Psychological Review
2013, Vol. 120, No. 2, 297–328

0033-295X/13/$12.00 DOI: 10.1037/a0031541

A Probabilistic Clustering Theory of the Organization of Visual Short-Term Memory

A. Emin Orhan and Robert A. Jacobs
University of Rochester

Experimental evidence suggests that the content of a memory for even a simple display encoded in visual short-term memory (VSTM) can be very complex. VSTM uses organizational processes that make the representation of an item dependent on the feature values of all displayed items as well as on these items' representations. Here, we develop a probabilistic clustering theory (PCT) for modeling the organization of VSTM for simple displays. PCT states that VSTM represents a set of items in terms of a probability distribution over all possible clusterings or partitions of those items. Because PCT considers multiple possible partitions, it can represent an item at multiple granularities or scales simultaneously. Moreover, using standard probabilistic inference, it automatically determines the appropriate partitions for the particular set of items at hand and the probabilities or weights that should be allocated to each partition. A consequence of these properties is that PCT accounts for experimental data that have previously motivated hierarchical models of VSTM, thereby providing an appealing alternative to hierarchical models with prespecified, fixed structures. We explore both an exact implementation of PCT based on Dirichlet process mixture models and approximate implementations based on Bayesian finite mixture models. We show that a previously proposed 2-level hierarchical model can be seen as a special case of PCT with a single cluster. We show how a wide range of previously reported results on the organization of VSTM can be understood in terms of PCT. In particular, we find that, consistent with empirical evidence, PCT predicts biases in estimates of the feature values of individual items and also predicts a novel form of dependence between estimates of the feature values of different items. We qualitatively confirm this last prediction in 3 novel experiments designed to directly measure biases and dependencies in subjects' estimates.

Keywords: visual short-term memory, probabilistic model, perceptual grouping

Supplemental materials: http://dx.doi.org/10.1037/a0031541.supp

Questions about the capacity and precision of visual short-term memory (VSTM) have attracted much attention in recent years (Bays & Husain, 2008; Luck & Vogel, 1997; Rouder et al., 2008; Wilken & Ma, 2004; Zhang & Luck, 2008). Understanding these properties is important due to their theoretical (Cowan, 2001) and practical implications (Fukuda, Awh, & Vogel, 2010). However, there is a more fundamental and often neglected issue that bears directly on memory capacity and precision, namely, the content and organization of VSTM (Brady, Konkle, & Alvarez, 2011; Jiang, Olson, & Chun, 2000; Vidal, Gauchou, Tallon-Baudry, & O'Regan, 2005).

When subjects are presented with a display containing multiple items for a brief period of time, what exactly do they encode in VSTM? What would a complete description of the content of their visual memory for the display include and how is this content organized in VSTM? Do subjects only encode information about individual items or do they also encode more global information about the ensemble of items in the display? Is the information encoded about an item independent of the information encoded about other items? These and other questions about the content and the organization of VSTM are, in a sense, more fundamental than questions about the capacity and precision of VSTM because *how much* information can be encoded in VSTM (capacity) and *how precisely* it can be encoded (precision) depend on exactly *what* information is encoded. For instance, the finding that subjects encode information about ensemble statistics of items in a display (Brady & Alvarez, 2011) could have a significant impact on our estimate of how much information subjects encode about individual items in VSTM.

Here, we introduce a probabilistic modeling approach that attempts to address these questions about the content and the organization of VSTM. Although, as we discuss in the section titled *General Discussion,* our approach has implications for the nature

This article was published Online First January 28, 2013.

A. Emin Orhan and Robert A. Jacobs, Department of Brain & Cognitive Sciences, University of Rochester.

This work was supported by National Science Foundation Grant DRL-0817250 and Air Force Office of Scientific Research Grant FA9550-12-1-0303 awarded to Robert A. Jacobs. We thank C. Sims for many helpful discussions and for commenting on an earlier version of this article. We also thank P. Wilken and W. Ma for sharing their experimental data with us. Parts of this work were presented at the 33rd Annual Meeting of the Cognitive Science Society and at the 25th Annual Conference on Neural Information Processing Systems.

Correspondence concerning this article should be addressed to A. Emin Orhan, Department of Brain & Cognitive Sciences, University of Rochester, CPU 270268, Rochester, NY 14627. E-mail: eorhan@bcs.rochester.edu

of capacity limitations in VSTM, it is intended to be a more general theory of the content and organization of VSTM. We call our approach the probabilistic clustering theory (PCT) of the organization of VSTM. PCT states that VSTM infers probability distributions over partitions or clusterings of visual items. Probabilistic clustering of items gives rise to biases in, and dependencies among, VSTM representations. Representations of items belonging to the same cluster share parameters and thus are dependent. Representations of items belonging to different clusters do not share parameters, and thus are independent. However, VSTM does not infer a single partition. Rather, it infers a probability distribution over all possible partitions. As we discuss below, this property allows it to represent items at multiple granularities or scales.

