Running Head: ENSEMBLES CONSTRAIN INDIVIDUAL MEMORIES

Individual representations in visual working memory inherit ensemble properties

Igor S. Utochkin

National Research University Higher School of Economics, Russia

Timothy F. Brady

University of California San Diego, USA

Corresponding author: Igor S. Utochkin

Address:

Psychology Department, Higher School of Economics

101000, Armyansky per., 4, room 419

Moscow, Russian Federation

E-mail: isutochkin@inbox.ru

Word Count (total): 12,730

Abstract

Prevailing theories of visual working memory assume that each encoded item is stored or forgotten as a separate unit independent from other items. Here, we show that items are not independent, and that the recalled orientation of an individual item is strongly influenced by the summary statistical representation of all items (ensemble representation). We find that not only is memory for an individual orientation substantially biased towards the mean orientation, but the precision of memory for an individual item also closely tracks the precision with which people store the mean orientation (which is, in turn, correlated with the physical range of orientations). Thus, individual items are reported more precisely when items on a trial are more similar. Moreover, the narrower the range of orientations present on a trial, the more participants appear to rely on the mean orientation as representative of all individuals. This can be observed not only when the range is carefully controlled, but also shown even in randomly generated, unstructured displays. Our results suggest that the information about a set of items is represented hierarchically, and that ensemble information can be an important source of information to constrain uncertain information about individuals.

Keywords: visual working memory; ensemble summary statistics; hierarchical encoding

Public Significance Statement:

When we need to remember multiple items at a time, we do not remember these items independently. Instead, properties of the entire set of items, like how variable it is, impact how precisely we remember each individual item.

Visual working memory is the cognitive system that maintains visual information to make it accessible for use in ongoing tasks (Baddeley, 1986; Baddeley & Hitch, 1974). This system has a severely limited capacity in terms of individual item information that can be held at a time (Alvarez & Cavanagh, 2004; Cowan, 2001; Luck & Vogel, 1997). The nature of this limited capacity is a highly debated topic (Brady, Konkle, & Alvarez, 2011; Luck & Vogel, 2013; Suchow, Fougnie, Brady, & Alvarez, 2015). For example, one important issue is whether visual working memory contains a fixed number of items in a discrete "slot" fashion (Luck & Vogel, 1997; Zhang & Luck, 2008) or can be allocated among variable number of items in a continuous "resource" fashion depending of the complexity of these units or task requirements (Bays & Husain, 2008; Bays, Catalao, & Husain, 2009; Ma, Husain, & Bays, 2014). Another interesting line of debate is whether objects in visual working memory are stored (and forgotten) as monolithic units with well bound features (e.g., Luck & Vogel, 1997, 2013; Raffone & Wolters, 2001) or as relatively independent features that can be swapped (Bays et al., 2009; Bays, Wu, & Husain, 2011; Pertzov, Dong, Pech, & Husain, 2012) or lost separately from other features (Fougnie & Alvarez, 2011) and require attention for binding (Wheeler & Treisman, 2002).

Importantly, both these and other areas of the literature are based on interpreting data from experimental paradigms in terms of the number of items that are stored and how precisely they are stored (Cowan, 2001; Luck & Vogel, 1997; Wilken & Ma, 2004; Zhang & Luck, 2008). The basic assumption of such theories and models is that every item is stored as a single representation of a given fidelity in visual working memory or not stored at all and that representations of different items are independent. However, this assumption has been challenged. For example, the framework called *hierarchical encoding* (Brady et al., 2011) argues for the idea of structured memory representations; in other words, it argues that information about the same set of memorized items is simultaneously stored at several levels of abstraction and, when recalled, information from a combination of these levels of abstractions are used. This idea relaxes the dichotomy between "independent features" and "bound objects" by suggesting that both features

and objects can be stored in visual working memory at different levels. For example, feature memory benefits when several features belong to the same object rather than different objects, but increasing the number of to be remembered features within an object can lead to interference and independent forgetting (Fougnie, Cormiea, & Alvarez, 2013; Fougnie, Asplund, & Marois, 2010). In a hierarchical framework, this is because instantiating new objects at the object-level has some form of capacity constraint, yet features within objects -- at a lower-level of abstraction -- are ultimately stored independently (Brady et al. 2011).

In other studies, it has been shown that hierarchical encoding goes beyond features and objects. In particular, people seem to represent higher-order structures of many items at one time (Brady & Alvarez, 2015a, 2015b; Brady & Tenenbaum, 2013; Jiang, Olson, & Chun, 2000; Nassar, Helmers & Frank, 2018; Orhan & Jacobs, 2013; Morey, Cong, Zheng, Price, & Morey, 2015; Son, Oh, Kang & Chong, 2019). These higher-order structures can be compared to a long-known concept of "chunks" (Miller, 1956). However, the hierarchical memory structures are proposed to be simultaneous represented, rather than an all-or-none combination of low-level features. Thus, whereas chunks are thought to represent an extended single-level structure (e.g., you can memorize stimuli "C", "A", and "T" as a word "CAT" without necessarily having any information about each individual letter or information about each individual line in each individual letter: Cowan, 2001), the hierarchical representation assumes that representations of both individual items and of a group of items are held in memory simultaneously and both are influential at retrieval. For example, Brady and Alvarez (2011) showed that when asked to remember the sizes of a set of colored dots, their participants seemed to rely on a combination of information about the individual sizes of each dot, as well as the mean size of the set of same-colored dots and the mean size of the set of all dots. This resulted in a relatively accurate memory that was nonetheless biased towards the mean size of all dots with the same color, and also biased toward the mean size of all dots. Brady and Alvarez (2011) concluded that along with individual size representations, the observers stored in visual working memory *ensemble summary statistics*

(for review, see Alvarez, 2011; Haberman & Whitney, 2012; Whitney & Yamanashi Leib, 2018): compressed and, hence, less memory-demanding descriptions of multiple objects (e.g., Ariely, 2001; Chong & Treisman, 2003, 2005) Remarkably, the results of Brady and Alvarez (2011) revealed at least three hierarchical levels: individual sizes, ensemble summaries for same-color subsets, and ensemble summary of all items. In her later study, Corbett (2017) showed that many basic Gestalt grouping factors can also give birth to hierarchical representations of this sort, and these biases have been found in other feature domains as well (e.g., in memory for faces: Corbin & Crawford, 2018; Griffiths, Rhodes, Jeffery, Palermo, & Neumann, 2018).

The bias towards the mean of sets of items found in previous studies (Brady & Alvarez, 2011; Corbett, 2017; Corbin & Crawford, 2018; Dubé, Zhou, Kahana, & Sekuler, 2014; Griffiths et al., 2018) is not the only consequence of hierarchical encoding in visual working memory. The mean is potentially the best descriptor of multiple items (Alvarez, 2011) but how well it represents the items depends on the overall feature distribution. The accuracy of the mean as a summary of the items decreases with more variable items, and it is known from the ensemble literature that the accuracy of computing the mean in a visual averaging task also tends to decrease as a function of the variability of individual features (Im & Halberda, 2013; Marchant, Simons, & De Fockert, 2013; Maule & Franklin, 2015; Utochkin & Tiurina, 2014). Consequently, if ensemble information is indeed used as a component of a hierarchical representation then the recalled trace of an individual item should also inherit the imprecision from the corresponding ensemble representation. That is, if people do rely on summaries of the entire set of items in their memory for each individual item, we predict that the features of an individual item should be recalled less precisely if all items are variable. By contrast, if items are fundamentally stored independently, there should be no effect of the feature values of other items on memory for any individual item.

To test this prediction, we designed an experiment using continuous recall of orientation (e.g., Bays et al., 2011; Fougnie & Alvarez, 2011; Fougnie et al., 2010, 2013; Zhang & Luck, 2008). In this experiment, we presented participants with sets of four items with different

orientations and manipulated their variability (the range of orientations). We asked participants to memorize the orientation of a particular precued triangle or the mean orientation of all items to obtain baseline parameters for individual and ensemble representations. In a critical condition, the participants had to memorize the individual orientations of all four items. We measured participants' performance using both non-parametric methods designed to measure the imprecision of memory and any bias toward the mean of the set of items, as well as via a mixture model (Zhang & Luck, 2008). Our primary question was how the imprecision of memory and any biases in memory were affected by the variability of the set of items as a whole. If ensemble information is used to in some way constrain individual item memories, than the variability of the set should affect how accurate memory for individual items is; if items are stored independently, as in many influential models, then the variability of the set should be relevant for ensemble judgments but not item memory.

Experiment 1

All experimental materials, raw data, and the results of the analysis for this and subsequent experiments are available on OSF: https://osf.io/v7yde/.

Method

Participants

Sixteen students of the Higher School of Economics (10 female; age 18-21) took part in the experiment for course credit. With this sample size, we could detect effect size estimates as small as $\eta^2 = .4$ and Cohen's $d_z = .9$ given the Holm correction (where applicable) for the $\alpha = .05$ and a power of .8 (Faul, Erdfelder, Lang, & Buchner, 2007). All participants reported having normal or corrected-to-normal vision and no neurological problems. Before the beginning of the experiment, participants gave informed consent.

Apparatus and stimuli

Stimuli were presented using PsychoPy (Peirce, 2007) for Linux on a standard VGA monitor (75 Hz at 1024 × 728 resolution) on a homogeneous gray field. Participants sat

approximately 47 cm from the monitor. From that distance, each pixel subtended 0.054º of visual angle.

Memory displays consisted of four white isosceles triangles with different apex orientations. The bases and altitudes of the triangles were 2.7º and 3.9º, respectively. The triangles were centered on an imaginary circle with a radius of 10.8º and occupied cardinal positions corresponding to 45º, 135º, 225º, and 315º of rotation on that circle; random positional jitter between –10º and 10º was added to each of the locations.

The orientations of the sample triangles were generated on each trial according to the following rule. A random angle was chosen from between 1º and 360º to serve as the mean orientation of all triangles. The individual orientations were then constrained to a particular range of orientations around this value, which we varied across conditions, such that the orientations covered a range of either 30º, 60º, or 120º (always centered at this mean orientation). The values were equally spaced in this range and jittered by ±3º. Therefore, for the 30º range, the individual values were –15º, –5º, 5º, and 15º away from the mean (±jitter); for the 60º range, the individual values were –30º, –10º, 10º, and 30º away from the mean (±jitter); for the 120º range, the individual values were –60º, –20º, 20º, and 60º away from the mean (±jitter). Therefore, the mean orientation was never physically present as a member of the set. On average, individual items were 10º, 20º and 40º from the mean orientation in the 3 conditions. The individual values were randomly assigned to the four locations on the sample screen. Examples of a sample display as a function of the range are given in Figure 1A. Each of the individual orientations was equally likely to be probed at test in the tasks demanding memory for individual triangles (see task descriptions below).

