KL divergence estimates for each participant are shown alongside corresponding baseline estimates in Fig. 3C (results for color reports are overlayed in gray for comparison). [2] In contrast to results for color (Fig. 2), this analysis revealed that for most participants, deviations from uniform were either greatly reduced or absent.

We performed repeated-measures ANOVAs using estimate type (ie. data vs baseline) and report pair as within-subject factors. This analysis found a significant effect of estimate type on KL divergence in both the particpant-ordered (F(1,29) $= 39.10$, $p = 0.01 \times 10^{-4}$) and randomly-ordered (F(1,18) $= 34.12$, $p = 1.6 \times 10^{-5}$) conditions, but no significant effects of report order or interaction between report order and estimate type. Although this suggests some dependence between orientation reports, this effect was greatly reduced in comparison to results for color.

[2] Note that baseline KL divergence estimates are higher for the color distributions—this is because participants in the orientation task completed 200 trials per set size (vs 99 for the color task).

Hierarchical ensemble encoding predicts biases not present in whole-report data

As described above, we found clear evidence that within-trial color reports are not independent in the whole-report task. Participants appear to group stimuli by color when reporting, raising the possibility that they are encoding higher-order statistics of color displays.

This is a proposal made by "ensemble" encoding models of VWM, where visual scenes are hierarchically encoded at multiple levels of abstraction and individual stimulus representations are marginally dependent due to shared ensemble statistics (T. F. Brady, Konkle, and Alvarez, 2009; T. F. Brady and Alvarez, 2011; Orhan and Jacobs, 2013). This introduces potentially unwanted biases, but could reduce variance in stimulus estimates (Orhan and Jacobs, 2013) and improve encoding efficiency (Nassar, 2018).

We implemented two hierarchical encoding models to simulate the whole-report task and test this possibility. The first is a simple adaptation of T. F. Brady and Alvarez, 2011: a two-level hierarchical Bayesian model (HBM) that encodes both individual stimuli and the ensemble statistics—mean and precision—of all stimuli in the display. This model makes the simple, testable prediction that color reports will be biased toward the mean color of the display.

To address the possibility that participants may be encoding multiple groups of similar colors, we also adapted a hierarchical encoding model known as a Bayesian finite mixture model (BFMM) (Orhan and Jacobs, 2013). The BFMM has another level of abstraction where stimuli are grouped by similarity, and the statistics of each group are also encoded. This model predicts that color reports will be biased toward the mean of the group to which they belong.

Hierarchical Bayesian model (HBM)

The biases produced by a HBM with two levels of representation are illustrated in Fig. 4A. In the HBM, an observer assumes that all stimuli in a trial are samples from the same underlying ensemble distribution, and this internal model effectively forms a prior over the values of all stimuli. When integrating over the uncertainty in individual stimulus encodings and this prior, the observer produces posterior estimates that are biased toward the ensemble mean.

The HBM implemented here assumes that individual stimulus and ensemble encodings take the form of a von Mises—or "circular normal"—distribution, each parameterized by a mean (μ) and precision (κ). These parameters are all inferred from presented stimuli via Bayesian inference (see Methods for full equations and priors). The model also assumes that stimulus values are only available as a noisy observation sampled from a von Mises distribution with precision κ_{obs} centered on the true value. κ_{obs} is the HBM's only free parameter, and is assumed to represent the combined effect of sensory and memory noise (Orhan and Jacobs, 2013). Intuitively, lower values of κ_{obs} result in noisier individual representations and a greater bias toward the ensemble mean.

To estimate reasonable values for κ_{obs}, we took advantage of continuous whole-report error distributions at set size 1, where there are theoretically no higher-order statistics to encode. We fit a von Mises distribution to each participant's report errors at set size 1, giving us a range of plausible κ_{obs} values (Fig. 4B). For all HBM and BFMM results presented here, simulations were conducted in triplicate using $\kappa_{obs} = 5, 10, 20$.

Here, we focus on simulation results from set size 3, reasoning that the HBM would tend to infer a more precise generative distribution—and therefore predict stronger biases—with fewer stimuli. HBM simulations of the continous and discrete whole-report tasks are presented in Fig. 5A and C, respectively. Report biases can be visualized by comparing two angles: (1) the angular distance between a presented color and the trial ensemble mean and (2) the error between presented and reported colors. A correlation between the signs of these two angles provides evidence that reported colors are biased toward the ensemble mean.

As expected, HBM simulations of both tasks demonstrate a clear bias toward the ensemble mean. Bias magnitude was a function of the distance to the mean color—the farther a stimulus was from the ensemble mean, the greater the reported bias toward the mean. Bias magnitude was also dependent on the choice of κ_{obs}—the lowest κ_{obs} value resulted in the greatest bias and

