American Psychological Association

Journal of Experimental Psychology:
Human Perception and Performance

ISSN: 0096-1523

2020, Vol. 46, No. 5, 458–473
http://dx.doi.org/10.1037/xhp0000727

Individual Representations in Visual Working Memory Inherit Ensemble Properties

Igor S. Utochkin
National Research University Higher School of Economics

Timothy F. Brady
University of California, San Diego

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 toward 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, and after accounting for the possibility of location-based 'swap' errors. 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.

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.

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

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, 2014). 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, Catalao, & Husain, 2009; Bays & Husain, 2008; 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, Peich, & 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 article was published Online First March 19, 2020.

Igor S. Utochkin, Department of Psychology, National Research University Higher School of Economics; Timothy F. Brady, Department of Psychology, University of California, San Diego.

Igor S. Utochkin made experimental scripts and ran Experiments 1 and 2. Timothy F. Brady made the experimental script and ran Experiment 3. Igor S. Utochkin and Timothy F. Brady designed the experiments, analyzed data, and wrote the manuscript.

Experiments 1 and 2 were supported by Russian Science Foundation Grant 18–18-00334 to Igor S. Utochkin. Experiment 3 was supported by NSF CAREER Grant BCS-1653457 to Timothy F. Brady. We thank Natalia Tiurina, Yuri Markov, and Vladislav Khvostov for their assistance in data collection.

Correspondence concerning this article should be addressed to Igor S. Utochkin, Department of Psychology, National Research University Higher School of Economics, Armyansky pereulok 4, Room 419, Moscow 101000, Russian Federation. E-mail: isutochkin@inbox.ru

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 abstraction 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, Asplund, & Marois, 2010; Fougnie, Cormiea, & Alvarez, 2013). 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; Morey, Cong, Zheng, Price, & Morey, 2015; Nassar, Helmers, & Frank, 2018; Orhan & Jacobs, 2013; Son, Oh, Kang, & Chong, 2019). These higher-order structures can be compared with a long-known concept of "chunks" (Miller, 1994). 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 the 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 toward the mean size of all dots with the same color and 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 toward 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 a set of experiments using continuous recall of orientation (e.g., Bays et al., 2011; Fougnie & Alvarez, 2011; Fougnie et al., 2010, 2013; Zhang & Luck, 2008). In the first two experiments, 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 nonparametric 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. In a third experiment, we address how this relates to performance in standard working memory tasks where the variability of the set is uncontrolled but varies randomly from trial to trial.

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 women; age range = 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 .80 (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 1,024 × 728 resolution) on a homogeneous gray field. Participants sat approximately 47 cm from the monitor. From that distance, each pixel subtended 0.054° of visual angle.

Memory displays consisted of four white isosceles triangles with different apex orientations. The bases and altitudes of the triangles were 2.7° and 3.9°, respectively. The triangles were centered on an imaginary circle with a radius of 10.8° and occupied cardinal positions corresponding to 45°, 135°, 225°, and 315° of rotation on that circle; random positional jitter between −10° and 10° 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 1° and 360° 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 30°, 60°, or 120° (always centered at this mean orientation). The values were equally spaced in this range and jittered by ±3°. Therefore, for the 30° range, the individual values were −15°, −5°, 5°, and 15° away from the mean (±jitter); for the 60° range, the individual values were −30°, −10°, 10°, and 30° away from the mean (±jitter); for the 120° range, the individual values were −60°, −20°, 20°, and 60° away from the mean (±jitter). Therefore, the mean orientation was never physically present as a member of the set. On average, individual items were 10°, 20°, and 40° from the mean orientation in the three 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 1, Panel A. Each of the individual orientations was equally likely to be probed at test in the tasks demanding memory for individual triangles (see task descriptions in the following text).