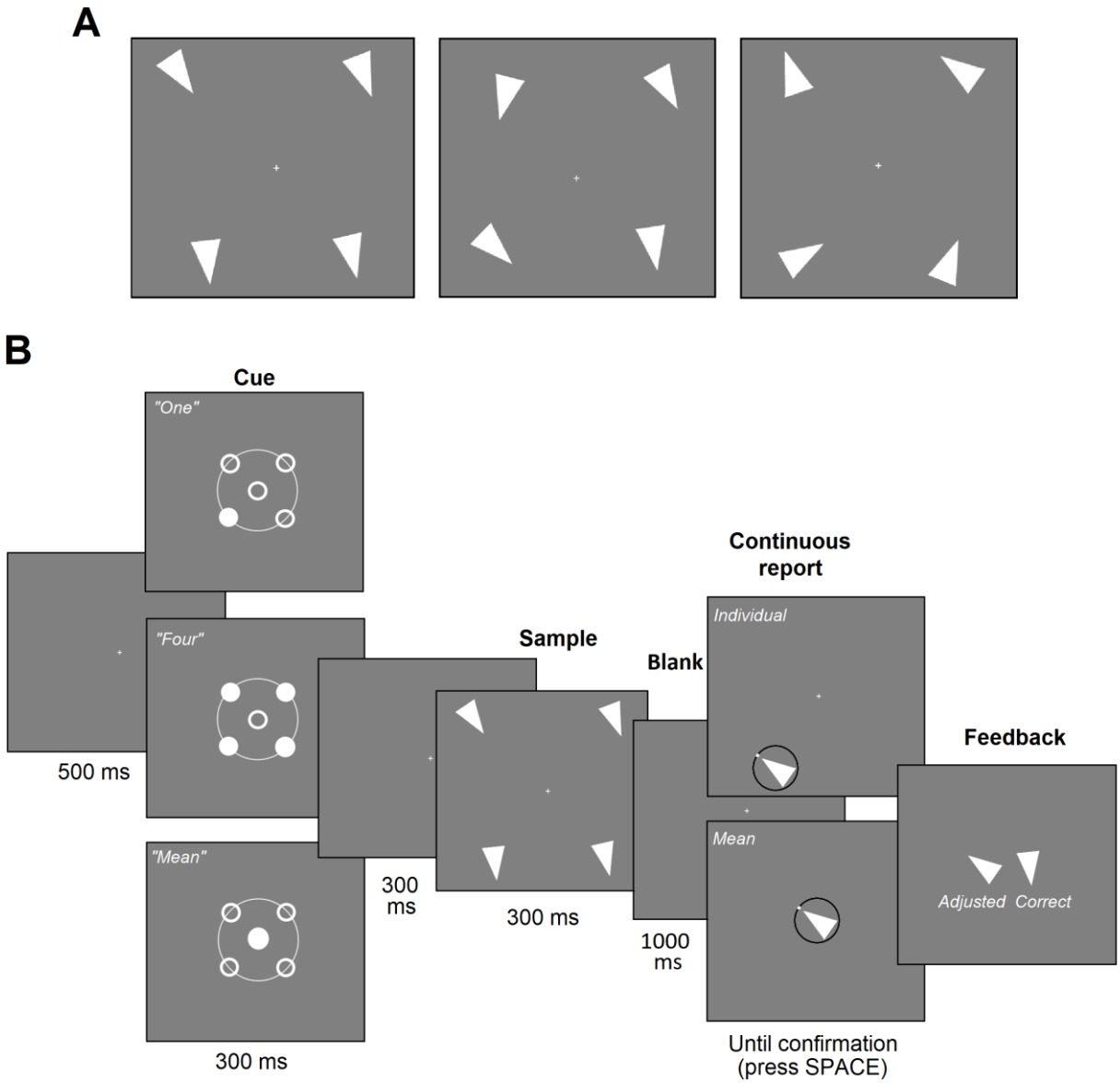

Figure 1. Methods of the experiment. (A) Example stimuli of different ranges (left to right: ranges of 30º, 60º, and 120º); (B) The time course of a typical trial. Participants first saw a cue indicating whether they needed to remember 1 item, 4 items, or the mean orientation; followed by a brief delay and then the study display. After a 1000ms retention interval, participants were then probed on either an individual item or the mean orientation, followed by feedback. Cues to which item would be probed (or whether the mean would be probed) were always 100% valid and the tasks were done in separate blocks.

Procedure

The experiment consisted of three visual working memory tasks. Depending on the task, participants were instructed to remember and recall (1) the orientation of a single triangle precued in advance ("Remember One" task), (2) the orientations of all four triangles ("Remember Four" task), or (3) the mean orientation of all four triangles ("Remember Mean" task). The tasks were run in separate blocks of 120 trials. Each task was presented in two blocks arranged in a mirror

order (e.g., block 1 and 6, or2 and 5, or 3 and 4). The order of tasks within the first set of 3 blocks was randomly assigned to each participant. Each block of 120 trials was preceded by six practice trials that served as familiarization with the next task.

Trials (Figure 1B) started with a fixation cross in the center of a screen for 500 ms. This was followed by a 300-ms cue informing participants about the memorized attribute. The cue was a white icon (approximately 9.2º×9.2º) depicting the four locations of sample items in a circular arrangement and a central position for the mean orientation. Depending on the task, certain circles would be filled indicating the relevant attribute to report. In the "Remember One" task, one of the randomly chosen circle locations was cued to indicate which particular triangle was to be memorized. In the "Remember Four" task, all four individual locations were cued. In the "Remember Mean" task, the central circle was cued. Although cues provided no new information in the "Remember Four" and "Remember Mean" tasks (given the blocked design of the experiment), we used them to make the sequence of events in a trial the same across all three tasks.

After a 300 ms delay following the cue offset, a sample display was presented for 300 ms. It was followed by a 1,000-ms retention interval, and then the test display was presented. In that display, a single white triangle with an adjustable orientation was presented either at one of the sample locations randomly assigned for that trial ("Remember One" or "Remember Four" tasks), or in the center of the screen ("Remember Mean" task). The initial orientation of the test was set randomly. The test triangle was surrounded by a black orientation wheel with a white slider that could be rotated with the mouse to make the triangle rotate (Figure 1B). Participants were instructed to adjust the orientation of the test triangle to be as close as possible to the individual orientation or the mean orientation. To confirm their response, participants had to press the spacebar. Response confirmation was followed by feedback showing the adjusted orientation on the left and a correct orientation on the right. The feedback remained on the screen until the participant pressed the spacebar to start the next trial. The feedback screen could be used by participants to have a break any time they needed.

Design and analysis

The experiment had a 3 (Task: Remember One vs. Remember Four vs. Remember Mean) × 3 (Range: 30º vs. 60º vs. 120º) within-subject design. Within each cell of the design, a participant was exposed to 80 trials. Therefore, the total number of trials was 720 per participant (without considering the practice trials at the beginning of each block).

Error distributions. The principal measure of visual working memory performance on each trial was the difference in degrees between the adjusted orientation and the correct answer. Given the circular nature of orientation and the directional nature of the triangle stimuli we used, these errors covered a 360º range and thus fell between –180º and 180º. Traditionally, positive errors are clockwise and negative errors counterclockwise. However, this essentially eliminates any capacity to detect a systematic effect of the ensemble mean, because the direction of such a bias depends the position of the item relative to the mean. Thus, for each trial, we unified the directionality of errors in relation to the mean orientation: We reversed the sign of the error in trials where tested items were clockwise relative to the mean. This transformation was applied only to the "Remember One" and the "Remember Four" tasks. We did not apply this transformation to the "Remember Mean" task. Thus, in the individual item memory tasks, positive error always indicates errors towards the mean and negative error always indicates error away from the mean. Importantly, this unitization changed mostly the sign of the errors but did not strongly affect the dispersion of data in overall distributions. Obviously, flipping the sign of some errors would have no effect whatsoever on some measures of error, like root mean square error, which simply ask about the magnitude of difference from 0 separately for each error. However, for the measure we use, the the angular deviation, which is the analogue of the standard deviation in circular data (Berens, 2009; Zar, 1999; equation 26.20), there is a small effect of this unitization because it impacts the clustering of the errors, which is what is measured by this index. Nevertheless, angular deviation did not differ a substantial amount between the unitized and non-unitized error distributions in all ranges of the “Remember One” task (mean differences < 0.6º, *t*’s

< 2.3, p's > .03, Bonferroni corrected α = .17, d_z < .6) and were slightly though consistently smaller in all ranges of measure of the "Remember Four" task (mean differences = 2-3º, t's > 4.9, p's < .001, Bonferroni corrected α = .17, d_z > 1.8). Importantly, non-unitized and unitized angular deviations were very highly correlated in both tasks and all ranges (r's > .95, p's < .001). This suggests that error unitization we used to be capable of detecting a bias towards or away from the mean did not strongly distort the rest of the information necessary to judge other critical distributional parameters.

Memory performance from error distributions. We applied two methods of evaluating visual working memory performance from the error distribution. The first method is non-parametric and is based on summary statistics: the circular mean as a measure of bias and the angular deviation as a measure of imprecision. The angular deviation is a circular analogue of the standard deviation (Berens, 2009; Zar, 1999) and has been recommended as an ideal measure because despite being straightforward and non-parametric, it is closely related to model-based measures of performance like *d'* (Schurgin, Wixted & Brady, 2018). In addition to these non-parametric methods, we also used a three-parameter mixture model to estimate visual working memory performance from the error distributions (Zhang & Luck, 2008) to assess the robustness of the conclusions using a more common method (although see Schurgin et al., 2018, who argue that the parameters estimated from this method do not reflect distinct psychological constructs). The model fits two distributional components: a von Mises component (Gaussian-like distribution for circular dimensions) which describes responses nearby the correct answer as noisy item-based responses, and a uniform component describing 'guess' responses to elements that are assumed not to be successfully stored. The two parameters extracted from the von Mises component are the mean (μ) and the standard deviation (*SD*) reflecting the systematic bias and the imprecision of the memory representation. The third parameter is extracted from the uniform component and is usually interpreted as the probability of 'guesses' (P_{guess}), an estimate of how many of presented

items cannot be retrieved. The mixture models were implemented in MemToolbox (Suchow, Brady, Fougnie, & Alvarez, 2013).

A 3×3 repeated measure ANOVA was used to statistically estimate the effects of the Task and the Range on the parameters obtained from these summaries.

Results

The pattern of errors for each condition and each range of orientation is plotted in Figure 2. In all plots, the errors are flipped such that errors toward the mean are plotted as positive and errors away from the mean are plotted as negative.

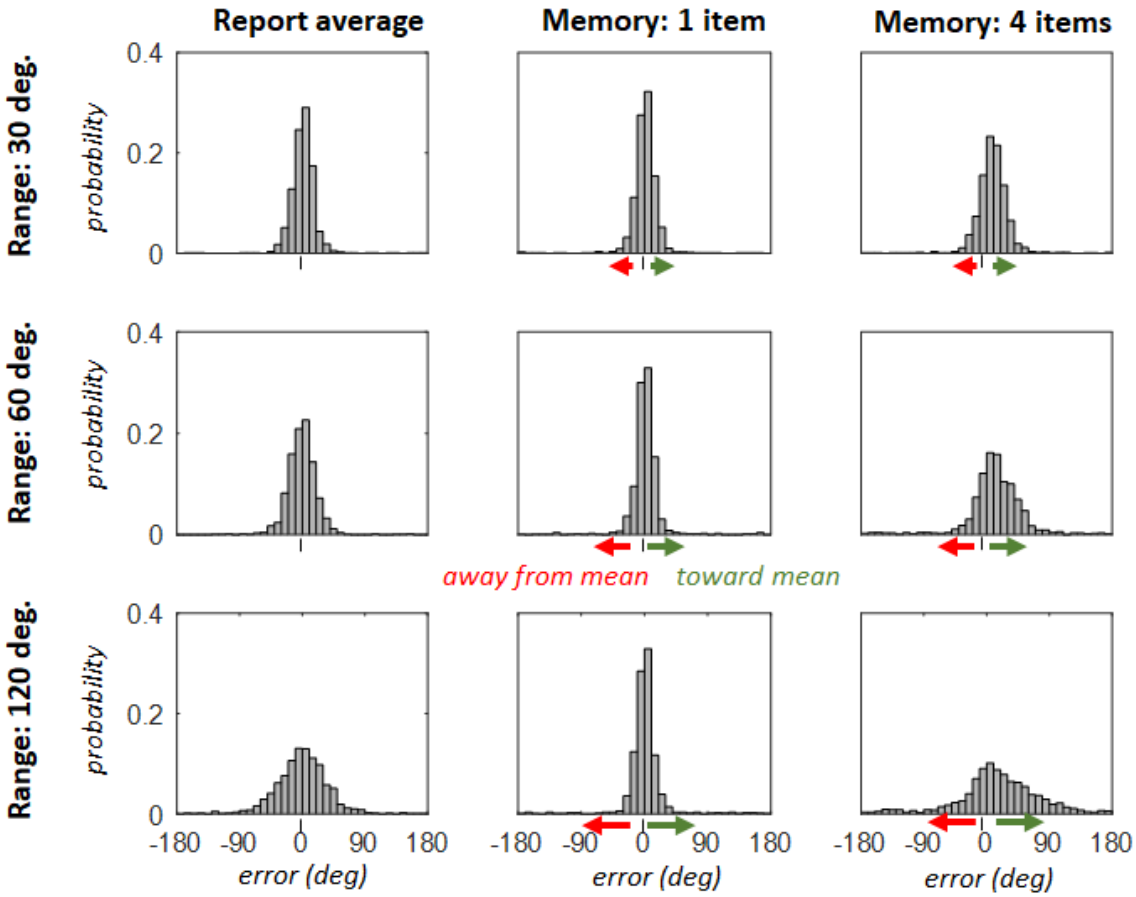

Figure 2. Results of Experiment 1. Histogram of data from each condition pooled across participants, with all errors in the two individual-item memory conditions flipped so that errors toward the mean of the set of items are positive and errors away from the mean of the set of items are negative. Each column represents a different task (Report the average; Memory for 1 item; Memory for 4 items), and each row represents a different range condition (all 4 items within 30º; within 60º; within 120º).