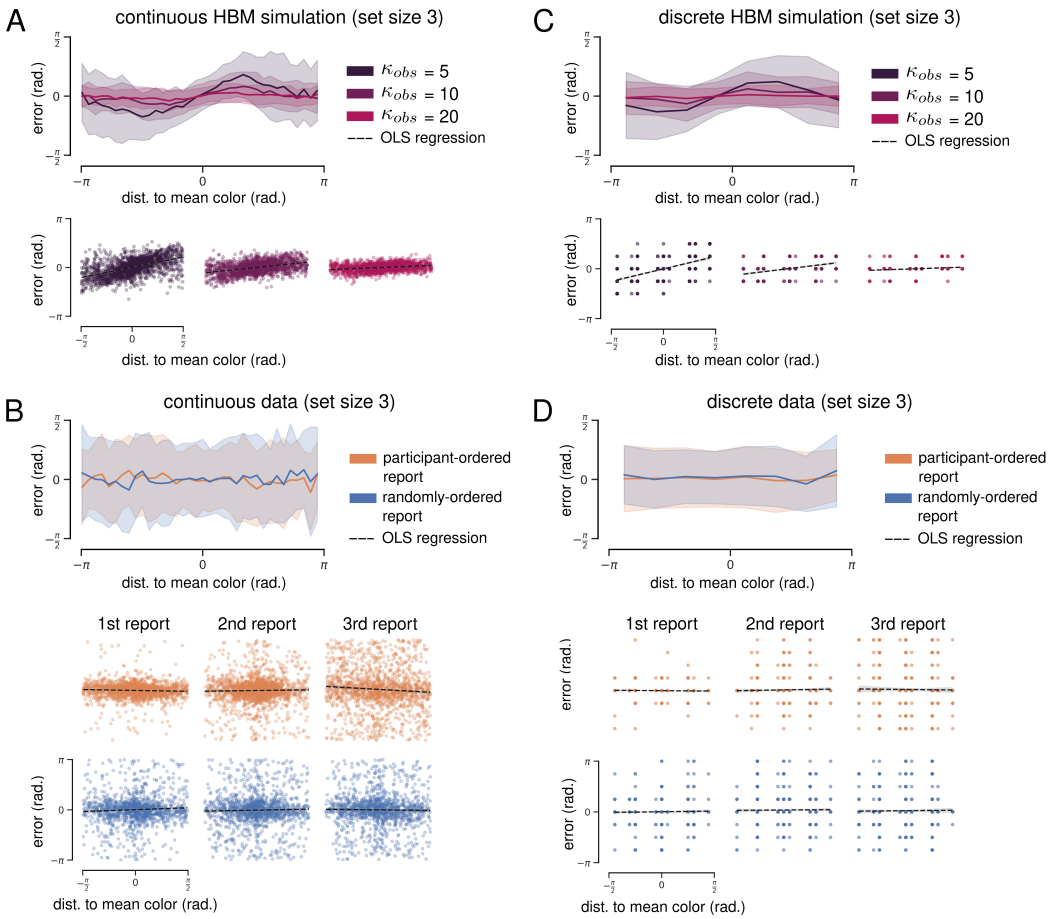

Figure 5. Hierarchical Bayesian model (HBM) simulation results for set size 3. **A.** *Upper:* Biases predicted by HBM simulations of the continuous task. Mean report error plotted as a function of the reported color's distance to the ensemble mean. 360 error values are divided into 90 bins for visualization, and shaded area shows standard deviation. *Lower:* OLS regression of simulated report error on distance to ensemble mean (dashed black lines). **B.** *Upper:* Empirical biases for the continuous task at set size 3 (collapsed across all reports). *Lower:* OLS regression of empirical report error on distance to ensemble mean. Each column shows results for a different report number. **C.** Same as panel A, but for simulations of the discrete task. **D.** Same as panel B, but for empirical results of the discrete task.

vice versa. We quantified bias using ordinary least squares (OLS) regression of report error on the distance to ensemble mean. This analysis was restricted to include only mean distances between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ where the bias function was approximately linear (see Methods for details). Regression was done at two levels: (1) all simulated errors for a given set size, collapsed across reports; and (2) simulated errors separated by report number. [3]

OLS regression was significant for all κ_{obs} in continuous whole-report HBM simulations at set size 3 (collapsed across reports). The largest estimated effect was for $\kappa_{obs} = 5$ ($R^2 = 0.278$, mean dist. coefficient $= 0.428$, $p = 2.97 \times 10^{-124}$), and the smallest estimated effect for $\kappa_{obs} = 20$ ($R^2 = 0.053$, mean dist. coefficient $= 0.078$, $p = 3.63 \times 10^{-63}$. OLS regressions were significant when simulated reports were separated by report order.

OLS regression was also significant for all κ_{obs} in discrete whole-report HBM simulations at set size 3 (collapsed across reports). The largest estimated effect was for $\kappa_{obs} = 5$ ($R^2 = 0.252$, mean dist. coefficient $= 0.454$, $p = 4.03 \times 10^{-104}$), and the smallest estimated effect for $\kappa_{obs} = 20$ ($R^2 = 0.024$, mean dist. coefficient $= 0.053$, $p = 2.73 \times 10^{-10}$). OLS regressions were significant when simulated reports were separated by report order.

Empirical bias functions for the continuous and discrete whole-report tasks are shown in Fig.

[3] Note that reports are not simulated sequentially—neither the HBM or the BFMM include a temporal or sequential component. For visualization of both simulations and data, the bias functions in Fig. 5 include all reports, but the simulation regression plots in Fig. 5A and C include only the "first" simulated report to facilitate comparison with data regressions, which are separated by report order.

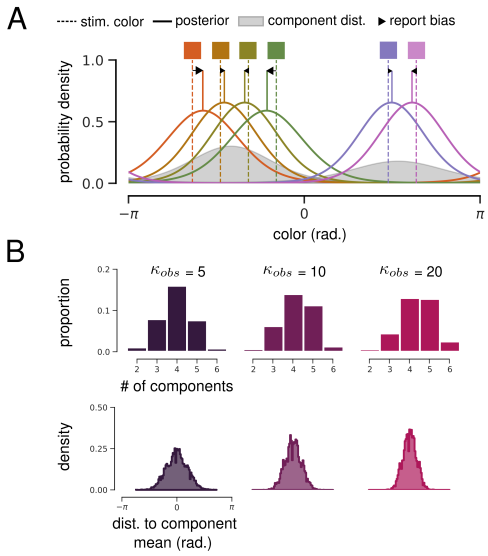

Figure 6. Bayesian finite mixture model (BFMM). **A.** Illustration of a BFMM with $K = 2$ components encoding 6 color stimuli (vertical dashed lines). The model infers that colors were generated from a mixture of 2 von Mises components (grey). Solid lines show posterior estimates for each stimulus value, and black arrows illustrate resulting biases toward component means. **B.** Impact of κ_{obs} on clustering in a BFMM with $K = 6$ components. *Upper:* Distribution of the number of components that the BFMM infers generated colors on each trial of the simulated continuous whole-report task at set size 6. Higher values of κ_{obs} cause the BFMM to assume a higher number of more precise components. *Lower:* Distribution of distance to inferred component mean for all simulated reports.

5B and D, respectively. Empirical report errors were more broadly distributed (less accurate) than simulated results, but report error was comparable when only the first report in each trial was considered (Fig. 5B and D; first column).

We repeated the regression analysis using experimental data from both report conditions of both tasks at set size 3. Regression was significant for the participant-ordered continuous data (collapsed across reports), but the estimated mean distance coefficient was negative, indicating a bias *away* from the ensemble mean ($R^2 = 0.001$, mean dist. coefficient $= -0.050$, $p = 0.004$). We also found significant negative biases for the 1st and 3rd reports when the data were separated by report order. There were no significant regression results for the randomly-ordered condition of the continuous whole-report task at set size 3, or for either condition of the discrete whole-report task.

We also simulated both tasks and report conditions at set size 6, and repeated the OLS regression analysis for both simulated and experimental reports. In comparison to set size 3 simulations, the HBM tended to infer less precise generative distributions, resulting in smaller predicted biases. These biases were still significant for all continuous task simulations, but were not significant for the majority of discrete task simulations—likely because the predicted biases were smaller than the discrete stimulus increment ($\frac{\pi}{4}$) and were lost when estimates were rounded for report.