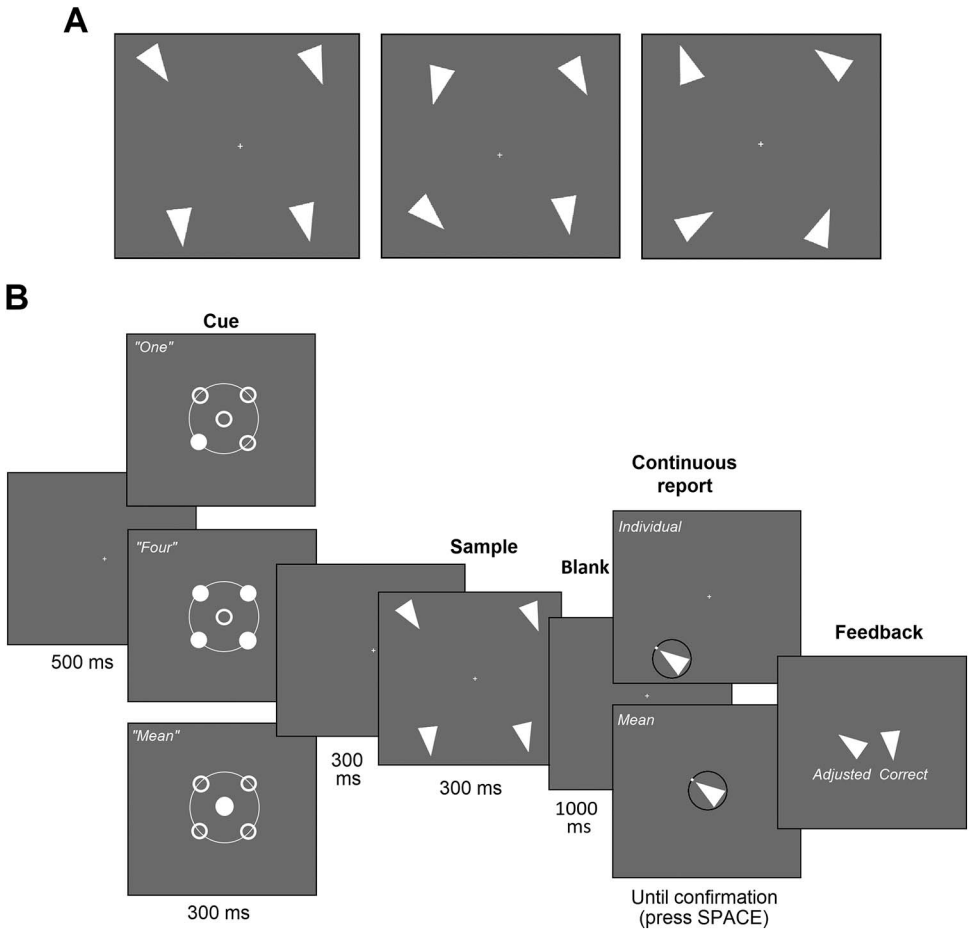

Figure 1. Methods of the experiment. (Panel A) Example stimuli of different ranges (left to right: ranges of 30°, 60°, and 120°). (Panel B) The time course of a typical trial. Participants first saw a cue indicating whether they needed to remember one item, four items, or the mean orientation, followed by a brief delay and then the study display. After a 1,000-ms 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 (i.e., the remember one task), (2) the orientations of all four triangles (i.e., the remember four task), or (3) the mean orientation of all four triangles (i.e., 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., Blocks 1 and 6, or 2 and 5, or 3 and 4). The order of tasks within the first set of three 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 (see 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 9.2° × 9.2°) 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 (see 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: 30° vs. 60° vs. 120°) 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 360° range and thus fell between −180° and 180°. 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 on 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 toward 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 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 nonunitized error distributions in all ranges of the remember one task (M differences <0.6°, ts < 2.3, ps > .03, Bonferroni corrected $\alpha = .017$, $d_z < .6$) and were slightly though consistently smaller in all ranges of measure of the remember four task (M differences = 2–3°, ts > 4.9, ps < .001, Bonferroni corrected $\alpha = .017$, $d_z > 1.8$). Importantly, nonunitized and unitized angular deviations were very highly correlated in both tasks and all ranges (rs > .95, ps < .001). This suggests that error unitization we used to be capable of detecting a bias toward or away from the mean did not strongly distort the rest of the information necessary to judge other critical distributional parameters. Indeed, as expected from this strong correlation, in this experiment as well as Experiments 2 and 3, the analyses of precision and how it is affected by item variability result in the same conclusions using unitized or nonunitized error distributions.

Memory performance from error distributions. We applied two methods of evaluating visual working memory performance from the error distribution. The first method is nonparametric 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 nonparametric, it is closely related to model-based measures of performance like d' (Schurgin, Wixted, & Brady, 2018). In addition to these nonparametric 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 argued 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-measures analysis of variance (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.

Precision. We found strong effects of the task on our nonparametric estimate of error, $F(2, 30) = 63.91$, $p < .001$, $\eta^2 = .81$, and on the standard deviation 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 (nonparametric: $t[47] = 3.51$, $p < .001$, $d_z = .58$; 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 (nonparametric: $t[47] = 11.62$, $p < .001$, $d_z = 1.67$; mixture model: $t[47] = 7.41$, $p < .001$, $d_z = 1.07$). The effect of the range was also strong (nonparametric: $F[2, 30] = 178.98$, $p < .001$, $\eta^2 = .92$; mixture model: $F[2, 30] = 160.60$, $p < .001$, $\eta^2 = .92$), reflecting the overall standard deviation growth with range. In fact, this growth was task-specific (Task × Range effect for Nonparametric Error: $F[4, 60] = 45.05$, $p < .001$, $\eta^2 = .37$; for the mixture model standard deviation: $F[4, 60] = 68.36$, $p < .001$, $\eta^2 = .82$). The nonparametric error or standard deviation did not differ between ranges in the remember one task ($ts[15] = [0.89, 1.78]$, $ps \geq .096$, $d_z = [.22, .45]$) but grew steadily with the range in the remember four and the remember mean tasks ($ts[15] = [7.81, 24.29]$, $ps < .001$, $d_z = [1.95, 6.07]$); though for the nonparametric error, there was no substantial growth between the 30° and the 60° ranges: $t(15) = 1.72$, $p = .11$, $d_z = .43$. Overall, both nonparametric and mixture-model estimates of the error showed basically same patterns. The strong similarity between the nonparametric 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 = 11–12°); 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, whereas the nonparametric 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: $F[2, 30] =$