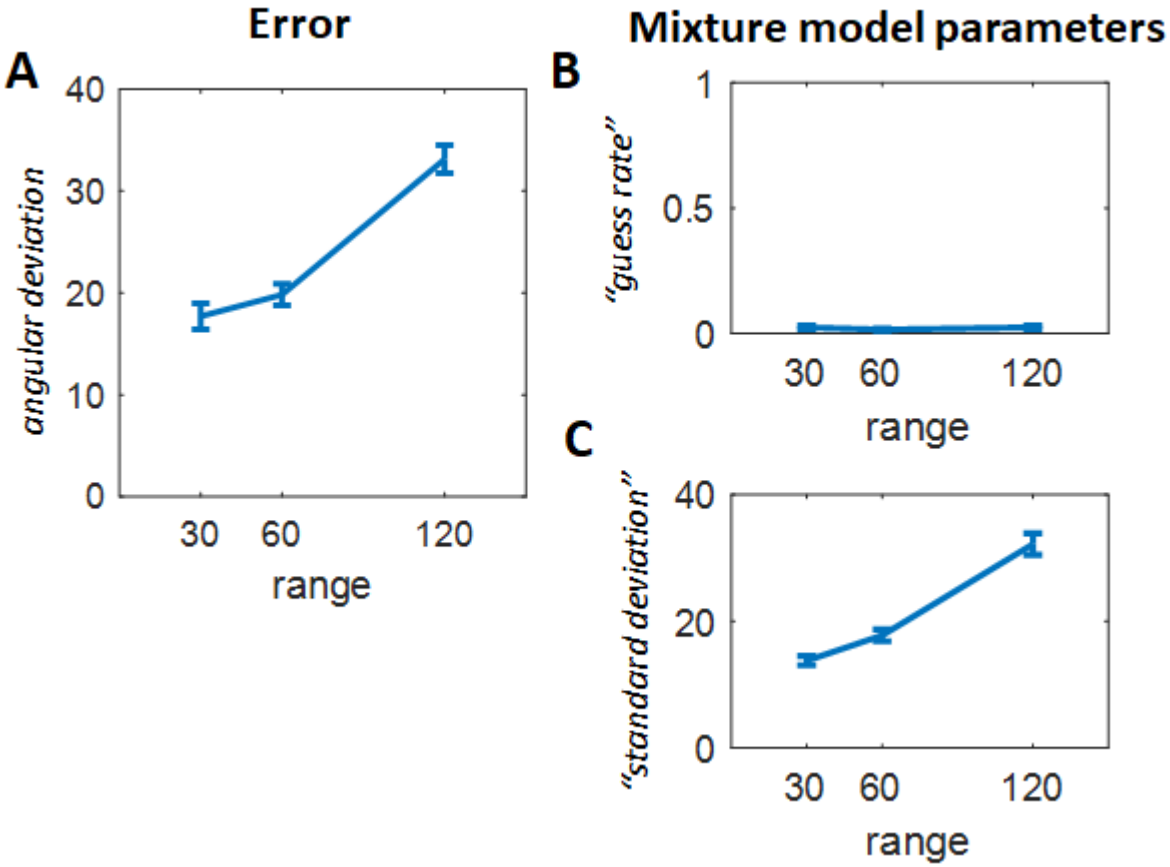

Figure 3. Performance at the ensemble (report the average orientation) task. (A) Performance assessed using a non-parametric measure of error, the angular deviation, shows that performance is much better when the range is small than when the range is large. (B) Similar results are obtained using the mixture model, which shows very few lapses (P_{guess}) and (C) a similar effect to the non-parametric analyses in terms of the standard deviation (*SD*) of the von Mises distribution component of the mixture model. Error bars denote ±1 standard error of the mean.

Precision. We found strong effects of the task on our non-parametric estimate of error ($F(2,30) = 63.91$, $p < .001$, $\eta^2 = .81$) and on the *SD* parameter of the mixture models ($F(2,30) = 199.89$, $p < .001$, $\eta^2 = .93$) showing that orientation reports in the "Remember Mean" task were overall noisier than in the "Remember One" task (non-parametric: $t(47) = 3.51$, $p < .001$, $d_z = .58$; mixture model: $t(47) = 7.37$, $p < .001$, $d_z = 1.06$), and orientation reports in the "Remember Four" task were noisier than in the "Remember Mean" task (non-parametric: $t(47) = 11.62$, $p < .001$, $d_z = 1.67$, mixture model: $t(47) = 7.41$, $p < .001$, $d_z = 1.07$). The effect of the range was also strong (non-parametric: $F(2,30) = 178.98$, $p < .001$, $\eta^2 = .92$, mixture model: $F(2,30) = 160.60$, $p < .001$, $\eta^2 = .92$) reflecting the overall *SD* growth with range. In fact, this growth was task-specific (task × range effect for non-parametric error: $F(4,60) = 45.05$, $p < .001$, $\eta^2 = .37$; for the mixture model

SD: $F(4,60) = 68.36$, $p < .001$, $\eta^2 = .82$). The non-parametric error or *SD* did not differ between ranges in the "Remember One" task (*t*'s(15) = [0.89, 1.78], *p*'s ≥ .096, d_z = [.22, .45]) but grew steadily with the range in the "Remember Four" and the "Remember Mean" tasks (*t*'s(15) = [7.81, 24.29], *p*'s < .001, d_z = [1.95, 6.07]; though for the non-parametric error, there was no substantial growth between the 30º and the 60º ranges: ($t(15) = 1.72$, $p = .11$, $d_z = .43$). Overall, both non-parametric and mixture-model estimates of the error showed basically same patterns. The strong similarity between the non-parametric and the mixture model errors can be seen comparatively in Figures 3A and 3C ("Remember Mean" task), and also in Figures 4A and 5A ("Remember One" and "Remember Four" tasks). In Figure 7 with the aggregated model fits, it can be seen that all "Remember One" (thin solid lines) representations have approximately the same width (*SD* = 11-12º); in contrast, in the "Remember Four" and the "Remember Mean" tasks the width of the distributions tend to increase along with the bias away from zero, with both increasing with range.

Guess rate. The mixture model claims a dissociation between precision and 'guess rate'. Thus, while the non-parametric error combines all the data, the mixture model separately estimates the effect of the manipulation on the central part of the distribution and the tail of the distribution. Broadly, however, we find the same effects on 'guess rate' as on 'precision': In particular, the main effects of the task and the range on P_{guess} were strong (task: $F(2,30) = 19.36$, $p < .001$, $\eta^2 = .57$; range: $F(2,30) = 15.95$, $p < .001$, $\eta^2 = .52$) that in fact were provided by the range effect specific for the "Remember Four" task (task × range effect: $F(4,60) = 11.31$, $p < .001$, $\eta^2 = .43$). Indeed, the range had no effect on P_{guess} in the "Remember One" task (*t*'s(15) = [1.35, 2.16], *p*'s = [.05, .20], all *p*-values were larger than Holm corrected α, d_z = [.34, .54], Figure 5A, lower panel) or in the "Remember Mean" task (*t*'s(15) = [.25, 1.65], *p*'s = [.12, .81], d_z = [.06, .41], Figure 3B). The overall P_{guess} in these two tasks was within .02-.06 on average. In the "Remember Four" task, P_{guess} tended to increase with the range from 0.05 to 0.17 (*t*'s(15) = [2.28, 5.04], *p*'s ≤ .005, all *p*-values were smaller than Holm corrected α, d_z = [.82, 1.26], Figure 5C) such that with higher

orientation variability, not only did imprecision as measured by the mixture model increase but so did P_{guess}.

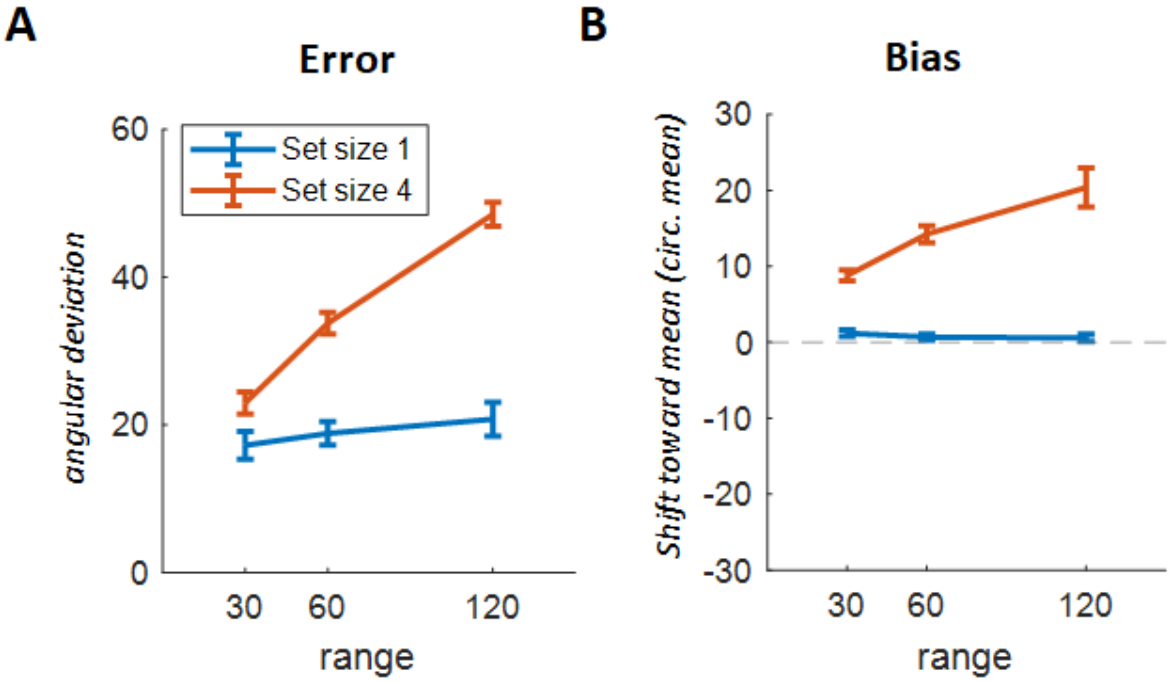

Figure 4. Performance at the individual item memory task measured non-parametrically. (A) Non-parametric measures of error show that despite the task being to report memory for a single item, increases in the range of all of the items resulted in large changes in error, particularly at set size 4. (B) Non-parametric measures of bias (the circular mean of the error distribution) showed a reliable bias for participants to report items as closer to the mean than they really were at set size 4. In absolute terms, this bias increased as a function of the range of the items, though relative to the actual location of the mean of the items (+10º; +20º; +40º), it decreased with range (see also simulation results).