There were several significant regression results for experimental data from both tasks at set size 6. Similar to set size 3, these regressions all estimated small negative mean distance coefficients. Finally, we repeated all analyses above with experimental data from the continuous task with orientation stimuli (Supplementary Fig. 12). We found no significant regressions with positive coefficients in either report condition at set size 3 or set size 6. In summary, we found no empirical evidence for a report bias toward within-trial mean stimulus values.

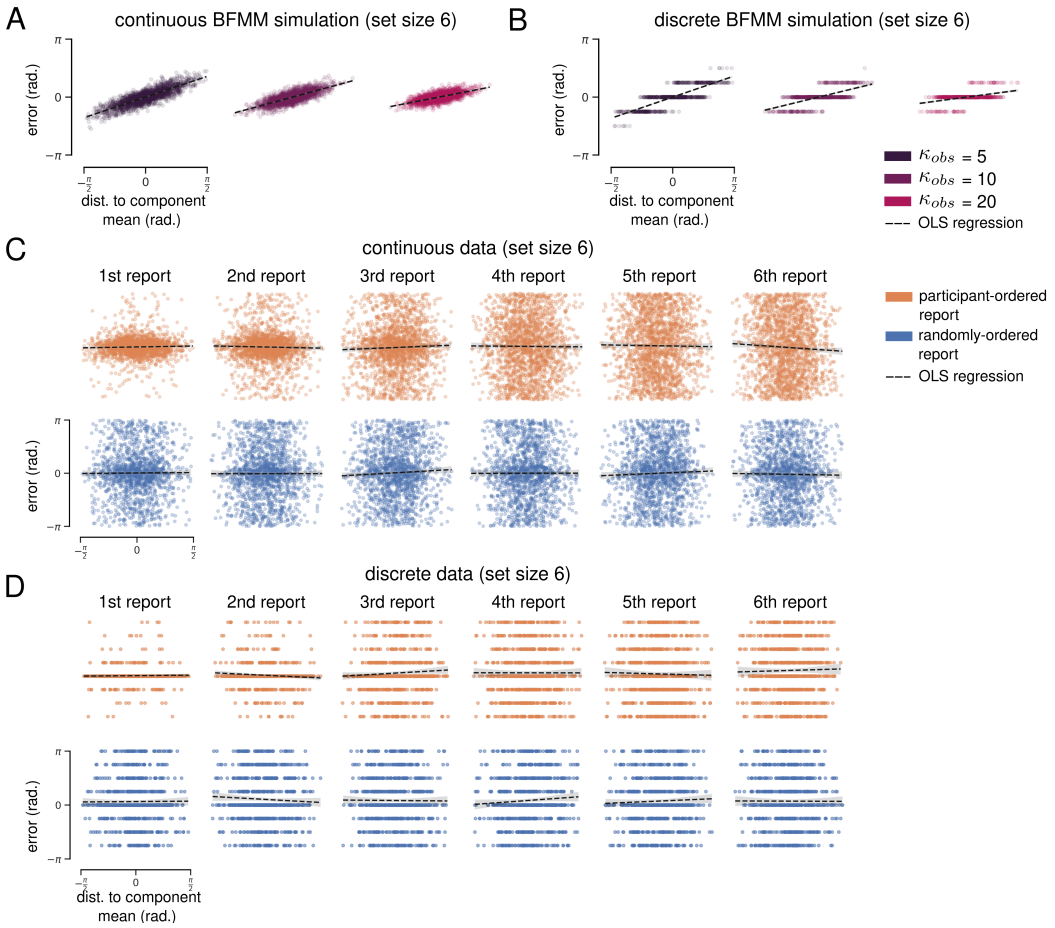

Figure 7. Bayesian finite mixture model (BFMM) simulation results for set size 6. **A.** OLS regression of simulated continuous report error on distance to component mean (dashed black lines). **B.** Same as panel A, but for discrete task simulations. **C.** OLS regression of empirical report error on distance to component mean. Each column shows results for a different report number. **D.** Same as C, but for empricial results of the discrete task.

Bayesian finite mixture model (BFMM)

The HBM presented above assumes that all stimuli are generated from a single distribution. The limitations of this approach are clear when considering the HBM illustration in Fig. 4: although encoding orange and red together is intuitive, it is not obvious that the blue stimulus should be included. It is possible that participants cluster similar stimuli together, and encode ensemble statistics of these clusters in addition to statistics of the entire display (T. F. Brady and Alvarez, 2011; Orhan and Jacobs, 2013).

To address this, we adapted a flexible generalization of Brady and Alvarez's HBM called a Bayesian finite mixture model (BFMM) (Orhan and Jacobs, 2013). The BFMM assumes that each stimulus value is generated from one of K weighted von Mises components (Fig. 6A shows an example where $K = 2$). [4] Component weights and parameters are determined via Bayesian inference, so sampling from the posterior of the BFMM can be thought of as performing data-driven clustering (see Methods for full model and priors).

K (the maximum number of components) was fixed to equal the simulated set size, and the same values of κ_{obs} (the precision of noisy observations) were used. Here, κ_{obs} has a subtly different impact on simulation results than in the HBM (Fig. 6B). At lower values of κ_{obs}, the model tends to infer that stimuli were generated from a small number of low-precision components, whereas at higher values of κ_{obs}, the model tends to infer a greater number of higher-precision components

[4] Orhan and Jacobs, 2013 also implemented a non-parametric model that allows for infinite groups. The BFMM is a constrained version, but is equivalent when K is equal to the set size.

(up to a maximum of K).

BFMM simulations were very similar to HBM simulations described above (Fig. 7). To visualize bias, we plot report error as a function of the distance to the mean of the component that the model assumes is responsible for generating a given stimulus. As expected, simulated reports for both the continuous and discrete tasks demonstrate a clear bias toward their respective component means. Similar to the HBM simulations above, bias magnitude was greater for larger distances to the component mean, and was greater for smaller values of κ_{obs}. Overall, BFMM report biases were greater than HBM biases.

To quantify bias, we used ordinary least squares (OLS) regression of report error on the distance to component mean. As above, we restricted this analysis to include only mean distances between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ (in practice, this included nearly all reports; see Fig. 6B).

OLS regression was significant for all κ_{obs} in continuous whole-report BFMM simulations at set size 6 (collapsed across reports). The largest estimated effect was for $\kappa_{obs} = 5$ ($R^2 = 0.746$, mean dist. coefficient $= 0.725$, $p < 1.0 \times 10^{-300}$), and the smallest estimated effect for $\kappa_{obs} = 20$ ($R^2 = 0.465$, mean dist. coefficient $= 0.429$, $p < 1.0 \times 10^{-300}$). OLS regressions were also significant when simulated reports were separated by report order.