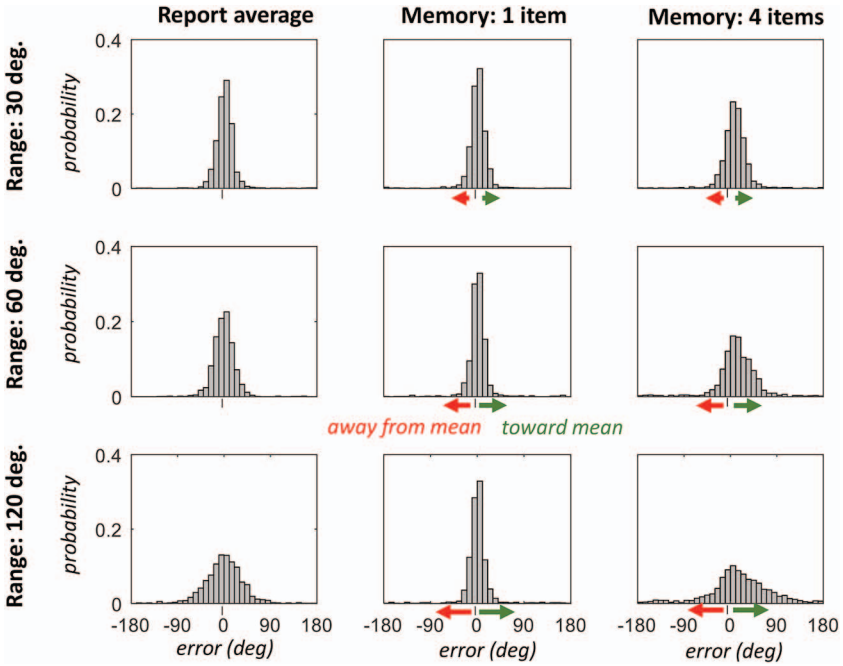

Figure 2. Results of Experiment 1. Histogram of data from each condition pooled across participants, with all errors in the two individual-item memory conditions flipped so that errors toward the mean of the set of items are positive, and errors away from the mean of the set of items are negative. Each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

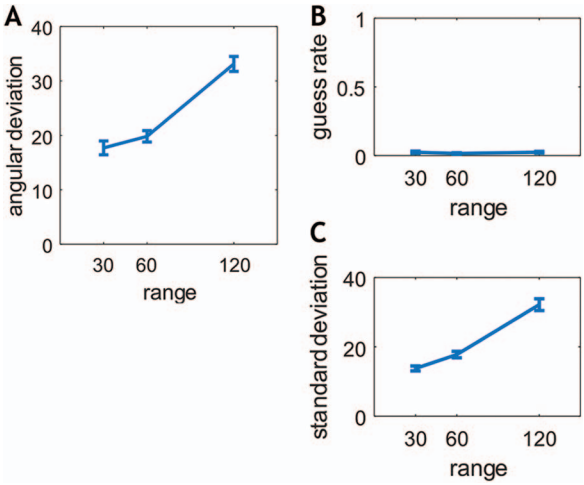

Figure 3. Performance at the ensemble (report the average orientation) task. (Panel A) Performance assessed using a nonparametric measure of error, the angular deviation, shows that performance is much better when the range is small than when the range is large. (Panel B) Similar results are obtained using the mixture model, which shows very few lapses (P_{guess}) and (Panel C) represents a similar effect to the nonparametric 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. See the online article for the color version of this figure.

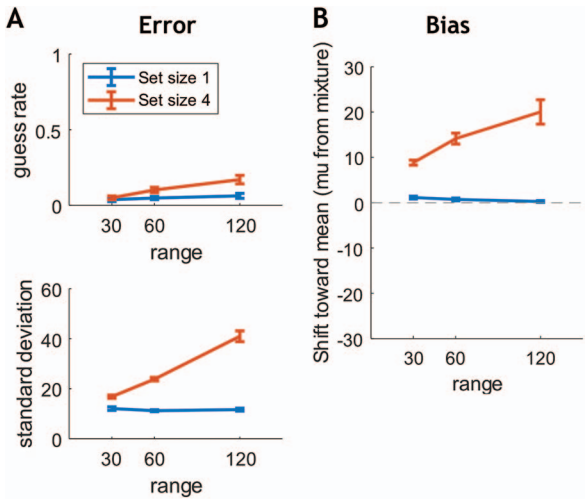

Figure 5. Performance at the individual item memory task measured via a mixture model. (Panel A) P_{guess} and standard deviation (*SD*), a measure of imprecision, at set size one and four. (Panel B) Bias of the von Mises distribution component at set sizes one and four. Error bars denote ±1 standard error of the mean. See the online article for the color version of this figure.

19.36, $p < .001$, $\eta^2 = .57$; range: $F[2, 30] = 15.95$, $p < .001$, $\eta^2 = .52$), though this was largely driven by the large 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 ($ts[15] = [1.35, 2.16]$, $ps = [.05, .20]$, all p values were larger than Holm corrected α, $d_z = [.34, .54]$; see Figure 5A) or in the remember mean task ($ts(15) = [.25, 1.65]$, $ps = [.12, .81]$, $d_z = [.06, .41]$; see 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 ($ts[15] = [2.28, 5.04]$, $ps \leq .005$, all p values were smaller than Holm corrected α, $d_z = [.82, 1.26]$; see Figure 5C) such that with higher orientation variability, not only did imprecision as measured by the mixture model increase but so did P_{guess}.

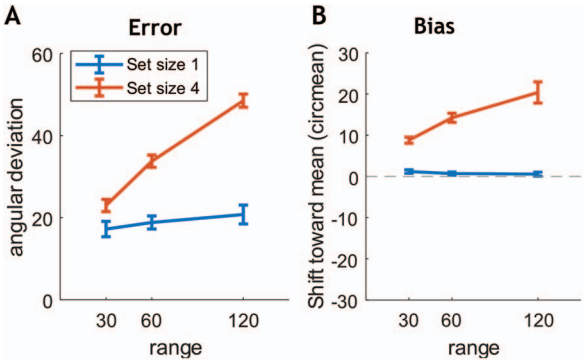

Figure 4. Performance at the individual item memory task measured nonparametrically. (Panel A) Nonparametric measures of error show that despite the task being to report memory for a single item, increases in the range of all of the items resulted in large changes in error, particularly at set size four. (Panel B) Nonparametric 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 four. 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 (+10°, +20°, +40°), it decreased with range (see also simulation results). See the online article for the color version of this figure.