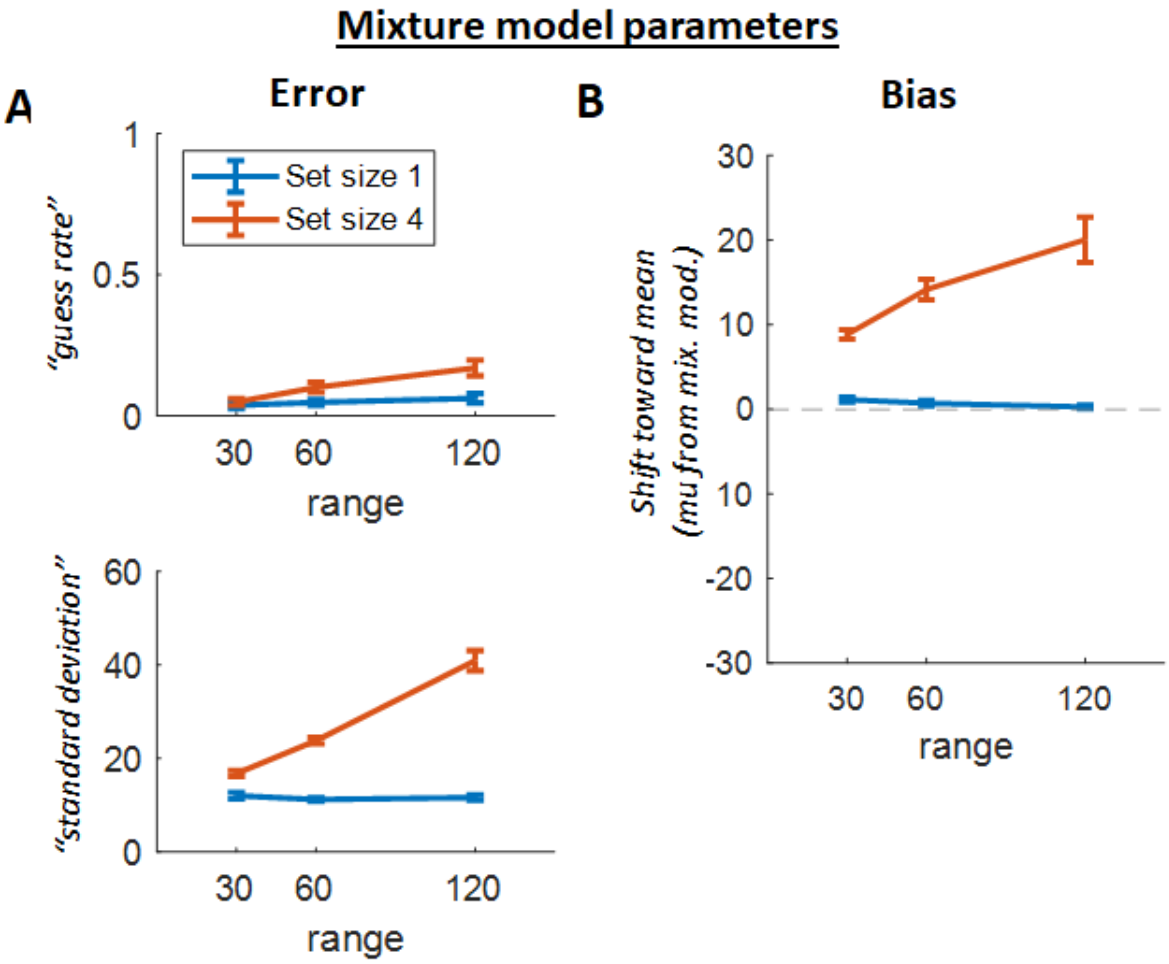

Figure 5. **Performance at the individual item memory task measured via a mixture model.** (A) P_{guess} and standard deviation (*SD*), a measure of imprecision, at set size 1 and 4; (B) Bias of the von Mises distribution component at set size 1 and 4. Error bars denote ±1 standard error of the mean.

Biases. As the bias relative to the mean was not informative in the "Remember Mean" task, we estimated the effects on bias only for the "Remember One" and for "Remember Four" tasks. We found that the task had a strong effect both on the non-parametric bias measure ($F(1,15)$ = 136.68, $p < .001$, η^2 = .90) and on the parametric one extracted from the mixture model ($F(1,15)$ = 133.11, $p < .001$, η^2 = .90). This is driven by the "Remember Four" task in which the responses were overall substantially biased in the positive (towards the mean) direction (M = 15º for the non-parametric bias, M = 14º for the parametric bias, one-sample comparisons with the null bias: $t(47)$ = [12.10, 12.37], $p < .001$, d_z = [1.75, 1.79]). By contrast, the "Remember One" task had an extremely small magnitude of bias, though this bias was systematic (M = .7º-.8º, $t(47)$ = [3.35, 4.02], $p < .001$, d_z = [.47, .58]). The main effect of the range also was significant in both non-

parametric ($F(2,30) = 12.99$, $p < .001$, $\eta^2 = .46$) and parametric ($F(2,30) = 10.46$, $p < .001$, $\eta^2 = .41$) methods. In fact, this effect was provided by a strong range effect within the "Remember Four" task that is supported by the task × range interaction ($F(2,30) = [15.10, 17.14]$, $p < .001$, $\eta^2 = [.50, .53]$). In this task, the bias increased with the range (t's(15) = [2.28, 6.07], p's ≤ .038, all p-values are less than Holm corrected α's, Cohen's d_z's = [.57, 1.52]); in the "Remember One" task, there were no range effect on the bias (t's(15) = [.18, 1.98], p's ≥ .067, none of the p-values are less than Holm corrected α's, Cohen's d_z's = [.04, .50]). The strong similarity between the non-parametric and the mixture model biases can be seen comparatively in Figures 4B and 5B.

Overall, then, we find that contrary to the assumption of independent representation, not only do judgments of the mean orientation become less precise with increasing range, but memory for individual items, particularly at set size 4, also become less precise with increasing range. In addition, memory for individual items is reliably biased toward the mean orientation.

Simulation 1: Simulation of responding based only on the mean

What would participants' errors in the individual item task look like if they relied solely on the mean orientation rather than any information about individuals? Understanding this can help contextualize the extent to which participants used individual information vs. relied on information about the mean orientation in the "Remember Four" condition. In addition, since the average distance between the tested item and the mean increases with the range, it is possible, although unlikely, that the range effect in the "Remember Four" task could largely be explained by a simple strategy of reporting the mean orientation instead of an individual one. In particular, because the items are similar to each other, participants could in theory have chosen to simply report the mean orientation in both the "Remember Mean" and "Remember Four" task rather than making an effort to remember individual item information in the "Remember Four" task.

To assess what performance would look like if participants relied solely on their knowledge of the mean orientation, we compared errors in the "Remember Four" task with those found in the "Remember Mean" task. In particular, we asked what errors for an individual item would look like if participants relied solely on their knowledge of the mean orientation, as assessed in the "Remember Mean" task. Thus, we added the average distances between the actually tested orientation and the mean orientations to the error values obtained in the "Remember Mean" task. This transformation was applied individually for each range condition in each participant to simulate what the "Remember Four" condition would look like if people only used information about the mean orientation ("Simulated-From-Mean-Only").

This transformation created responses from the viewpoint of an average individual tested item in the "Remember Four" condition. We then compared the "Remember Four" responses (Figure 6, left) with the "Simulated-From-Mean-Only" responses (Figure 6, middle). As can be seen in Figure 6 (right column), even at the smallest range, which shows a proportionally large bias toward the mean, the simulation resulted in error distributions that underestimated the number of responses near the individual item (near 0) and overestimated the number of responses near the mean. We next summarized the "Simulated-From-Mean-Only" responses in terms of the prediction they make about how biased toward the mean responses should be, and comparing them with the "Remember Four" biases in the three ranges (2 × 3 repeated-measure ANOVA). We found the strong effect of the task ($F(1,15) = 37.81$, $p < .001$, $\eta^2 = .716$) showing that "Simulated-From-Mean-Only" responses were more biased than the "Remember Four" responses . We also found a strong task × range effect ($F(1,15) = 21.00$, $p < .001$, $\eta^2 = .583$) reflecting the increasing divergence between the transformed "Remember Mean" and "Remember Four" with range.

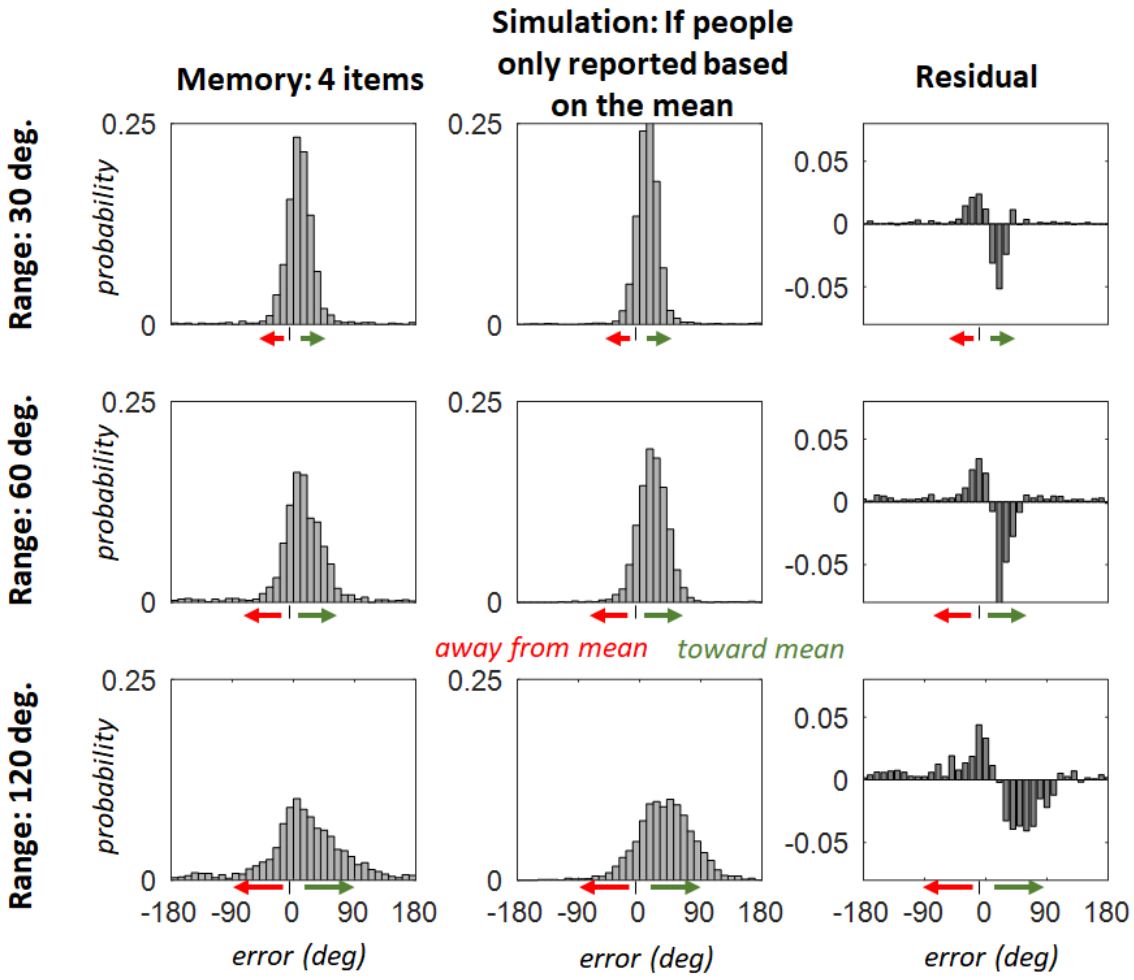

Figure 6. Simulation of data if people reported only based on their (measured) knowledge of the mean. **Left**: Data from "Remember Four" condition. **Middle**: Data pattern expected if participants solely used their knowledge of the mean (assessed via the "Remember Mean" condition) to respond in the Remember Four condition; **Right**: Residual of this model; e.g., Column 1 minus Column 2. As can clearly be seen, responses based solely on the mean would be too biased and contain too few responses near 0 (e.g., near the correct individual item answer) if they were based solely on the mean.

Specifically, in the 30º range the biases were similar between these two tasks (M = 10º in the "Simulated-From-Mean-Only" vs. M = 9º in the "Remember Four"; comparison: $t(15) = 1.82$, $p = .089$, $d_z = .45$), but the difference substantially increased in the 60º range (M = 21º vs. 14º, respectively; comparison: $t(15) = 4.07$, $p < .001$, $d_z = 1.02$) and especially in the 120º range (M = 38º vs. 20º, respectively; comparison: $t(15) = 5.78$, $p < .001$, $d_z = 1.44$).

While the full histogram (Figure 6) shows that even at low variance (range 30º) participants data is not totally compatible with reporting only the mean, the average biases suggests that reporting the mean instead of the individual orientations is similar to participants data in low-

variance trials but cannot account for their behavior at all in trials with higher variance. Although the responses to individual items are biased to the mean, the intermediate bias values suggest that these responses are also strongly affected by real individual orientations. The intermediate biases in "Remember Four" reports are also visualized in Figure 7, which depicts the distribution fits aggregated across participants and aligned along a scale having an individual tested item as reference point (Error = 0): It can be seen that the peaks of "Remember Four" distributions (thick solid lines) in the 60º and the 120º ranges are shifted to the right from "Remember One" distributions (thin solid lines) and to the left from the corresponding "Simulated-From-Mean-Only" distributions (thin dashed lines).