The article is organized as follows. The next section, titled *Biases and Dependencies in VSTM*, lays out the general framework and reviews experimental evidence for biases and dependencies in VSTM representations. The phenomena reviewed in this section are the type of phenomena our theory is primarily intended to explain. *Hierarchical Encoding of Items in VSTM* discusses previous attempts at explaining some of these phenomena, focusing, in particular, on hierarchical encoding schemes. Although these schemes have many attractive properties, we argue that they also have important shortcomings. *Probabilistic Clustering Theory* introduces PCT. We motivate PCT as a natural generalization of hierarchical encoding approaches in VSTM that addresses the shortcomings of these approaches discussed in *Hierarchical Encoding of Items in VSTM*. We then discuss the relationships between PCT and previous works on hierarchical encoding in human memory. *Models* describes the computational models that will be used in the remainder of the article. As we discuss in this section, these models can all be regarded as specific implementations of PCT with varying degrees of generality. *Simulations* demonstrates that PCT accounts for a variety of phenomena observed in previous visual short-term recall and recognition experiments. *Experiments* presents three new experiments designed to directly measure dependencies and biases in subjects' VSTM representations. These experiments reveal a hitherto unrecognized form of dependence between VSTM representations of different items that is qualitatively predicted by PCT. Finally, the *General Discussion* section provides a summary, discusses connections with related ideas, and suggests avenues for future research.

Biases and Dependencies in VSTM

In this section, we first lay out the general framework and the mathematical notation that we use throughout the article and then review experimental evidence for biases and dependencies in VSTM.

Probabilistic Encoding in VSTM

Consider an observer that briefly views a display containing N visual items. The observer is asked to remember the feature values of these items and, after a brief delay interval, to report one or more of them. We denote the actual feature values of the items by the random variables $\theta_1, \ldots, \theta_N$. We assume that the observer only has access to noisy internal observations of these features, denoted by the variables $x_1, \ldots, x_N$, that are assumed to be corrupted by both sensory and memory noise. The generation of these noisy observations can be described by a likelihood function

$$p\left(\{x_i\}_{i=1}^N \middle| \{\theta_i\}_{i=1}^N\right),$$

which we assume to be a normal distribution in this article. In addition, the observer might have prior assumptions about the feature values of the items. These assumptions can be described by a prior distribution $p(\{\theta_i\}_{i=1}^N)$. Given the likelihood and the prior, the observer's goal is to compute the posterior distribution over the feature values $\theta_1, \ldots, \theta_N$ in accordance with Bayes' rule:

$$p\left(\{\theta_i\}_{i=1}^N \middle| \{x_i\}_{i=1}^N\right) \propto p\left(\{x_i\}_{i=1}^N \middle| \{\theta_i\}_{i=1}^N\right) p\left(\{\theta_i\}_{i=1}^N\right) \quad (1)$$

In a recall task, the observer then makes point estimates of the feature values of the items based on the posterior distribution. We denote the observer's estimates of the feature values by the random variables $\hat{\theta}_1, \ldots, \hat{\theta}_N$. Note that $\hat{\theta}_1, \ldots, \hat{\theta}_N$ are random variables (i.e., they are stochastic) even when conditioned on a specific $\theta_1, \ldots, \theta_N$, because they depend on the noisy observations $x_1, \ldots, x_N$. In this article, we use the posterior mean as the observer's estimate of the feature values of the items in recall tasks, although we found that the results presented here were robust to the choice of a specific estimator so long as the estimator was reasonable. If, for example, the observer is asked to report the feature value of a single target item t, the marginal posterior corresponding to that item, $p(\theta_t|\{x_i\}_{i=1}^N)$, is computed from the joint posterior in Equation 1 and the observer's estimate is taken to be the mean of the marginal posterior: $\hat{\theta}_t = E[\theta_t|\{x_i\}_{i=1}^N]$.