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 the remember four tasks. We found that the task had a strong effect both on the nonparametric bias measure, $F(1, 15) = 136.68$, $p < .001$, $\eta^2 = .90$, and on the parametric one extracted from the mixture model, $F(1, 15) = 133.11$, $p < .001$, $\eta^2 = .90$. This is driven by the remember four task in which the responses were overall substantially biased in the positive (toward the mean) direction ($M = 15°$ for the nonparametric bias; $M = 14°$ 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 ($M = .7°–.8°$, $t[47] = [3.35, 4.02]$, $p < .001$, $d_z = [.47, .58]$). The main effect of the range also was significant in both nonparametric, $F(2, 30) = 12.99$, $p < .001$, $\eta^2 = .46$, and parametric methods, $F(2, 30) = 10.46$, $p < .001$, $\eta^2 = .41$. 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 ($ts[15] = [2.28, 6.07]$, $ps \leq .038$; 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 ($ts[15] = [.18, 1.98]$, $ps \geq .067$, none of the p values are less than Holm corrected α's, Cohen's d_z's = [.04, .50]). The strong simi-

larity between the nonparametric and the mixture model biases are shown 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 four, 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? Although it is unlikely participants used this strategy per se, given the results of Experiment 1, understanding this can help contextualize the extent to which participants used individual information versus relied on information about the mean orientation in the remember four condition. In particular, because the items are similar to each other, particularly at the smallest range, 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. Visualizing what the results of such a strategy would look like can help us think more clearly about the results we did observe.

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. To assess this, we (1) took the average distance between the actually tested orientation and the mean orientation and (2) perturbed this by the error values obtained in the remember mean task, which reflect how accurately participants know the mean. 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 (i.e., simulated from mean only).

We then compared the remember four responses (see Figure 6) with the simulated from mean only responses (see Figure 6). As is shown in Figure 6, 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.

Thus, as expected, comparing the bias predicted from the simulated from mean only responses to the actual remember four biases in the three ranges (2 × 3 repeated-measure ANOVA) reveals that the biases in the actual data are substantially smaller

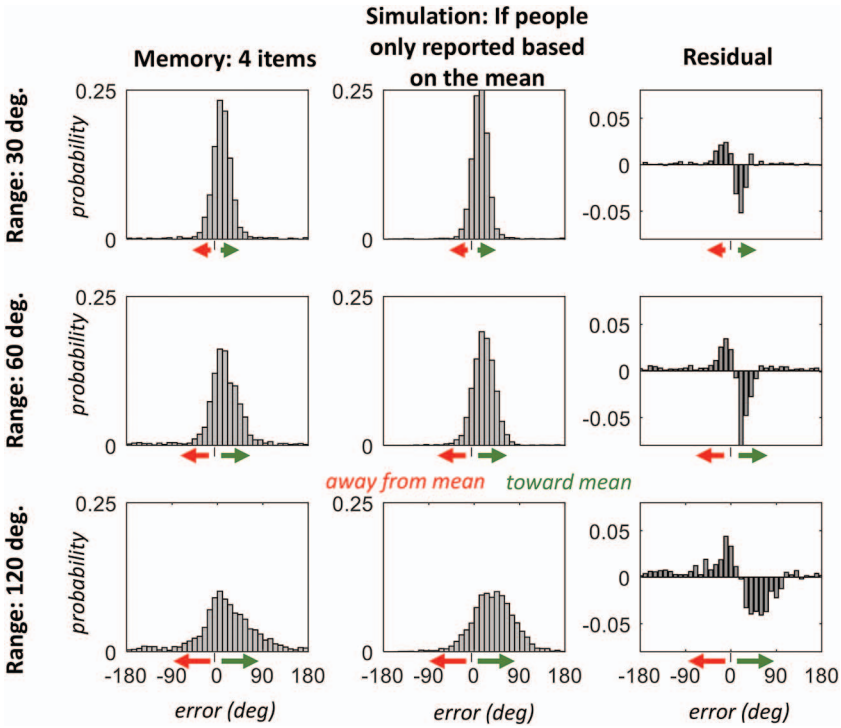

Figure 6. Simulation of data if people reported only based on their (measured) knowledge of the mean. Left panel: Data from the "Remember Four" condition. Middle panel: Data pattern expected if participants solely used their knowledge of the mean (assessed via the "Remember Mean" condition) to respond in the "Remember Four" condition. Right panel: Residual of this model (e.g., Column 1 minus Column 2). Clearly, 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. See the online article for the color version of this figure.

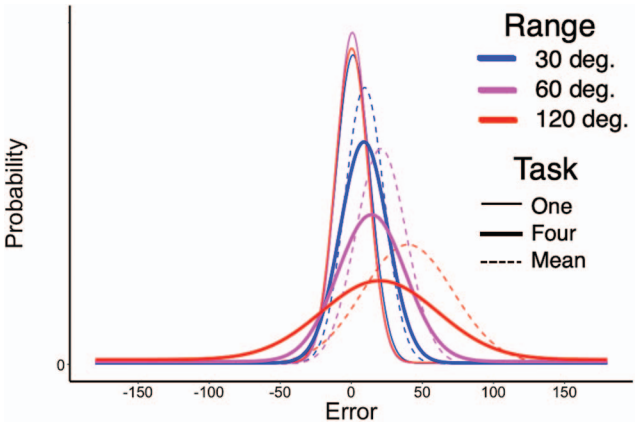

Figure 7. Distribution fits of the mixture models centered on individual tested orientations averaged across participants. The "Remember Mean" task reflects the fits from the "Simulated-From-Mean-Only" (e.g., what errors would be found for the tasks if participants solely used their knowledge of the 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. See the online article for the color version of this figure.

than the simulated from mean only prediction, $F(1, 15) = 37.81$, $p < .001$, $\eta^2 = .716$. 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.