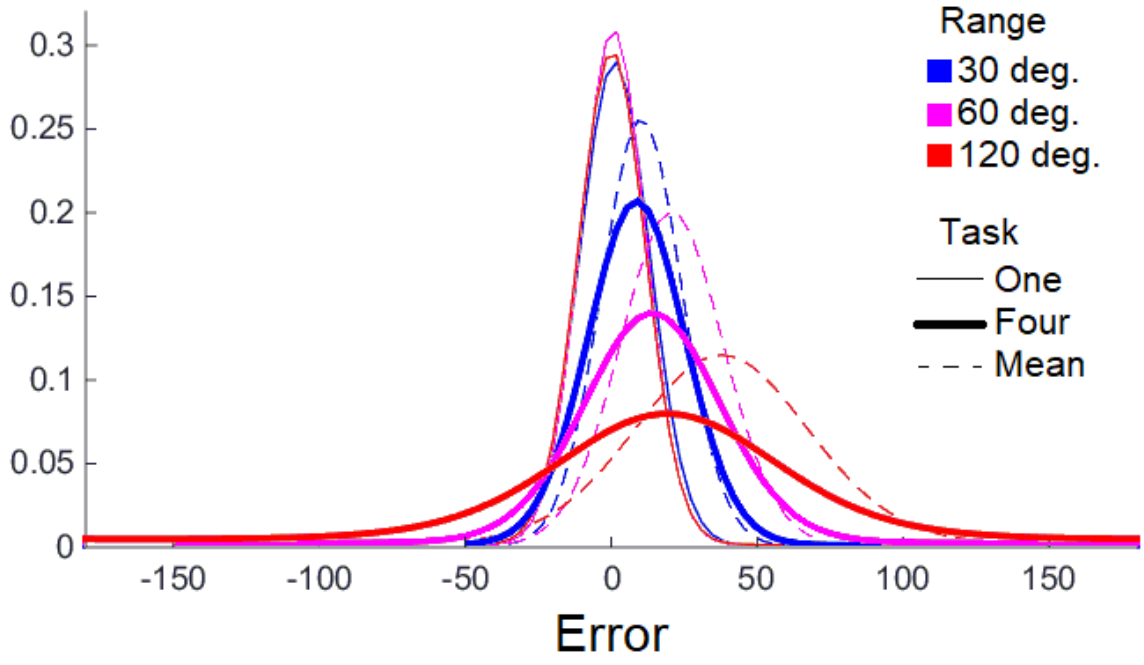

Figure 7. Distribution fits of the mixture models centered on individual tested orientations averaged across participants. The 'Mean" task reflects the fits from the "Simulated-From-Mean-Only"; e,g., what errors would be found for the tasks if participants solely used their knowledge of the mean to respond. Colors are used for different orientation ranges; line types are used for different tasks. Importantly, even in the "Remember Four" condition, participants are far less biased than expected from a strategy of relying solely on the mean.

Simulation 2: Location-based confusions with other items

Is it possible that people do not use the range or mean of the display at all, but simply have a fixed rate of "swaps"? Mistakenly reporting incorrect items would be expected to introduce

greater error at larger ranges, consistent with the direction of our results, and models based on "swaps" are popular in the literature on visual working memory (e.g., Bays et al., 2009).

When items are tightly clustered, as in our displays, it is quite difficult to use the pattern of errors in any condition alone to distinguish the most general form of "swap" model from a model based on a mixture of mean-based and item-based responding ('hierarchical encoding'). However, the more general proposal of "swaps" can be instantiated in various forms, and there are at least two (dissociable) theoretical accounts that can be thought of as "swapping": one is the original proposal of Bays et al. (2009), where swaps are largely based on spatial confusions about which location is being probed, which is consistently the version of 'swapping' found in data with randomly generated displays (e.g., Emrich & Ferber, 2012; Oberauer & Lin, 2017). In this account, since spatial distance to the probed item was unrelated to the range and unrelated to the similarity to the target and mean, the 'swap rate' should be the same for items across all ranges and for items that are similar or dissimilar to the target. The other kind of "swap" model is quite different, and effectively another way of saying people take into account the distribution of the features of the other items in making their responses: In particular, one possible response strategy that participants could use -- an ensemble-based strategy -- would be to simply limit the responses to be within the plausible range of the display, which is similar to a swap-based account but based on an efficient encoding strategy, rather than an error from location noise. Under this account, we would predict different rates of estimated "swaps" to items close and far from the mean, and different rates on displays with different ranges.

To tell these apart, we asked whether a model (Bays et al., 2009) that estimates "swap rate" -- assuming all errors arise from either correct responses, 'guesses' or swaps, with no role for direct representations of the display mean or range -- finds a fixed swap rate across ranges. If this "swapping" did not take into account the range or distribution of other items on the display at all, as in the case where it arose primarily from location uncertainty, we would expect this swap rate to be very similar across all ranges. If instead it reflects some form of strategic responding based

on the ensemble of the display, then we would expect this "swap rate" to be higher when the items are more clustered in feature space. In this way we can distinguish whether ensemble-based responding is occurring without directly attempting to distinguish between ensemble mean and item-based responding or a more "swap"-based version of an ensemble strategy.

Consistent with the ensemble-based account, we find that the swap rate estimates are much higher for displays where the items are more tightly clustered in orientation (e.g., smaller ranges): at range 30º, 60º, and 120º, respectively, the estimated swap rates are 56.8% (SEM: 3.2%), 47.1% (SEM: 3.6%) and 33.0% (SEM: 4.2%), a significant difference ($F(2,30)=17.2$, $p<0.0001$) and each larger range has a significantly lower swap rate than the tighter range (e.g., 30 vs. 60: $t(15)=3.28$, $p=0.005$, $dz=0.82$; 60 xs. 120: $t(15)=3.27$, $p=0.005$, $dz=0.82$). Furthermore, the precision estimates, even after partially out such a huge number of putative 'swaps', are still less precise at larger ranges even in this mixture model: (M = 12.4º, 14.4º, 26.0º; $F(2,30)=18.7$, $p<0.0001$). Thus, even if we assume no direct effect of the ensemble mean of the display, but simply reports of other items, we find that the range of items on the display must mediate how likely subjects are to rely on other items in their reports.

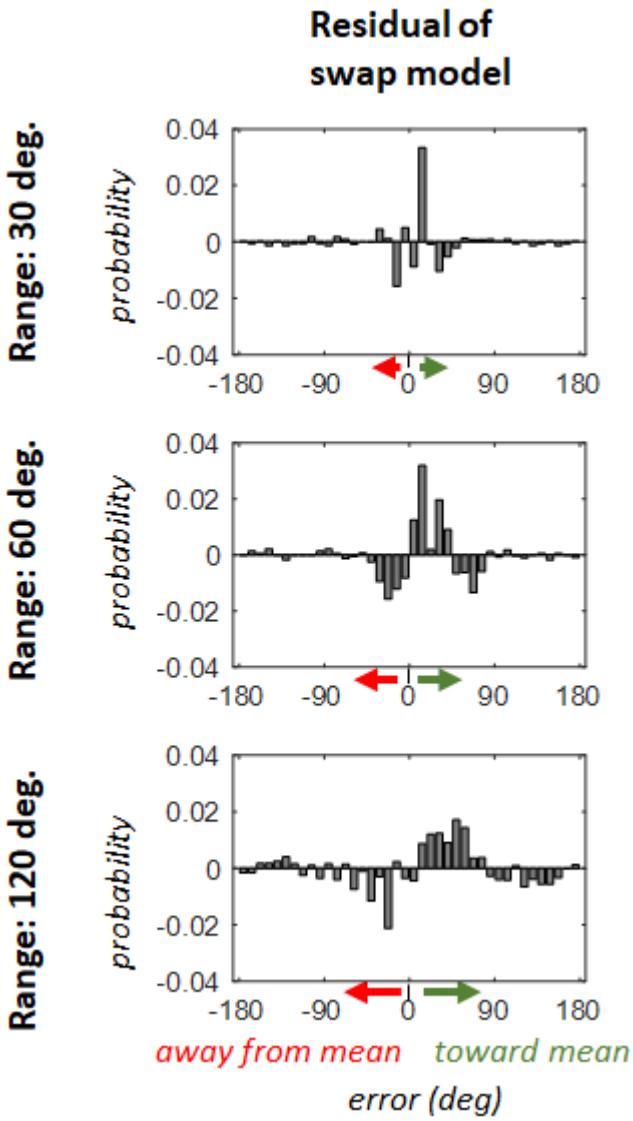

Figure 8. Residual of swap model fits to the data. A model based solely on swaps can account for some aspects of the data, if it allowed to propose different swap rates for displays with different ranges of orientations. However, even such a model has quite systematic residuals: In particular, the data has far more responses relatively near the mean than the swap model predicts, and fewer responses to distractor items away from the mean or far from the mean than is predicted by the swap account.

Importantly, the swap rate estimates we find are also much larger than would be expected from a general swapping account that was not augmented by an ensemble-based strategy. In general, with only 4 items previous studies have found extremely low swap rates (<10%; e.g., Bays et al. 2009; Bays et al. 2011), and in our data, with long encoding times and fixed, substantially distinct positions, we would expect these rates to be even lower. In addition, while the model based solely on swaps can account for some aspects of the data if allowed to propose

different swap rates for displays with different ranges of orientations, this model does still have systematic residuals (Figure 8). In particular, the data has far more responses relatively near the mean than the swap model predicts, and fewer responses to distractor items that happen to be in the direction away from the mean orientation of the display (e.g., left in Fig. 8) or far from the mean (the far right in Fig. 8) than is predicted by the swap account. This seems broadly consistent with the idea that the putative 'swaps' recovered by the swap model are only a rough proxy for how participants use the ensemble properties of the display to limit their responding to items within the general range of items on the display.

Thus, overall, we conclude that in some sense the data here could be thought of as arising from 'swaps': people do respond selectively near the other items orientations. However, this may be an artifact of relying on the display mean and range to limit responding. However, regardless of the cause, this is nevertheless a form of ensemble-based responding, since participants make such responses because they are aware of the feature distribution of the items, rather than as an artifact based on location confusion.

Experiment 2

In Experiment 1, we tested how memory both for individual orientations and for the mean orientation changed with the overall range. This allowed us to directly establish their resemblance: participants were more accurate in item memory when they had a more accurate estimate of the mean. This did not appear to arise from location-based 'swaps' or from solely relying on the mean, but instead seemed to reflect a kind of hierarchical encoding where participants made use of both item information and ensemble information. However, using both a working memory task and an ensemble task in one experiment with the same group of participants (although in separate blocks) could have biased the observers to strategically use ensemble information for remembering individuals more than they would normally use it. That is, the experience of performing the "Remember Mean" task could be transferred to the "Remember Four" task: Relying more on the

mean orientation instead of trying their best to memorize four items. Thus, in Experiment 2, we eliminated the "Remember Mean" task and tested our participants only in the "Remember Four" task with the three ranges of orientations, as in Experiment 1. Moreover, in order to encourage remembering individual objects we added filler trials with range of 360º where no "averageable" ensemble information is available, to further discourage any ensemble-based strategy on the critical fixed range trials.

Method

Participants

Sixteen students of the Higher School of Economics (10 female; age 18-21) took part in the experiment for course credit. None of them took part in Experiment 1. All participants reported having normal or corrected-to-normal vision and no neurological problems. Before the beginning of the experiment, participants gave informed consent.

Apparatus and stimuli

Apparatus and stimuli were the same as those used in Experiment 1. The only addition included displays with orientations spanning the full 360º-range with a step size between items of 90º±3º. Such displays have no orientation ensemble as they have no defined mean orientation.

Procedure

The experiment consisted of a single block with the "Remember Four" task, as described in Experiment 1. Trials with four orientation ranges (30º, 60º, 120º, and 360º) were randomly mixed. There were 80 trials per range resulting in 320 trials in the whole block. It was preceded by 16 practice trials.

Design and analysis

The experiment had a within-subject design with three range conditions (30º, 60º, and 120º). The 360º-range trials were considered to be fillers and were not included in analysis since they provide no information about bias toward the mean (since there was no ensemble mean in the 360º range), and so no analysis comparable with the rest of the range conditions could be applied

to these trials. As in Experiment 1, we estimated memory performance based on both non-parametric statistics and mixture model. A one-way repeated-measures ANOVA was applied to the obtained measures of performance.