OLS regression was also significant for all κ_{obs} in discrete whole-report BFMM simulations at set size 6 (collapsed across reports). The largest estimated effect was for $\kappa_{obs} = 5$ ($R^2 = 0.644$, mean dist. coefficient $= 0.722$, $p < 1.0 \times 10^{-300}$), and the smallest estimated effect for $\kappa_{obs} = 20$ ($R^2 = 0.207$, mean dist. coefficient $= 0.282$, $p = 1.41 \times 10^{-247}$). OLS regressions were also significant when simulated reports were separated by report order.

We repeated the OLS regressions using empirical data from both report conditions of both tasks at set size 6. Regressions were not significant for either report condition of either task when data were collapsed across reports.

When separated by report order, three regressions reached significance for the continuous whole-report task: the 3rd and 6th report in the participant-ordered condition and the 3rd report of the randomly-ordered condition. These regressions all estimated small *negative* coefficients for the distance to component mean. There were no significant regression results for either condition of the discrete whole-report task when data were separated by report order.

Finally, we repeated all analyses above with experimental data from the continuous task with orientation stimuli (Supplementary Fig. 13) and found no significant regressions with positive coefficients. Taken together, the many regression analyses we performed suggest that color reports were not consistently biased toward inferred ensemble mean in any condition of the whole-report task.

Discussion

Here, we analyzed data from two whole-report visual working memory (VWM) experiments to characterize within-trial behavior and test the applicability of delayed estimation models. In contrast to models that assume independent encoding, we found that within-trial color reports were not independent. We also showed that this effect is either weaker or absent in task conditions using orientation stimuli, calling into question the assumption that color and orientation are equivalently encoded variables. Finally, we implemented two hierarchical Bayesian ensemble encoding models that explicitly include dependence between encoded stimuli and found no empirical evidence for the biases that they predict.

We began by replicating the continuous whole-report results reported by Adam, Vogel, and Awh, 2017 with data from our own discrete whole-report dataset. Although we successfully replicated several key results, our analyses of within-trial behavior in both datasets raise a potential issue with their conclusions. Specifically, Adam et al. found that late-trial report distributions were best described as uniform, and that participants typically self-reported that late-trial color reports were guesses. Reasoning that participants had no information about several items on each trial,

they concluded that this was clear evidence for a fixed item limit in VWM. This has proven controversial, as some have argued that apparently uniform distributions could reflect very low-precision retrieval (Schneegans, Taylor, and Bays, 2020), and others have failed to reproduce late-trial uniform distributions altogether (Oberauer, 2022). Our results raise a different issue: we found that late-trial color reports were consistently biased away from early reports, which suggests that participants had *some* information about the colors they were reporting—even when self-reporting guesses.

Our analyses addressed the possibility that within-trial dependencies were indicative of hierarchical Bayesian ensemble encoding, but it is also possible that participants were using explicit or implicit task strategies that introduced dependency. For example, discrete task participants tended to consecutively report similar but not identical colors, suggesting that they were exploiting the absence of stimulus repeats within each trial. It has also been shown that certain whole-report stimuli can induce structured non-uniform guessing in late-trial reports (Ngiam et al., 2023), and we consider it probable that whole-report error distributions reflect both "memory" and "non-memory" processes. Unfortunately, post-hoc attribution of behavior to one process or another is not always straightforward—what if a participant intentionally guesses a color that they *do not* remember?—and distinguishing between behavior resulting from encoding structure versus task strategy raises challenges for the interpretation of VWM tasks that we will revisit below.

Adam, Vogel, and Awh, 2017 also reported very similar results for whole-report task conditions using color stimuli and those using orientation stimuli, and fit a set of models that assume a generic circular variable is encoded. Modeling color and orientation as equivalent is common (van den Berg et al., 2012; Schneegans, Taylor, and Bays, 2020), but recent work on has challenged the treatment of color as a circular variable (Schurgin, Wixted, and T. F. Brady, 2020). Although our results do not directly address the nature of color representations, the behavioral differences we identified between color and orientation conditions highlight the need to consider stimulus-dependent effects. It is also worth noting that there is considerable variation in visual working memory stimuli even when using the "same" stimulus type—few studies use standardized screen calibration methods for color (Bae, Olkkonen, et al., 2014), and orientation stimuli include Gabor patches ranging from $-\pi/2$ to $\pi/2$ (van den Berg et al., 2012) as well as various oriented "clock hands" and triangles ranging from $-\pi$ to π (Adam, Vogel, and Awh, 2017; Bae and Luck, 2017; Utochkin and T. F. Brady, 2020).

Our focus on ensemble encoding as a potential cause of within-trial report dependencies was motivated by a rich body of VWM literature. The assumption of independent stimulus encoding has been challenged before using a variety of single-report working memory tasks (Jiang, Olson, and Chun, 2000; Kahana and Sekuler, 2002; Alvarez and Oliva, 2009; T. F. Brady and Alvarez, 2011; Orhan and Jacobs, 2013; Lew and Vul, 2015; Bae and Luck, 2017; Nassar, 2018) and at least one whole-report task (Utochkin and T. F. Brady, 2020), so it was unsurprising that we found evidence for dependence in the whole-report task. More surprising was that the hierarchical encoding models we implemented predicted report biases that were not present in either whole-report dataset. Our regression analyses of empirical reports failed to reproduce the consistent positive regression coefficents (report bias toward inferred ensemble or component mean values) that were present in HBM and BFMM simulations. In fact, a small number of regressions resulted in significant *negative* coefficents—this could be interpreted as bias away from ensemble or component means, but the effect was not consistent enough for us to draw such a conclusion. These results contrast with previous studies showing report biases toward mean stimulus size (T. F. Brady and Alvarez, 2011), mean horizontal position (Orhan and Jacobs, 2013), and mean 2-dimensional spatial position (Lew and Vul, 2015) that were accounted for by very similar hierarchical Bayesian encoding models. This discrepancy could possibly be explained by differences in stimulus type and task structure, but this interpretation is complicated by the results of Utochkin and T. F. Brady, 2020, where reports were biased toward mean orientation in a whole-report task.