Specifically, in the 30° range the biases were similar between these two tasks ($M = 10°$ in the simulated from mean only vs. $M = 9°$ in the remember four comparison: $t(15) = 1.82$, $p = .089$, $d_z = .45$), but the difference substantially increased in the 60° range ($M = 21°$ vs. 14°, respectively; comparison: $t[15] = 4.07$, $p < .001$, $d_z = 1.02$) and especially in the 120° range ($M = 38°$ vs. 20°, respectively; comparison: $t[15] = 5.78$, $p < .001$, $d_z = 1.44$).

The intermediate value of these biases—not 0, but not all the way to the mean—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 60° and the 120° 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).

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 have simply 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 (i.e., 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 30°, 60°, and 120°, 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, $F(2, 30) = 17.2$, $p < .0001$, and each larger range has a significantly lower swap rate than the tighter range (e.g., 30° vs. 60°: $t[15] = 3.28$, $p = .005$, $d_z = 0.82$; 60° vs. 120°: $t[15] = 3.27$, $p = .005$, $d_z = 0.82$). Furthermore, the precision estimates, even after partialing out such a huge number of putative swaps, are still less precise at larger ranges even in this mixture model: ($M = 12.4°$, 14.4°, 26.0°; $F[2, 30] = 18.7$, $p < .0001$). 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 observers are to rely on other items in their reports.

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 four items previous studies have found extremely low swap rates (<10%; e.g., Bays et al., 2009, 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 (see 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 (see the left in Figure 8) or far from the mean (see the far right in Figure 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, because participants make such responses because they are aware of the feature distribution of the items, rather than as an artifact based on location confusion.

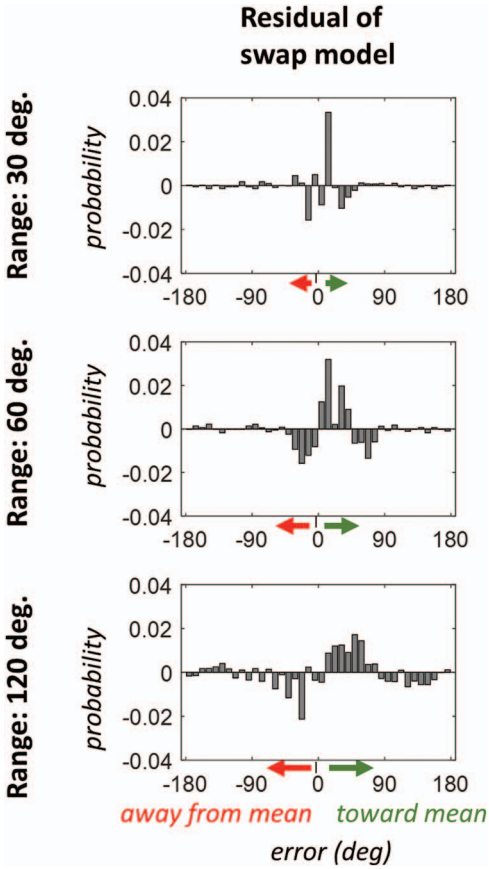

Figure 8. Residual of swap model fits to the data. A model based solely on swaps can account for some aspects of the data, if it allowed to propose different swap rates for displays with different ranges of orientations. However, even such a model has quite systematic residuals: In particular, the data 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. See the online article for the color version of this figure.

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, that is 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 360° 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 women; age range = 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 360°-range with a step size between items of 90° ± 3°. 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 (30°, 60°, 120°, and 360°) 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 (30°, 60°, and 120°). The 360°-range trials were considered to be fillers and were not included in analysis because they provide no information about bias toward the mean (because there was no ensemble mean in the 360° 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 nonparametric statistics and mixture model. A one-way repeated-measures ANOVA was applied to the obtained measures of performance.

Results

Precision. The raw error distributions broken down by range are shown in Figure 9. Our nonparametric 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 standard deviation 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 standard deviation than a previous one ($ts[15] = [3.14, 10.53]$, $ps \leq .007$, all values were smaller than Holm corrected α, $d_z = [.78, 2.63]$; see Figure 10). This pattern replicates the pattern found in the remember four task in Experiment 1.

Guess rate. We found the strong effect of the range on P_{guess} extracted from the mixture model, $F(2, 30) = 22.17$, $p < .001$, $\eta^2 = .19$. Specifically, we found that P_{guess} in the 120° range was greater than in the 30° and 60° ranges ($ts[15] = [5.02, 5.77]$, $ps < .001$, all values were smaller than Holm corrected α, $d_z = [1.26, 1.44]$; see 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 toward the mean in all range conditions (see Figure 10C), as shown by both nonparametric (M = 9°–16°) and mixture model (M = 9°–14°) bias measures (one-sample comparisons with the null bias: $ts[15] = [3.54, 23.13]$, $p \leq .003$, $d_z = [.89, 5.78]$). 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 nonparametric 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 120°: 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 toward the mean orientation, and the nonparametric 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 toward 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.

Experiment 3

In a third experiment, we used completely randomly generated displays of sets of three 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 (2015a), we showed each display to 300 participants and asked participants to do whole-report of all three 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. Participants were 300 people recruited from Amazon's Mechanical Turk. All participants reported having normal or corrected-to-normal vision. Before the beginning of the experiment, participants gave informed consent. Four participants' data were lost, or failed to save, leaving a final sample of 296 participants.