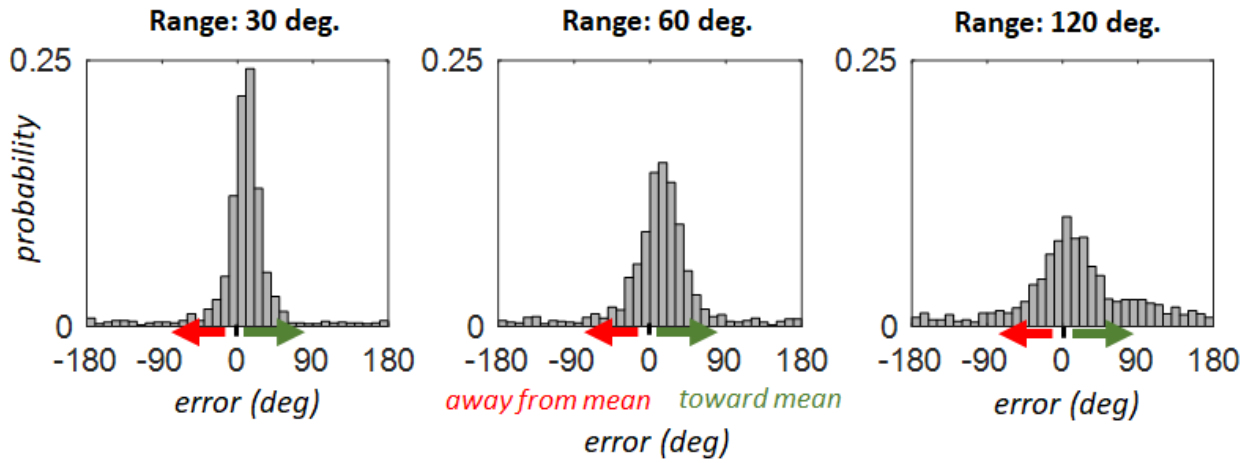

Figure 9. Results of Experiment 2. Histogram of data from each condition pooled across participants, with all errors flipped so that errors toward the mean of the set of items are positive and errors away from the mean of the set of items are negative. Each column represents a different range condition (all 4 items within 30º; within 60º; within 120º).

Results

Precision. The raw error distributions broken down by range are shown in Figure 9. Our non-parametric estimate of error showed strong growth as a function of the range ($F(2,30) = 79.46$, $p < .001$, $\eta^2 = .47$), which was mirrored by the *SD* parameter of the mixture models ($F(2,30) = 18.33$, $p < .001$, $\eta^2 = .47$). This growth was steady, with each range bringing significantly greater *SD* than a previous one (t's(15) = [3.14, 10.53], p's ≤ .007, all values were smaller than Holm corrected α, d_z = [.78, 2.63]; Figure 10A-B). This pattern replicates the pattern found in the "Remember Four" task in Experiment 1.

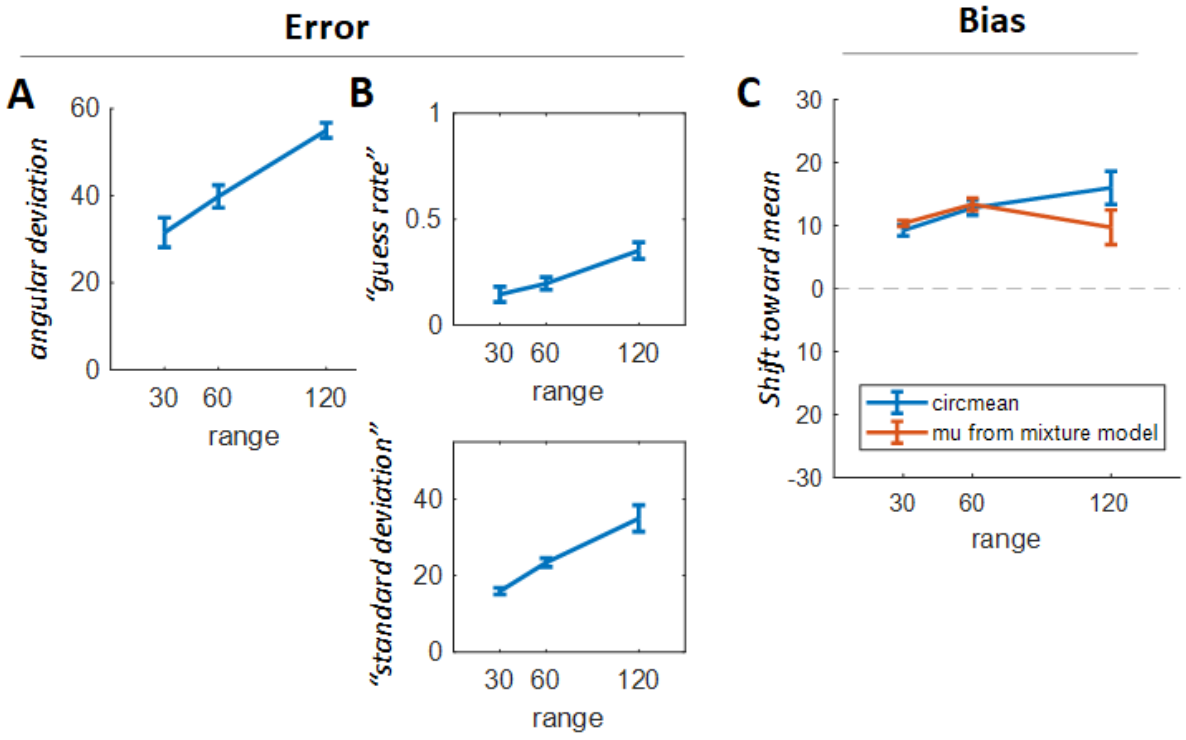

Figure 10. Performance at the memory task in Experiment 2. (A) Non-parametric measures of error show that despite the task being to report memory for a single item, increases in the range of all of the items resulted in large changes in error. (B) Performance at the individual item memory task measured via a mixture model, P_{guess} and standard deviation (SD), a measure of imprecision; (C) Non-parametric measures of bias (the circular mean of the error distribution) and the bias of the von Mises distribution showed a reliable bias for participants to report items as closer to the mean than they really were. Error bars denote ±1 standard error of the mean.

Guess rate. We found the strong effect of the range on P_{guess} extracted from the mixture ($F(2,30) = 22.17$, $p < .001$, $\eta^2 = .19$). Specifically, we found that P_{guess} in the 120º range was greater than in the 30º and 60º ranges (t's(15) = [5.02, 5.77], p's < .001, all values were smaller than Holm corrected α, d_z = [1.26, 1.44]; Figure 10B). This pattern basically repeats the pattern of *SD* changes. Importantly, it also replicates the pattern of P_{guess} changes in the "Remember Four task of Experiment 1, although the absolute guess rates are overall higher in Experiment 2.

Biases. The error distributions were substantially biased towards the mean in all range conditions (Figure 10C), as shown by both non-parametric (M = 9º-16º) and mixture model (M = 9º-14º) bias measures (one-sample comparisons with the null bias: t's(15) = [3.54, 23.13], $p \leq .003$, d_z = [.89, 5.78]). This finding replicates the results of Experiment 1. However, in contrast

with Experiment 1, evidence for a range effect on the bias was inconsistent across the measures: The non-parametric bias measure grew with the range, as in Experiment 1 ($F(2,30) = 4.13$, $p = .026$, $\eta^2 = .15$), whereas the mixture model bias measure showed no evidence for such a growth ($F(2,30) = 1.59$, $p = .22$, $\eta^2 = .05$). The distributions in Figure 9 make clear why this is, in particular why the mixture model 'bias' parameter is so low at range 120º: there are a substantially larger number of responses on the side toward the mean, but they largely occur in the 'tail' of the distribution, not in the central part, so the mixture model discounts them as part of its 'guessing' parameter, which is not allowed to be asymmetric (given the way this model is specified; see Figure 11 for a plot of the model fit to see this). Thus, the mixture model provides a poor fit to this particular distribution and does not capture the shift in responses toward the mean. In general, then, although we replicated the robust absolute bias towards the mean orientation, and the non-parametric bias measure showed this changed with range, this effect may not have been as strong as in Experiment 1.

Overall, the results of Experiment 2 rather closely replicate the results of Experiment 1, "Remember Four" task. To summarize, we found that error distributions became wider as the physical range increased, which can be interpreted as growing imperfection of the retrieved representation (whether coming from the noisy trace or random guesses); we also found the systematic error bias towards the mean. As our participants were performing only the "Remember Four" task in this experiment, we can conclude that the observed pattern was not explicitly informed by their experience of doing an averaging task. Rather, the use of ensemble information in retrieving individual features in working memory appears to be more mandatory.

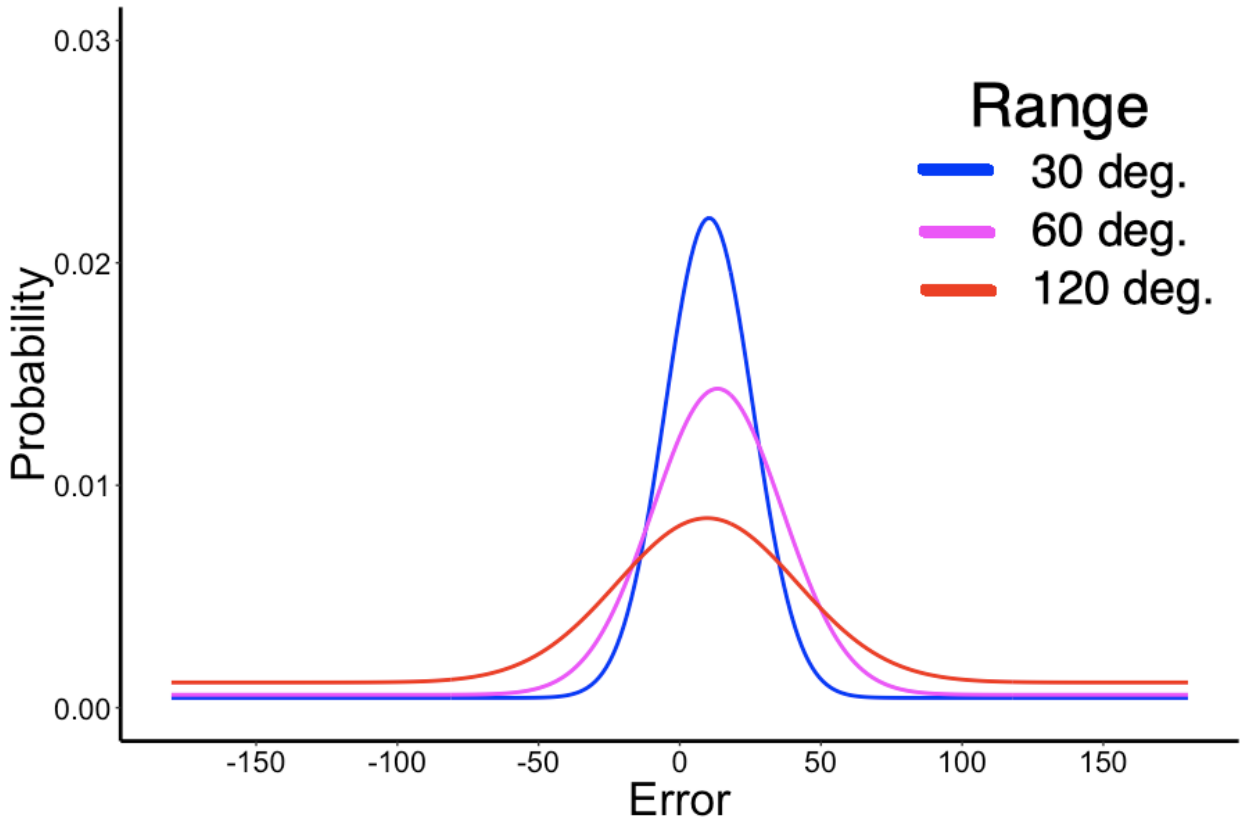

Figure 11. Distribution fits of the mixture models centered on individual tested orientations averaged across participants in Experiment 2. Colors are used for different orientation ranges.