It is important to note that ensemble encoding can take many forms, and that our results only

contradict ensemble encoding models that assume a specific hierarchical Bayesian generative process. There are types of report bias—such as repulsion from dissimilar colors (Nassar, 2018)—that are not accounted for by hierarchical encoding but can still be considered evidence for ensemble encoding. Furthermore, a recent analysis of the same continuous whole-report dataset from Adam, Vogel, and Awh, 2017 found evidence for systematic biases that were explained by a non-hierarchical chunking model (Chunharas and T. Brady, 2023). Specifically, Chunharas and T. Brady, 2023 show that in trials identified by the model as chunkable, early color reports were biased toward the "gist" color and later color reports were biased away from the "gist" color. The "gist" in their analysis differs subtly from the ensemble and component means inferred by the hierarchical Bayesian models we implemented in that it involves the convolution of an empirically-derived psychophysical similarity function (Schurgin, Wixted, and T. F. Brady, 2020) with presented stimulus values. This method could be considered an empirical or phenomenological approach to identifying ensemble effects, in contrast to hierarchical Bayesian encoding models (T. F. Brady and Alvarez, 2011; Orhan and Jacobs, 2013; Lew and Vul, 2015) that make specific assumptions about the encoding structure used by participants. In our view, our results are compatible with those reported by Chunharas and T. Brady, 2023—although we did not find evidence for hierarchical encoding, the dependencies we identified are consistent with ensemble encoding more broadly, and we agree with their rejection of independent-item-based accounts of visual working memory.

Conclusions

We analyzed data from two whole-report visual working memory tasks, and identified behavior that cannot be accounted for by models that assume independent encoding, equivalence between color and orientation stimuli, or hierarchical ensemble encoding. Given that these models were developed to fit single-report delayed estimation data and did not include sequential report mechanisms, it might be tempting to attribute such discrepancies to the idiosyncratic nature of the whole-report task. After all, the whole-report task plausibly suffers from the "task impurity" problem (Burgess, 1997) because it involves too many cognitive processes—attention, decision-making, strategies, etc—that are incidental to the structure of VWM encodings.

This problem is not unique to the current study, and raises a difficult dilemma. Should we reject more complex working memory tasks in the pursuit of "pure" working memory encoding models that may not generalize, or should we attempt to model the many processes responsible for idiosyncratic phenomena specific to the whole-report task? There is a growing recognition in the field that the proliferation of diverse VWM tasks, models, and empirical findings is outstripping our ability to develop theory capable of synthesizing them (Oberauer et al., 2018; Popov, 2023; Ngiam et al., 2023). In this context, neither option seems appropriate. Instead, we echo calls for a more integrative approach to theory development (Ngiam et al., 2023) and caution against inferring fundamental principles of visual working memory using data from any individual laboratory task.

Additional acknowledgments

This preprint was created using the LaPreprint template (https://github.com/roaldarbol/lapreprint) by Mikkel Roald-Arbøl.

Author contributions

Conceptualization: B.C., D.S., M.P., G.B.; Methodology: B.C., G.B.; Software and Analysis: B.C.; Writing - original draft: B.C.; Writing - review & editing: B.C., D.S., M.P., G.B.; Supervision: G.B., M.P.; Project administration: G.B.; Funding acquisition: G.B.

References

Adam, Kirsten C. S., Irida Mance, et al. (2015). "The Contribution of Attentional Lapses to Individual Differences in Visual Working Memory Capacity". In: *Journal of Cognitive Neuroscience* 27.8, pp. 1601–1616. ISSN: 0898-929X. DOI: 10.1162/jocn_a_00811. URL: https://doi.org/10.1162/jocn_a_00811 (visited on 2020).

Adam, Kirsten C. S. and Edward K. Vogel (2018). "Improvements to Visual Working Memory Performance with Practice and Feedback". In: *PLOS ONE* 13.8, e0203279. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0203279. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0203279 (visited on 2022).

Adam, Kirsten C. S., Edward K. Vogel, and Edward Awh (2017). "Clear Evidence for Item Limits in Visual Working Memory". In: *Cognitive Psychology* 97, pp. 79–97. ISSN: 00100285. DOI: 10.1016/j.cogpsych.2017.07.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010028517300634 (visited on 2019).

Alvarez, George A. and Aude Oliva (2009). "Spatial Ensemble Statistics Are Efficient Codes That Can Be Represented with Reduced Attention". In: *Proceedings of the National Academy of Sciences* 106.18, pp. 7345–7350. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0808981106. pmid: 19380739. URL: https://www.pnas.org/content/106/18/7345 (visited on 2020).

Bae, Gi-Yeul and Steven J. Luck (2017). "Interactions between Visual Working Memory Representations". In: *Attention, Perception, & Psychophysics* 79.8, pp. 2376–2395. ISSN: 1943-393X. DOI: 10.3758/s13414-017-1404-8. URL: https://doi.org/10.3758/s13414-017-1404-8 (visited on 2020).

Bae, Gi-Yeul, Maria Olkkonen, et al. (2014). "Stimulus-Specific Variability in Color Working Memory with Delayed Estimation". In: *Journal of Vision* 14.4, pp. 7–7. ISSN: 1534-7362. DOI: 10.1167/14.4.7. URL: https://jov.arvojournals.org/article.aspx?articleid=2121591 (visited on 2019).

Bays, Paul M., Raquel F. G. Catalao, and Masud Husain (2009). "The Precision of Visual Working Memory Is Set by Allocation of a Shared Resource". In: *Journal of Vision* 9.10, pp. 7–7. ISSN: 1534-7362. DOI: 10.1167/9.10.7. URL: http://jov.arvojournals.org/article.aspx?articleid=2122354 (visited on 2020).

Bays, Paul M. and Masud Husain (2008). "Dynamic Shifts of Limited Working Memory Resources in Human Vision". In: *Science* 321.5890, pp. 851–854. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1158023. pmid: 18687968. URL: https://science.sciencemag.org/content/321/5890/851 (visited on 2020).

Brady, Timothy F. and George A. Alvarez (2011). "Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items". In: *Psychological Science* 22.3, pp. 384–392. ISSN: 0956-7976. DOI: 10.1177/0956797610397956. URL: https://doi.org/10.1177/0956797610397956 (visited on 2019).

Brady, Timothy F., Talia Konkle, and George A. Alvarez (2009). "Compression in Visual Working Memory: Using Statistical Regularities to Form More Efficient Memory Representations". In: *Journal of Experimental Psychology. General* 138.4, pp. 487–502. ISSN: 1939-2222. DOI: 10.1037/a0016797. pmid: 19883132.

Burgess, Paul (1997). "Theory and Methodology in Executive Function Research". In: *Methodology of Frontal and Executive Function*. Routledge, pp. 87–121.