Apparatus and stimuli. Participants saw displays of three black triangles arranged around an invisible circle, each at a randomly and independently chosen orientation. We used set size

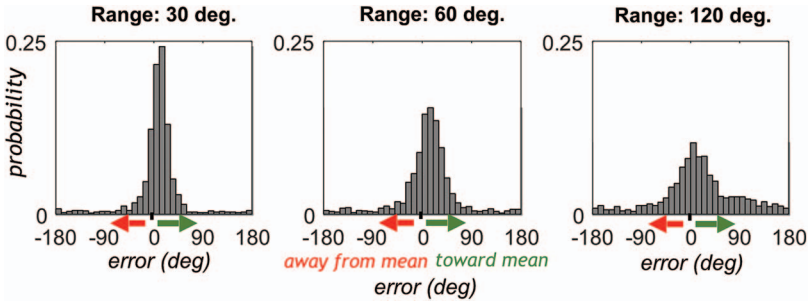

Figure 9. Results of Experiment 2. Histogram of data from each condition pooled across participants, with all errors flipped so that errors toward the mean of the set of items are positive and errors away from the mean of the set of items are negative. Each column represents a different range condition (all four items within 30°, within 60°, or within 120°). See the online article for the color version of this figure.

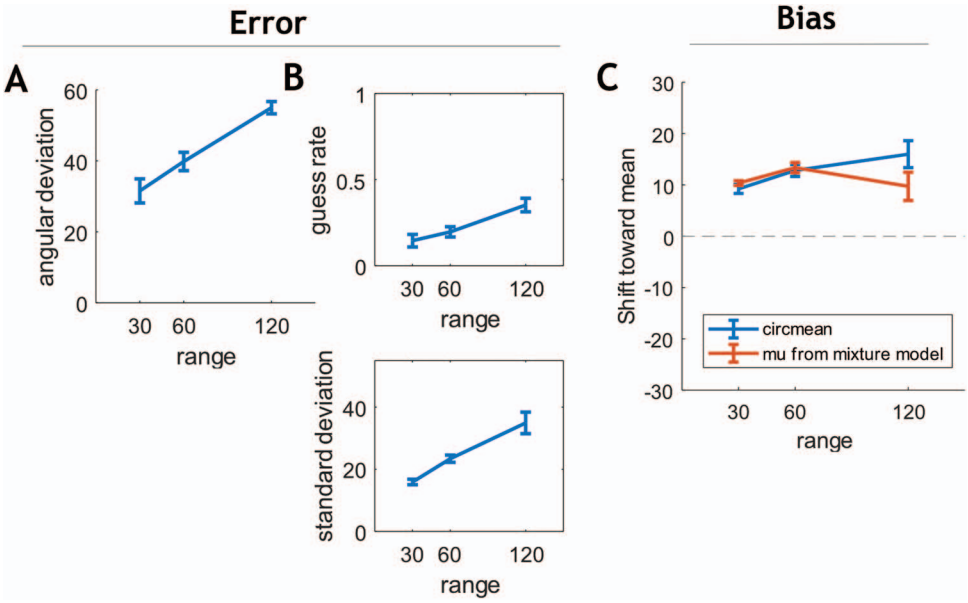

Figure 10. Performance at the memory task in Experiment 2. (Panel A) Nonparametric measures of error show that despite the task being to report memory for a single item, increases in the range of all of the items resulted in large changes in error. (Panel B) Performance at the individual item memory task measured via a mixture model, P_{guess} and standard deviation (*SD*), a measure of imprecision. (Panel C) Nonparametric 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. See the online article for the color version of this figure.

three rather than set size four because randomly generating four orientations nearly always results in a range >120° (only 15% of displays have a range <120°), whereas at set size three, nearly 34% of displays have a range <120°. 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.

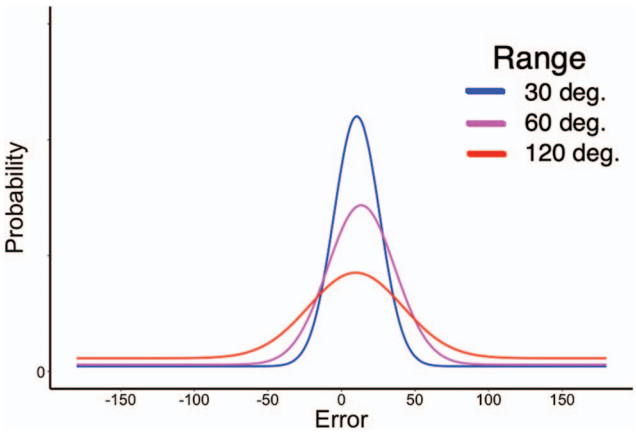

Figure 11. Distribution fits of the mixture models centered on individual tested orientations averaged across participants in Experiment 2. Colors are used for different orientation ranges. See the online article for the color version of this figure.

Procedure. The experiment consisted of a single block of 48 trials. On each trial, participants saw the three triangles for 1,000 ms, and then had a 1,000-ms interstimulus interval (ISI). After this ISI, they were probed on the orientation of each of the three 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 three 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 nonparametric 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

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, 2015a; 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 ($r = .72$, $p < .001$; Figure 12).

Note that in this task, participants reported all items. Thus, it is possible they could have relied more on the clustering in the display for second and third responses, where they had less individual item information. However, we find no evidence for this: examining only the first response each participant gave revealed a similar relationship between orientation similarity and performance ($r = .68$, $p < .001$).