Experiment 3

In a third experiment, we used completely randomly generated displays of sets of 3 orientations to probe the effect of the range of the display when there were no constraints on the displays at all and no suggestion of ensemble coding. Following Brady and Alvarez (2015), we showed each display to 300 participants and asked participants to do whole-report of all 3 orientations from each display. This allowed us to estimate how precisely items were remembered on a display-by-display basis, as a function of the range of the orientations on the display, yet with all items randomly generated as in a typical working memory study.

Method

Participants

300 participants from Amazon's Mechanical Turk participated. All participants reported having normal or corrected-to-normal vision. Before the beginning of the experiment, participants

gave informed consent. 4 participants data was lost or failed to save, leaving a final sample of 296 participants.

Apparatus and stimuli

Participants saw displays of 3 black triangles arranged around an invisible circle, each at a randomly and independently chosen orientation. We used set size 3 rather than set size 4 because randomly generating 4 orientations nearly always results in a range >120º (only 15% of displays have a range <120º), whereas at set size 3, nearly 34% of displays have a range <120º. This allows us a higher powered test of how the range of the display impacts performance.

A set of 48 displays was randomly generated once and these same displays -- with the exact same items in them -- was shown to each participant. This allowed us to look at performance as a function of each individual display.

Procedure

The experiment consisted of a single block of 48 trials. On each trial, participants saw the 3 triangles for 1000 ms, and then had a 1000-ms interstimulus interval (ISI). After this ISI, they were probed on the orientation of each of the 3 triangles in a random order. For each item, they had to adjust the triangle orientation and then click to lock in their answer. Once they locked it in, they were probed on another item until they had reported their remember orientation for all 3 items. This allowed us to estimate not only their accuracy at a single item but also their accuracy with reproducing the entire display. Although each participant saw the same displays, each participant saw the displays in a different randomized order and were probed on items from the display in a random order.

Design and analysis

Our main question was whether the variation in items in the display predicts the accuracy of performance. Thus, we took the range of the items in the randomly generated displays and compared this using a correlation with our non-parametric index of performance -- the angular deviation of responses averaged across all items in the display.

In addition, we compare the bias toward the mean with the range of the display, both as an absolute bias and as a proportion of distance to the mean. This is because in displays with items tightly clustered and also in displays with no ensemble structure, we would predict little absolute bias, but proportionally we expect a large bias in the first case but none in the second case. Thus, using the bias as a proportion of distance to the mean allows us to make a monotonic prediction.

Results

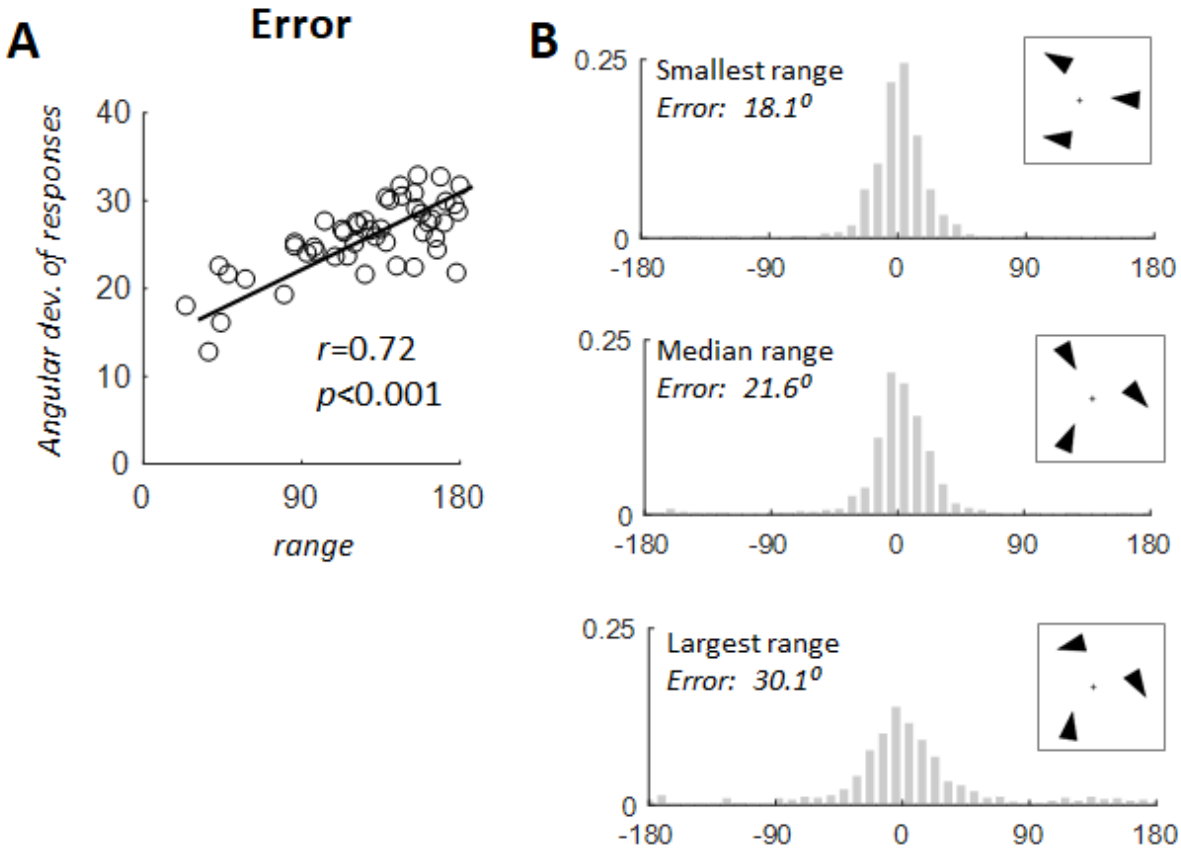

Figure 12. (A) There was a high correlation, in randomly generated displays, between how clustered the items were (in terms of their orientations; e.g., the range) and how accurately participants could reproduce the orientations of the items on the displays. (B) Examples of the error distributions and average error from the smallest range, median range and largest range displays.

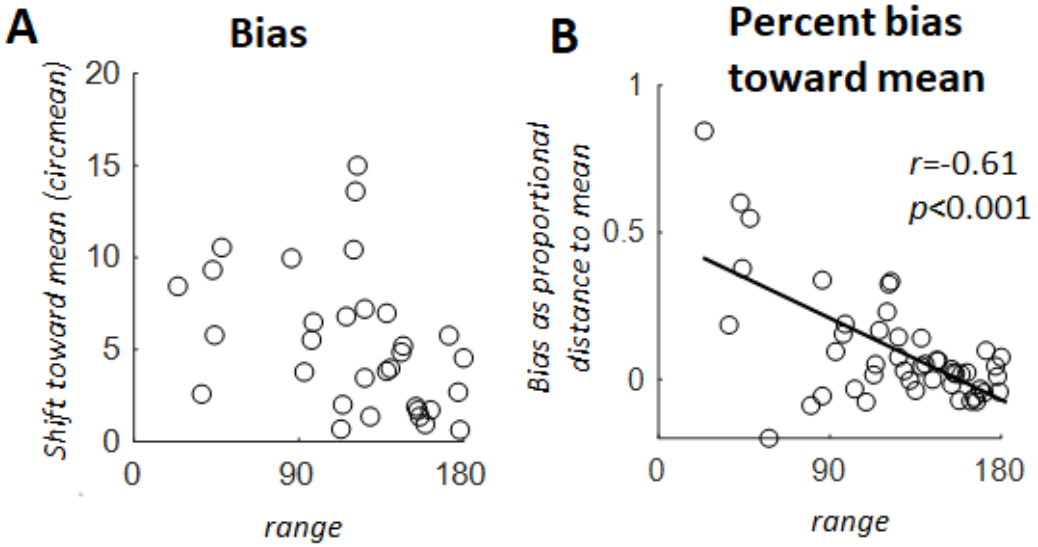

Figure 13. (A) Bias toward the mean as a function of how clustered the items are. The prediction for randomly generated displays is somewhat complicated when considered as raw bias, since we expect a small absolute bias when items are very similar (since they are all close to the mean), and a small absolute bias when items are completely distinct (since there is no real ensemble structure), but a greater absolute bias for intermediate values. (B) In terms of how far the bias takes average reports towards the mean (0 = unbiased; 1 = report the mean with no item influence), however, the predictions are relatively straightforward: Proportionally, participants should be more biased when the items are more tightly clustered. We find this is true, as in the previous experiments.

Precision. We found that even in randomly generated displays, there was a significant relationship between how accurately participants could remember the orientations of the items in the display and how clustered the orientations-to-be-remembered were. Rather than analyze data per participant, we instead analyze the data per display: thus, we collapsed across all ~300 participants and asked how accurately participants could remember each display (see Brady & Alvarez, 2015; Brady & Tenenbaum, 2013 for similar logic). Consistent with our main claim, we found that displays where the orientations were more similar resulted in better performance (r = 0.72, $p < 0.001$; Figure 12).

Bias. What bias towards the mean would we expect in randomly generated displays? The prediction of our ensemble-based account is somewhat complicated when considered as raw bias, since we expect a small absolute bias when items are very similar (since they are all close to the mean), and a small absolute bias when items are completely distinct (since there is no ensemble structure), but a greater absolute bias for intermediate values. In previous experiments, we never tested displays that have no ensemble structure at all, and so found that the bias increased as a

function of the range of the displays (Experiment 1). However, proportionally this bias decreased substantially with range: In Experiment 1, the responses are biased ~9/10 toward the mean in the 30º range, ~2/3 toward the mean in the 60º range, and ~1/2 toward the mean in the 120º range. The same was found in Experiment 2: the responses are biased ~1 toward the mean in the 30º range, ~2/3 toward the mean in the 60º range, and ~1/4 toward the mean in the 120º range. Our account predicts this proportionally smaller bias would continue with increased range.

Thus, our main analysis of bias in this experiment considers how large the bias is as a proportion of distance to the mean, where 0 is completely unbiased and 1 is what we would expect if people only reported the mean, with no influence of the actual shown item. In this case, the predictions are relatively straightforward: Proportionally, participants should be more biased when the items are more tightly clustered. We find this is true, as in the previous experiments, even for randomly generated displays ($r = -0.61$, $p < 0.001$; Figure 13).

Thus, together with Experiment 2, these results demonstrate that the effects we observe under carefully controlled display conditions in Experiment 1 are generalizable to normal visual working memory experiments where no attention is drawn to ensemble properties of the display.

General Discussion

We tested how information about the set of objects stored in visual working memory influences what people remember about individual objects. We directly compared the memory for individual objects with memory for the ensemble average of the entire set of memorized objects. We replicated findings from previous studies that participants memory for individual objects was biased towards the mean of all of the objects (Brady & Alvarez, 2011; Corbett, 2017; Corbin & Crawford, 2018; Dubé et al. 2014; Griffiths et al. 2018), even with only three or four items needing to be remembered. In addition, we found that the bias is not the only parameter that depends on the feature distribution of the whole set of items. Instead, inconsistent with models that suggest items are stored independently, we found that the accuracy of memory – quantified either in terms

of angular deviation or with mixture models – also strongly depend on the statistical structure of the whole set.

In Experiment 1, we used "Remember One" and "Remember Mean" tasks as baseline conditions to assess working memory for either the individual-level alone or the ensemble-level alone. We found that remembering the orientation of a single precued item was not affected by other items that had been present but required no memorization. The reports were always precise (comparable or even slightly better than in other studies using continuous report for orientation, e.g., Bays, Wu, & Husain, 2011; Fougnie & Alvarez, 2012; Fougnie et al., 2010; Zhang & Luck, 2009) and unbiased regardless of the other object orientations (that is, unaffected by ensemble properties). In the "Remember Mean" task, the critical finding was imprecision (the non-parametric deviation or the mixture-model *SD*) growing with the physical range, which was previously documented in averaging tasks in other sensory domains (e.g., Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Dakin, 2001; Fouriezos, Rubenfeld, & Capstick, 2008; Im & Halberda, 2013; Marchant, Simons, & De Fockert, 2013; Maule & Franklin, 2015; Solomon, Morgan, & Chubb, 2011; Sweeny, Haroz, & Whitney, 2012; Utochkin & Tiurina, 2014). Most interestingly, this range-related imprecision turned out to be reflected by reports in our critical memory condition, "Remember Four", where a greater range of items overall led to greater error for individual items despite the fact that participants were being tested only a single item. Taken together, these findings suggest an overall degradation of information about individual objects that has to do with the quality of ensemble representation.