Chunharas, Chaipat and Timothy Brady (2023). *Chunking, Attraction, Repulsion and Ensemble Effects Are Ubiquitous in Visual Working Memory*. DOI: 10.31234/osf.io/es3b8. URL: https://osf.io/preprints/psyarxiv/es3b8/ (visited on 2023). preprint.

Cuthbert, Benjamin et al. (2018). "Strategic Working Memory Performance May Confound the Interpretation of Cumulative Task Statistics". In: *Journal of Vision* 18.10, pp. 685–685. ISSN: 1534-7362. DOI: 10.1167/18.10.685. URL: http://jov.arvojournals.org/article.aspx?articleid=2699675 (visited on 2020).

deBettencourt, Megan T. et al. (2019). "Real-Time Triggering Reveals Concurrent Lapses of Attention and Working Memory". In: *Nature Human Behaviour* 3.8 (8), pp. 808–816. ISSN: 2397-3374. DOI:

10.1038/s41562-019-0606-6. URL: https://www.nature.com/articles/s41562-019-0606-6 (visited on 2020).

Fougnie, Daryl, Jordan W. Suchow, and George A. Alvarez (2012). "Variability in the Quality of Visual Working Memory". In: *Nature Communications* 3.1 (1), p. 1229. ISSN: 2041-1723. DOI: 10.1038/ncomms2237. URL: https://www.nature.com/articles/ncomms2237 (visited on 2020).

Hao, Yuting et al. (2021). "Free-Recall Benefit, Inhomogeneity and between-Item Interference in Working Memory". In: *Cognition* 214, p. 104739. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2021.104739. URL: https://www.sciencedirect.com/science/article/pii/S001002772100158X (visited on 2022).

Hoffman, Matthew D. and Andrew Gelman (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.47, pp. 1593–1623. ISSN: 1533-7928. URL: http://jmlr.org/papers/v15/hoffman14a.html (visited on 2021).

Jiang, Yuhong, Ingrid R Olson, and Marvin M Chun (2000). "Organization of Visual Short-Term Memory". In: *Journal of Experimental Psychology*, p. 20.

Kahana, Michael J and Robert Sekuler (2002). "Recognizing Spatial Patterns: A Noisy Exemplar Approach". In: *Vision Research* 42.18, pp. 2177–2192. ISSN: 0042-6989. DOI: 10.1016/S0042-6989(02)00118-9. URL: https://www.sciencedirect.com/science/article/pii/S0042698902001189 (visited on 2021).

Killebrew, Kyle W. et al. (2018). "Electrophysiological Correlates of Encoding Processes in a Full-Report Visual Working Memory Paradigm". In: *Cognitive, Affective, & Behavioral Neuroscience* 18.2, pp. 353–365. ISSN: 1531-135X. DOI: 10.3758/s13415-018-0574-8. URL: https://doi.org/10.3758/s13415-018-0574-8 (visited on 2022).

Kleiner, M. et al. (2007). "What's new in psychtoolbox-3". In: *Perception* 36.14, pp. 1–16. ISSN: 0301-0066. URL: https://nyuscholars.nyu.edu/en/publications/whats-new-in-psychtoolbox-3 (visited on 2020).

Lew, Timothy F. and Edward Vul (2015). "Ensemble Clustering in Visual Working Memory Biases Location Memories and Reduces the Weber Noise of Relative Positions". In: *Journal of Vision* 15.4, p. 10. ISSN: 1534-7362. DOI: 10.1167/15.4.10. pmid: 26360154.

MATLAB (2016). *Version 9.0 (R2016a)*. Natick, Massachusetts: The Mathworks Inc.

Nassar, Matthew R. (2018). "Chunking as a Rational Strategy for Lossy Data Compression in Visual Working Memory." In: *Psychological Review* 125.4, p. 486. ISSN: 1939-1471. DOI: 10.1037/rev0000101. URL: https://psycnet-apa-org.proxy.queensu.ca/fulltext/2018-30695-002.html#s15 (visited on 2022).

Ngiam, William X. Q. et al. (2023). "Distinguishing Guesses from Fuzzy Memories: Further Evidence for Item Limits in Visual Working Memory". In: *Attention, Perception, & Psychophysics* 85.5, pp. 1695–1709. ISSN: 1943-393X. DOI: 10.3758/s13414-022-02631-y. URL: https://doi.org/10.3758/s13414-022-02631-y (visited on 2023).

Oberauer, Klaus (2017). "An Interference Model of Visual Working Memory." In: *Psychological Review* 124.1, p. 21. ISSN: 1939-1471. DOI: 10.1037/rev0000044. URL: https://psycnet.apa.org/fulltext/2016-56587-001.html (visited on 2022).

— (2022). "Little Support for Discrete Item Limits in Visual Working Memory". In: *Psychological Science* 33.7, pp. 1128–1142. ISSN: 0956-7976. DOI: 10.1177/09567976211068045. URL: https://doi.org/10.1177/09567976211068045 (visited on 2023).

Oberauer, Klaus et al. (2018). "Benchmarks for Models of Short-Term and Working Memory." In: *Psychological Bulletin* 144.9, p. 885. ISSN: 1939-1455. DOI: 10.1037/bul0000153. URL: http://psycnet.apa.org/fulltext/2018-42071-001.pdf (visited on 2023).

Orhan, A Emin and Robert A Jacobs (2013). "A Probabilistic Clustering Theory of the Organization of Visual Short-Term Memory". In: *Psychological Review*, p. 32.

Peters, Benjamin, Benjamin Rahm, Stefan Czoschke, et al. (2018). "Sequential Whole Report Accesses Different States in Visual Working Memory". In: *Journal of Experimental Psychology. Learn-*

ing, Memory, and Cognition 44.4, pp. 588–603. ISSN: 1939-1285. DOI: 10.1037/xlm0000466. pmid: 28933904.

Peters, Benjamin, Benjamin Rahm, Jochen Kaiser, et al. (2019). "Differential Trajectories of Memory Quality and Guessing across Sequential Reports from Working Memory". In: *Journal of Vision* 19.7, pp. 3–3. ISSN: 1534-7362. DOI: 10.1167/19.7.3. URL: https://jov.arvojournals.org/article.aspx?articleid=2738012 (visited on 2020).

Popov, Vencislav (2023). "If God Handed Us the Ground-Truth Theory of Memory, How Would We Recognize It?" In: *Social and Behavioral Sciences; Cognitive Psychology; Memory* ay5cm (ay5cm). DOI: 10.31234/osf.io/ay5cm. URL: https://www.zora.uzh.ch/id/eprint/233473/ (visited on 2023).