Bias. What bias toward the mean would we expect in randomly generated displays? In Experiments 1 and 2, we considered only displays that have a coherent average orientation. However, at some range of variation between items, the items orientations must become so inconsistent that there is no bias possible: for some displays, the mean itself even becomes undefined if the items all point in contradictory directions. Thus, the prediction of an ensemble-based account is somewhat complicated when considered as raw bias, because we expect a small absolute bias when items are very similar (because they are all close to the mean), and a small absolute bias when items are completely distinct (because there is no ensemble structure to be biased toward), but a greater absolute bias for intermediate values. In previous experiments, we never tested displays that have no ensemble structure at all, and so simply found that the absolute 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 30° range, ~2/3 toward the mean in the 60° range, and ~1/2 toward the mean in the 120° range. The same was found in Experiment 2: The responses are biased ~1 toward the mean in the 30° range, ~2/3 toward the mean in the 60° range, and ~1/4 toward the mean in the 120° 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 zero is completely unbiased and one 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 ($r = -0.61$, $p < .001$; 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

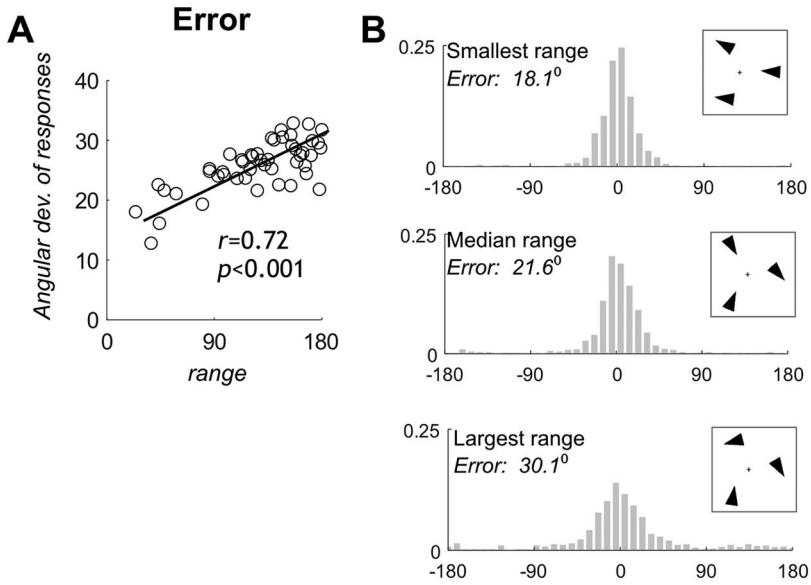

Figure 12. (Panel A) There was a high correlation, in randomly generated displays, between how clustered the items were (in terms of their orientations; e.g., the range) and how accurately participants could reproduce the orientations of the items on the displays. (Panel B) Examples of the error distributions and average error from the smallest range, median range and largest range displays. Note that at the smallest range, the items almost form a single coherent perceptual group, suggesting that the ensemble-based effects we observe at the much more common intermediate levels of variation between stimuli may be continuous with all-or-none perceptual grouping that occurs with identical or near-identical stimuli.

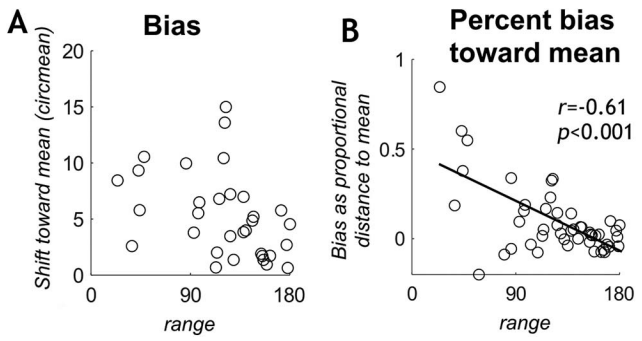

Figure 13. (Panel A) Bias toward the mean as a function of how clustered the items are. The prediction for randomly generated displays is somewhat complicated when considered as raw bias, since we expect a small absolute bias when items are very similar (since they are all close to the mean), and a small absolute bias when items are completely distinct (since there is no real ensemble structure), but a greater absolute bias for intermediate values. (Panel B) In terms of how far the bias takes average reports toward 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.

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 toward 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 et al., 2011; Fougnie & Alvarez, 2011; 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 nonparametric deviation or the mixture-model standard deviation) growing with the physical range, which was previously documented in averaging tasks in various sensory domains (e.g., Corbett, Wurnitsch, Schwartz, & Whitney, 2012; Dakin, 2001; Fouriezos, Rubenfeld, & Capstick, 2008; Im & Halberda, 2013; Marchant et al., 2013; Maule & Franklin, 2015; Solomon, Morgan, & Chubb, 2011; Sweeny, Haroz, & Whitney, 2013; 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 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 Experiments 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 (30° 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 although the biases numerically increased with greater range, the gap between the reported individual orientation and the mean orientation became proportionally greater with increasing range. 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 30° range, ~2/3 toward the mean in the 60° range, and ~1/2 toward the mean in the 120° range (see Figure 7). A similar picture was found in Experiment 2 where errors in large ranges were even less biased toward the mean than in Experiment 1 (see 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 proportionally. The intermediate bias between the correct response and the mean 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 (30°); but this coarse information would provide much less help at the broader ranges (e.g., 120°). 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 how newly learned category information affects the representation of items over short durations (Huttenlocher, Hedges, & Vevea, 2000) and in 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? (See, 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 four items previous studies have found extremely low swap rates (<10%; e.g., Bays et al., 2009, 2011), and the swaps that 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.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15,* 122–131. http://dx.doi.org/10.1016/j.tics.2011.01.003

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, 15,* 106–111. http://dx.doi.org/10.1111/j.0963-7214.2004.01502006.x

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12,* 157–162. http://dx.doi.org/10.1111/1467-9280.00327

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

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, NY: Academic.