In Experiment 2 and 3, we showed that less accurate memory when the items are more dispersed in feature space occurs even when participants are never probed on the average orientation of the display, and even in completely randomly generated displays. This suggests that even in standard working memory situations, items are not represented independently but the accuracy of memory for an item depends not only on its own feature value but also the distribution of all feature values in a display.

The combination of biases and changes in memory strength can give us useful insights about how observers might utilize individual and ensemble information during encoding and retrieval. We suggest that the relative contribution of an individual or an ensemble component to visual working memory strongly depends on the quality of the latter component. The smallest orientation range (30º in our experiment) yields the most precise representation of the mean; at the same time, the distribution of individual responses in the remember 4 condition at this range is proportionally extremely biased, such that the distribution is shifted nearly as far as would be predicted from responses based on the mean alone. This suggests that the very strong and reliable average representation has a strong influence on memory for an individual item (Alvarez, 2011) whose representation can be rather noisy and ambiguous when competing with other individual representations (Bays, 2014, 2015; Bays et al., 2009; Wilken & Ma, 2004).

When the range increases, the precision of the average representation decreases, making the mean orientation less reliably estimated and less precise as a summary of the items, but still an influential aspect of memory. One consequence of it is that while the biases numerically increased with greater range, the gap between the reported individual orientation and the mean orientation became proportionally greater. For example, if we put the distance between the correct answer and the mean as 1, then in Experiment 1, the responses are biased ~9/10 toward the mean in the 30º range, ~2/3 toward the mean in the 60º range, and ~1/2 toward the mean in the 120º range (Figure 7). A similar picture was found in Experiment 2 where errors in large ranges were even less biased towards the mean than in Experiment 1 (Figure 11), and Experiment 3 showed this proportional decrease in bias toward the mean held across a very wide set of orientation ranges. Therefore, despite the observation that the increased orientation range creates stronger biases in individual representations along the absolute scale, it is in fact less affected by the mean. The intermediate peak positions suggest that participants rely on a mixture of individual and ensemble information (Brady & Alvarez, 2011; Corbett, 2017). The growing role of individual representations can also explain why overall error for individual items increases with the range. At small ranges, if

observers rely more on the average as an approximate of all items in memory then their effective visual working memory set size tends to one and, hence, can be encoded almost for sure. Conversely, when observers rely on their individual representations more the effective set size tends to increase and some items may not be encoded or retrieved well or may be subject to greater noise.

If the increasing range makes observers rely on individual features more, then why does it also cause the growth of imprecision? We suggest that this can be explained by an interaction between individual memory and ensemble representation. Several models of this are possible. For example, if the same individual item information is combined as in Bayesian cue combination, then with less precise information from the ensemble there will be greater imprecision in responses (e.g., Brady & Alvarez, 2011). Similarly, if the perceived range is used to give a coarse impression of "alignment" around the mean orientation, this impression could affect how broad a deviation of an individual orientation is tolerated within this limit of alignment. As an extreme case, having only ensemble memory, an observer could choose a random orientation around the mean within a reasonable corridor set by the perceived ensemble range and be relatively accurate at the narrow range (30º); but this coarse information would provide much less help at the broader ranges (e.g., 120º). Broadly, then, it appears that in many cases observers have some coarse memory representation from the ensemble information, which they somehow use to constrain their individual item responses (either by Bayesian combination or by restricting their responses to the range defined by the ensemble, or some other ensemble-based strategy). A similar mechanism of Bayesian cue combination between imprecise individual representation and prior feature distribution in a category was previously suggested for category learning (Huttenlocher, Hedges, & Vevea, 2000) and long-term memory (Brady, Schacter, & Alvarez, 2018). Here, we show that an instantaneous impression of an ensemble in a single display can be used as such a prior, affecting the precision with which participants can recall imperfect individuals from working memory.

Might our results reflect so-called 'swap' errors, without any ensemble representation at all? (e.g., Bays et al. 2009). Mistakenly reporting incorrect items would be expected to introduce greater error at larger ranges, consistent with the direction of our results. However, as shown in Simulation 2, the effects we find are much larger than would be expected from a general swapping account that was not augmented by an ensemble-based strategy. In general, with only 4 items previous studies have found extremely low swap rates (<10%; e.g., Bays et al. 2009; Bays et al. 2011), and what swaps are present in such data appear to be largely based on spatial confusions (e.g., Emrich & Ferber, 2012; Oberauer & Lin, 2017). This was an important reason why we designed our experiment to minimize the possibility of location-based swap errors by presenting items in reliable spatial locations that are the same on each trial and are maximally different given the limits of the display (e.g., in different corners). Thus, spatial uncertainty -- and accompanying location-based confusions -- are unlikely to play any role in our results, suggesting that the pure 'swap' rate is likely to be near 0. However, this does not mean people may not be responding selectively near the other items but simply that they are doing so because they are aware of the feature distribution of the items, rather than as an artifact based on location confusion. In particular, , one possible response strategy that participants could use -- an ensemble-based strategy -- would be to simply limit the responses to be within the plausible range of the display, which is similar to a swap-based account but based on an efficient use of hierarchical encoding, rather than an error from location noise. As shown in Simulation 2, if we fit a simple swap model to the data, the so-called "swap rate" needs to depend on how clustered in feature space the items are -- it is not fixed, as would be expected of something like location noise, and even so, this model underestimates the reliance on the mean of the display (overpredicting errors away from the mean and underpredicting responses near the mean). Thus, rather than simply reports of items that happened to be in nearby locations, we find that when the items are more similar, participants tend to cluster their responses near the mean and/or range of items on the display, which can be thought of as a kind of swapping but only if the structure of the display is taken into account.

Overall, our results demonstrate that visual working memory for separate objects is strongly modulated by ensemble properties of the set, suggesting observers use representations stored at different levels of abstraction (i.e., item-based and ensemble-based). This is the essential statement of a framework called elsewhere *hierarchical encoding* (Brady & Alvarez, 2011; Brady et al. 2011). Critically, hierarchical encoding suggests that an item is not stored (or forgotten) as a single record in visual working memory but can be present in several different forms. These forms can be used together or interchangeably to reconstruct the item with an approximation allowed by the quality of the information conveyed by each set of these forms. The adaptive nature of such hierarchical representations is easy to see: A single representation of an individual object is precise, but several of them strongly interfere with each other in visual working memory leading to loss in precision or to forgetting. On the other hand, an ensemble representation is only a rough approximation of the individuals but it is less sensitive to limited capacity issues. That is, considering the ensemble representation can allow people to compensate for the loss of individual information. Our data shows that combining individual and ensemble information is a flexible process which depends on their validity as an estimate of an individual item. Both the hierarchical character of visual working memory representations and the flexibility caused by hierarchical storage should be considered for future theorizing about visual working memory.

Acknowledgements

Experiments 1 and 2 were supported by Russian Science Foundation grant 18-18-00334 to I.S.U. Experiment 3 was supported by NSF CAREER grant BCS-1653457 to T.F.B. We thank Natalia Tiurina, Yuri Markov, and Vladislav Khvostov for their assistance in data collection.

Contributions

I.S.U. made experimental scripts and ran Experiments 1 and 2. T.F.B. made the experimental script and ran Experiment 3. I.S.U. I.S.U. and T.F.B. designed the experiments, analyzed data, and wrote the manuscript.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Science, 15,* 122-131.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science : A Journal of the American Psychological Society / APS*, *15*(2), 106–111.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological science,* 12(2), 157-162.

Baddeley, A. D. & Hitch, G. J. (1974). Working memory. In G. A. Bower (ed). *The Psychology of Learning and Motivation: Advances in Research and Theory,* pp. 47–89. New York: Academic.

Baddeley, A. D. (1986). *Working memory.* Oxford, UK: Clarendon Press.

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

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431–438.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–4. http://doi.org/10.1126/science.1158023

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10): 7, 1–7.

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

Berens, P. (2009). CircStat: a MATLAB toolbox for circular statistics. J Stat Software, 31(10), 1-21.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *10*, 384–392.

Brady, T. F. & Alvarez, G.A. (2015a). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, 15(15):6.

Brady, T. F. & Alvarez, G.A. (2015b). No evidence for a fixed object limit in working memory: Ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 41(3), 921-9.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 4.

Brady, T. F., Schacter, D.L., & Alvarez, G.A. (2018). The adaptive nature of false memories is revealed by gist-based distortion of true memories. *PsyArXiv*.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, 43(4), 393-404.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. Vision research, 45(7), 891-900.

Corbett, J. E. (2017). The whole warps the sum of its parts: Gestalt-defined-group mean size biases memory for individual objects. *Psychological Science, 28(1),* 12–22.

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20*(2), 211–231.

Corbin, J. C., & Crawford, L. E. (2018). Biased by the group: Memory for an emotional expression biases towards the ensemble. *Collabra: Psychology*, 4(1).

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–185.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals.

Journal of Optical Society of America A, 18(5), 1016-1026.

Dubé, C., Zhou, F., Kahana, M. J., & Sekuler, R. (2014). Similarity-based distortion of visual short-term memory is due to perceptual averaging. *Vision research*, 96, 8–16. DOI: https://doi.org/10.1016/j.visres.2013.12.016

Emrich, S. M., & Ferber, S. (2012). Competition increases binding errors in visual working memory. *Journal of Vision, 12*(4): 12.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12), 3. http://doi.org/10.1167/11.12.3

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12), 27. http://doi.org/10.1167/10.12.27

Fougnie, D., Cormiea, S. M., Alvarez, G. A. (2013). Object benefits without object-based representations. *Journal of Experimental Psychology: General, 142,* 621-626.

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception and Psychophysics, 70,* 3, 456–464.

Griffiths, S., Rhodes, G., Jeffery, L., Palermo, R., & Neumann, M. F. (2018). The average facial expression of a crowd influences impressions of individual expressions. *Journal of Experimental Psychology: Human Perception and Performance*, 44, 311–319.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe and L. Robertson (Eds.), *From Perception to Consciousness: Searching with Anne Treisman.* Oxford University Press, 339-349.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception & Psychophysics, 75*(2), 278–286.

Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual-short term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26,* 683–702.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–81.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience,* 17(3), 347–356.

Marchant, A. P., Simons, D. J., & De Fockert, J. W. (2013). Ensemble representations: effects of set size and item heterogeneity on average size perception. *Acta Psychologica, 142*(2), 245–250.

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4), 6.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352.

Morey, C. C., Cong, Y., Zheng, Y., Price, M., & Morey, R. D. (2015). The color-sharing bonus: Roles of perceptual organization and attentive processes in visual working memory. Archives of Scientific Psychology, 3(1), 18.

Nassar, M.R., Helmers, J. & Frank, M.J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. Psychological Review, 125, 486-511.

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological Review, 24*(1), 21-59.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. Psychological Review, 120, 297–328.

Peirce, J. W. (2007). PsychoPy - psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8-1.

Pertzov, Y., Dong, M. Y., Peich, M. C., & Husain, M. (2012). Forgetting what was where: the fragility of object-location binding. *PLoS ONE*, *7*(10).

Raffone, A. & Wolters, G. (2001). A cortical mechanism for binding in visual working memory. *Journal of Cognitive Neuroscience, 13,* 766–785.

Schurgin, M. W., Wixted, J. T., and Brady, T.F. (2018). Psychophysical Scaling Reveals a Unified Theory of Visual Memory Strength. *bioRxiv*.

Solomon, J. A., Morgan, M. & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision, 11*(12): 13, 1-11.

Son, G., Oh, B. I., Kang, M. S., & Chong, S. C. (2019, in press). Similarity-based clusters are representational units of visual working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(10), 9. http://doi.org/10.1167/13.10.9

Suchow, J., Fougnie, D., Brady, T. F., & Alvarez, G.A. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception & Psychophysics*, 76(7), 2071-2079.

Sweeny, T. D., Haroz, S., & Whitney, D. (2012). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 329–337.

Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica, 146*, 7–18.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology. General*, *131*(1), 48–64.

Whitney, D. & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69,* 105-129.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of*

Vision, *4*(12), 11–11.

Zar, J. H. (1999). Biostatistical Analysis. Pearson Education, India.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233-U13.

Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science, 20*(4), 423-428.