Robison, Matthew K. and Nash Unsworth (2019). "Pupillometry Tracks Fluctuations in Working Memory Performance". In: *Attention, Perception, & Psychophysics* 81.2, pp. 407–419. ISSN: 1943-393X. DOI: 10.3758/s13414-018-1618-4. URL: https://doi.org/10.3758/s13414-018-1618-4 (visited on 2022).

Salvatier, John, Thomas V. Wiecki, and Christopher Fonnesbeck (2016). "Probabilistic Programming in Python Using PyMC3". In: *PeerJ Computer Science* 2, e55. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.55. URL: https://peerj.com/articles/cs-55 (visited on 2021).

Schneegans, Sebastian, Robert Taylor, and Paul M. Bays (2020). "Stochastic Sampling Provides a Unifying Account of Visual Working Memory Limits". In: *Proceedings of the National Academy of Sciences* 117.34, pp. 20959–20968. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2004306117. pmid: 32788373. URL: https://www.pnas.org/content/117/34/20959 (visited on 2020).

Schurgin, Mark W., John T. Wixted, and Timothy F. Brady (2020). "Psychophysical Scaling Reveals a Unified Theory of Visual Memory Strength". In: *Nature Human Behaviour* 4.11 (11), pp. 1156–1172. ISSN: 2397-3374. DOI: 10.1038/s41562-020-00938-0. URL: http://www.nature.com/articles/s41562-020-00938-0 (visited on 2023).

Swan, Garrett and Brad Wyble (2014). "The Binding Pool: A Model of Shared Neural Resources for Distinct Items in Visual Working Memory". In: *Attention, Perception, & Psychophysics* 76.7, pp. 2136–2157. ISSN: 1943-393X. DOI: 10.3758/s13414-014-0633-3. URL: https://doi.org/10.3758/s13414-014-0633-3 (visited on 2022).

Udale, Robert et al. (2021). "How Does Working Memory Store More Information at Larger Set Sizes? A Composite Code Model." In: DOI: 10.31234/osf.io/ndq9e. URL: https://psyarxiv.com/ndq9e/ (visited on 2022).

Utochkin, Igor S. and Timothy F. Brady (2020). "Individual Representations in Visual Working Memory Inherit Ensemble Properties". In: *Journal of Experimental Psychology. Human Perception and Performance* 46.5, pp. 458–473. ISSN: 1939-1277. DOI: 10.1037/xhp0000727. pmid: 32191111.

Van den Berg, Ronald et al. (2012). "Variability in Encoding Precision Accounts for Visual Short-Term Memory Limitations". In: *Proceedings of the National Academy of Sciences* 109.22, pp. 8780–8785. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1117465109. pmid: 22582168. URL: https://www.pnas.org/content/109/22/8780 (visited on 2019).

Wilken, Patrick and Wei Ji Ma (2004). "A Detection Theory Account of Change Detection". In: *Journal of Vision* 4.12, pp. 11–11. ISSN: 1534-7362. DOI: 10.1167/4.12.11. URL: https://jov.arvojournals.org/article.aspx?articleid=2192647 (visited on 2020).

Zhang, Weiwei and Steven J. Luck (2008). "Discrete Fixed-Resolution Representations in Visual Working Memory". In: *Nature* 453.7192, pp. 233–235. ISSN: 1476-4687. DOI: 10.1038/nature06860. URL: https://www.nature.com/articles/nature06860 (visited on 2019).

Supplementary figures

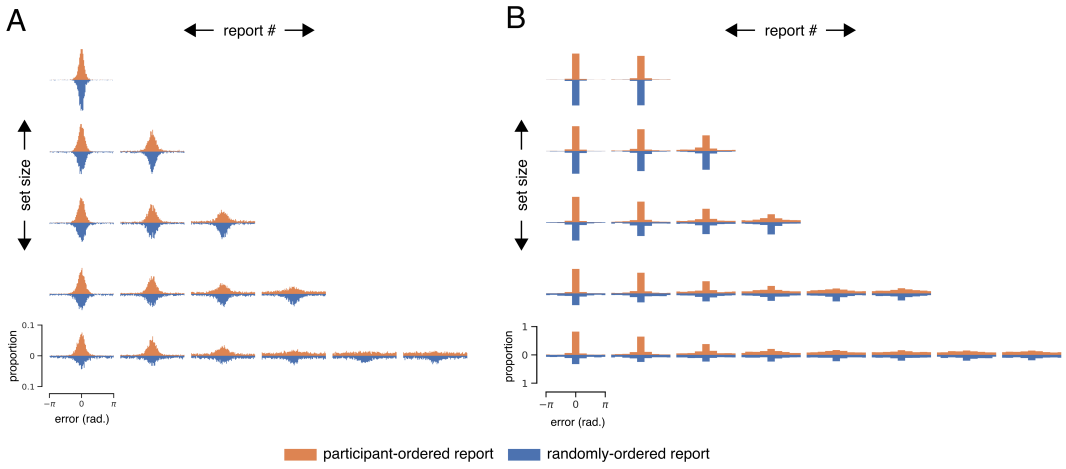

Figure 8. Color report error distributions for all set sizes, separated by report order. Includes all participants. **A.** Continuous error distributions. 360 possible error values divided into 90 bins for visualization. Orange and blue histograms show errors from participant-selected and randomly-generated report conditions, respectively. **B.** Discrete error distributions. Same as A, but bins correspond to all possible error values.

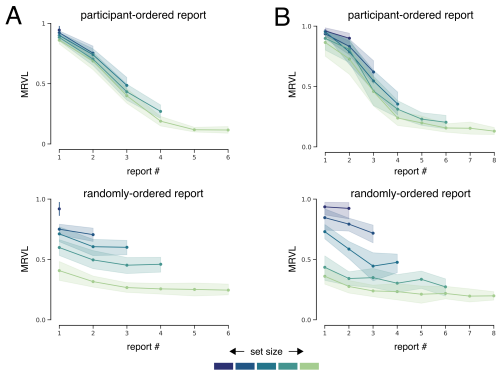

Figure 9. Mean resultant vector length (MRVL) of color report error distributions, separated by report order. **A.** Mean MRVL averaged across participants for the continuous whole report task. Shaded bars indicate +/- SEM. **B.** Same as A, but for the discrete task.

Figure 10. Within-trial joint color report distributions , separated by participant. **A.** Distribution of relative distances for the discrete task (set size 6). Each row shows data for one participant. **B.** Same as A, but for the continuous task.