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 34,* 3632–3645. http://dx.doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences, 19,* 431–438. http://dx.doi.org/10.1016/j.tics.2015.06.004

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. http://dx.doi.org/10.1167/9.10.7

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321,* 851–854. http://dx.doi.org/10.1126/science.1158023

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia, 49,* 1622–1631. http://dx.doi.org/10.1016/j.neuropsychologia.2010.12.023

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software, 31,* 1–21. http://dx.doi.org/10.18637/jss.v031.i10

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science, 22,* 384–392. http://dx.doi.org/10.1177/0956797610397956

Brady, T. F., & Alvarez, G. A. (2015a). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision, 15*(15), 6. http://dx.doi.org/10.1167/15.15.6

Brady, T. F., & Alvarez, G. A. (2015b). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 921–929. http://dx.doi.org/10.1037/xlm0000075

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. http://dx.doi.org/10.1167/11.5.4

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*. Advance online publication. http://dx.doi.org/10.31234/osf.io/zeg95

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,* 85–109. http://dx.doi.org/10.1037/a0030779

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393–404. http://dx.doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45,* 891–900. http://dx.doi.org/10.1016/j.visres.2004.10.004

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,* 12–22. http://dx.doi.org/10.1177/0956797616671524

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20,* 211–231. http://dx.doi.org/10.1080/13506285.2012.657261

Corbin, J. C., & Crawford, L. E. (2018). Biased by the group: Memory for an emotional expression biases towards the ensemble. *Collabra: Psychology, 4,* 33. http://dx.doi.org/10.1525/collabra.186

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114. http://dx.doi.org/10.1017/S0140525X01003922

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 18,* 1016–1026. http://dx.doi.org/10.1364/JOSAA.18.001016

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. http://dx.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. http://dx.doi.org/10.1167/12.4.12

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. http://dx.doi.org/10.3758/BF03193146

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://dx.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://dx.doi.org/10.1167/10.12.27

Fougnie, D., Cormiea, S. M., & Alvarez, G. A. (2013). Object-based benefits without object-based representations. *Journal of Experimental Psychology: General, 142,* 621–626. http://dx.doi.org/10.1037/a0030300

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics, 70,* 456–464. http://dx.doi.org/10.3758/PP.70.3.456

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. http://dx.doi.org/10.1037/xhp0000446

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), *From perception to consciousness: Searching with Anne Treisman* (pp. 339–349). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:osobl/9780199734337.003.0030

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129,* 220–241. http://dx.doi.org/10.1037/0096-3445.129.2.220

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception & Psychophysics, 75,* 278–286. http://dx.doi.org/10.3758/s13414-012-0399-4

Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 683–702. http://dx.doi.org/10.1037/0278-7393.26.3.683

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281. http://dx.doi.org/10.1038/36846

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences, 17,* 391–400. http://dx.doi.org/10.1016/j.tics.2013.06.006

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17,* 347–356. http://dx.doi.org/10.1038/nn.3655

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,* 245–250. http://dx.doi.org/10.1016/j.actpsy.2012.11.002

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4), 6. http://dx.doi.org/10.1167/15.4.6

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. 1956. *Psychological Review, 101,* 343–352. http://dx.doi.org/10.1037/0033-295X.101.2.343

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,* 18–29. http://dx.doi.org/10.1037/arc0000014

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review, 125,* 486–511. http://dx.doi.org/10.1037/rev0000101

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological Review, 124,* 21–59. http://dx.doi.org/10.1037/rev0000044

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review, 120,* 297–328. http://dx.doi.org/10.1037/a0031541

Peirce, J. W. (2007). PsychoPy - psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8–13. http://dx.doi.org/10.1016/j.jneumeth.2006.11.017

Pertzov, Y., Dong, M. Y., Peich, M. C., & Husain, M. (2012). Forgetting what was where: The fragility of object-location binding. *PLoS ONE, 7,* e48214. http://dx.doi.org/10.1371/journal.pone.0048214

Raffone, A., & Wolters, G. (2001). A cortical mechanism for binding in visual working memory. *Journal of Cognitive Neuroscience, 13,* 766–785. http://dx.doi.org/10.1162/08989290152541430

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2018). Psychophysical scaling reveals a unified theory of visual memory strength. *bioRxiv*. Advance online publication. http://dx.doi.org/10.1101/325472

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision, 11*(12), 13. http://dx.doi.org/10.1167/11.12.13

Son, G., Oh, B. I., Kang, M. S., & Chong, S. C. (2019). Similarity-based clusters are representational units of visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46,* 46–59. http://dx.doi.org/10.1037/xlm0000722

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://dx.doi.org/10.1167/13.10.9

Suchow, J. W., 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,* 2071–2079. http://dx.doi.org/10.3758/s13414-014-0690-7

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 329–337. http://dx.doi.org/10.1037/a0028712

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. http://dx.doi.org/10.1016/j.actpsy.2013.11.012

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General, 131,* 48–64. http://dx.doi.org/10.1037/0096-3445.131.1.48

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble Perception. *Annual Review of Psychology, 69,* 105–129. http://dx.doi.org/10.1146/annurev-psych-010416-044232

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

Zar, J. H. (1999). *Biostatistical analysis*. New Delhi, India: Pearson Education.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453,* 233–235. http://dx.doi.org/10.1038/nature06860

Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science, 20,* 423–428. http://dx.doi.org/10.1111/j.1467-9280.2009.02322.x

Received April 9, 2019
Revision received December 19, 2019
Accepted December 19, 2019 ■