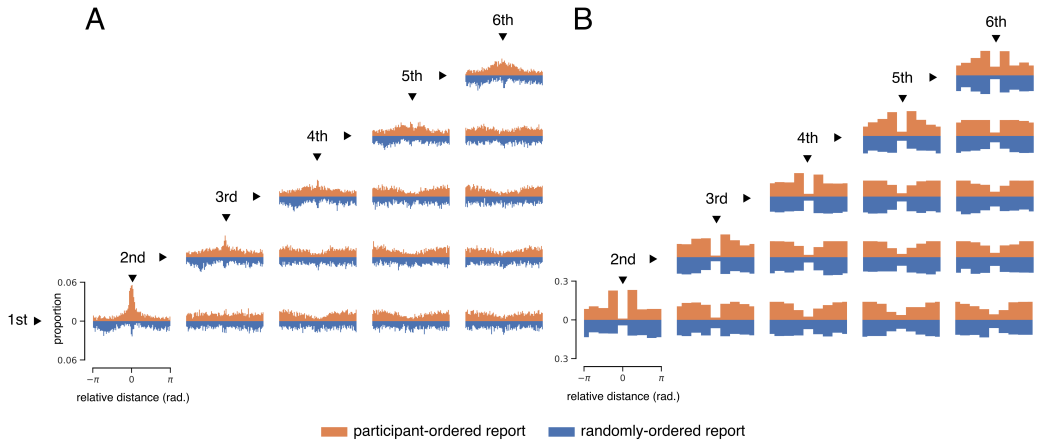

Figure 11. Within-trial joint color report distributions for all set sizes, collapsed across participants. **A.** Distribution of relative distances for the continuous task. Each row shows data for one set size. **B.** Same as A, but for the discrete task.

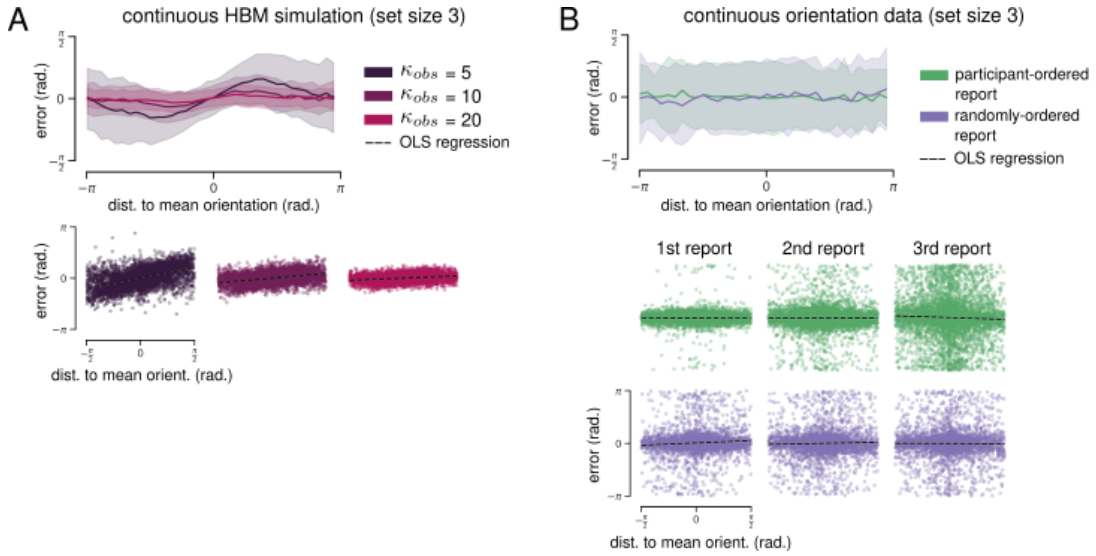

Figure 12. Hierarchical Bayesian model (HBM) simulation results for continuous task with orientation stimuli. **A.** *Upper:* Biases predicted by HBM simulations of the continuous task. Mean report error plotted as a function of the reported orientation's distance to the mean orientation. 360 error values are divided into 90 bins for visualization, and shaded area shows standard deviation. *Lower:* OLS regression of simulated report error on distance to mean orientation (dashed black lines). **B.** *Upper:* Empirical biases for the continuous task at set size 3 (collapsed across all reports). *Lower:* OLS regression of empirical report error on distance to mean orientation. Each column shows results for a different report number.

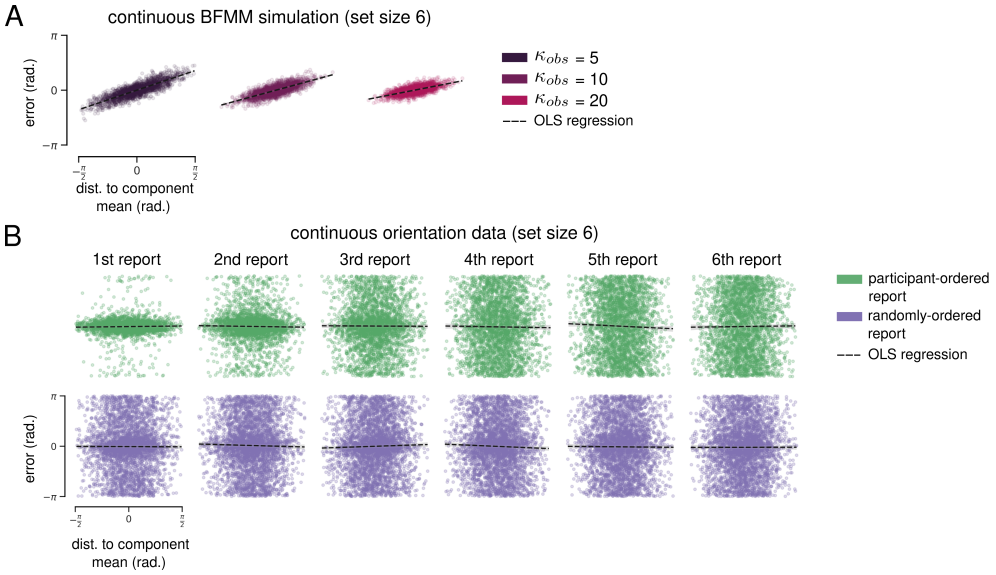

Figure 13. Bayesian finite mixture model (BFMM) simulation results for continuous task with orientation stimuli. **A.** *Upper:* Biases predicted by BFMM simulations of the continuous task. Mean report error plotted as a function of the reported orientation's distance to the component mean. 360 error values are divided into 90 bins for visualization, and shaded area shows standard deviation. *Lower:* OLS regression of simulated report error on distance to component mean (dashed black lines). **B.** OLS regression of empirical report error on distance to component mean. Each column shows results for a different report number.