Next consider the joint probability distribution over the estimates given the feature values of the visual items:

$$p\left(\left\{\hat{\theta}_i\right\}_{i=1}^N \middle| \{\theta_i\}_{i=1}^N\right) = \int p\left(\left\{\hat{\theta}_i\right\}_{i=1}^N \middle| \{x_i\}_{i=1}^N\right) p\left(\{x_i\}_{i=1}^N \middle| \{\theta_i\}_{i=1}^N\right) d\{x_i\}_{i=1}^N, \quad (2)$$

where the noisy internal observations $\{x_i\}_{i=1}^N$ are now integrated out. This joint distribution provides a complete characterization of how the observer represents the specific set of items $\theta_1, \ldots, \theta_N$ in his or her VSTM. We note that most of the previous works in the VSTM literature were mainly concerned with elucidating the encoding of individual items and how it changes with set size (e.g., how the encoding precision for individual items decreases with the number of displayed items). In our framework, this corresponds to characterizing only the marginals of the full joint distribution (e.g., the precision of the marginals and how it changes with set size). In contrast, we develop experimental and computational methods to characterize the properties of the full joint distribution (thereby focusing on the joint encoding of all items), instead of emphasizing only the marginals (i.e., focusing on the encoding of individual items).

For a given set of feature values $\theta_1, \ldots, \theta_N$, the joint distribution in Equation 2 can be determined empirically by presenting the same set of feature values over a number of trials and recording the observer's estimates $\hat{\theta}_1, \ldots, \hat{\theta}_N$ for each presentation. A contribution of this article is that we design novel short-term recall and recognition tasks to determine the properties of the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ experimentally. We say more about how to measure this distribution experimentally in the section titled *Experiments* below. For now, our discussion of the experimental results reviewed in the current section will limit the range of

suitable forms for the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ by ruling out some simple proposals.

Figure 1 schematically illustrates biases and dependencies that may arise in the joint estimates of the feature values of multiple items. The phenomena reviewed in the next subsection (*Biases in VSTM*) provide evidence for biases in VSTM and, hence, suggest that $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ should have a form similar to the example shown in Figure 1A. The phenomena reviewed in the following subsection (*Dependencies in VSTM*) provide evidence for dependencies among VSTM representations of multiple items and, hence, suggest that $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ should have a form similar to the example shown in Figure 1B. Together, these biases and dependencies paint a picture of $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ that has a form similar to the example shown in Figure 1C.

Biases in VSTM

A simple suggestion for the form of the joint distribution of the estimates is to assume that feature values of different items are represented independently in VSTM and that estimates of individual items are only affected by the actual feature values of the corresponding items. This corresponds to the assumption that the distribution can be factorized as $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N) = \Pi_{i=1}^N p(\hat{\theta}_i|\theta_i)$. This simple proposal is implicitly assumed in many studies on the capacity and precision of VSTM (Bays & Husain, 2008; Zhang & Luck, 2008). Assuming this factorized form for the joint distribution, one can then model individual distributions $p(\hat{\theta}_i|\theta_i)$ using, for example, univariate Gaussian distributions (Bays & Husain, 2008) or mixtures of Gaussian and uniform distributions (Zhang & Luck, 2008). However, there is extensive evidence against this simple factorized proposal. Here, we briefly review some of the evidence against it (for a more comprehensive review, see Brady et al., 2011).

Kahana and Sekuler (2002) showed that interitem similarity between stimuli influences subjects' performances in an old/new recognition task. In their Experiment 1, subjects were shown a set of study items consisting of a series of sinusoidal gratings with different spatial frequencies. After a blank interval, they were then shown a test grating that, on half of the trials, had the same spatial frequency as one of the study items (old) and, on the other half of the trials, had a novel spatial frequency (new). The task was to decide if the spatial frequency of the test probe was old or new. The authors fit subjects' data using a simple "noisy exemplar" model that included terms for the effects of both the probe-item similarity between the test probe and each of the study items and the interitem similarity among the study items. They found that interitem similarity had a significant effect on subjects' old/new decisions. In particular, when probe-item similarities were fixed, larger interitem similarities increased the likelihood of a "new" response. This result suggests that the estimate of each individual item $\hat{\theta}_i$ depends on the interitem similarity among all study items [contrary to the assumption that memories for individual items depend only on the feature values of their corresponding items; i.e., $p(\hat{\theta}_i|\{\theta_j\}_{j=1}^N) \neq p(\hat{\theta}_i|\theta_i)$]. In a later section (*Simulations*), we show how PCT explains the interitem similarity effect.

Kahana, Sekuler and colleagues replicated the interitem similarity effect in later works (Kahana, Zhou, Geller, & Sekuler, 2007; Viswanathan, Perl, Visscher, Kahana, & Sekuler, 2010; Zhou, Kahana, & Sekuler, 2004) and showed that the same qualitative interitem similarity effect can be observed in visual short-term memory for realistic-looking synthetic face stimuli (Yotsumoto, Kahana, Wilson, & Sekuler, 2007) as well as in auditory short-term memory (Visscher, Kaplan, Kahana, & Sekuler, 2007).

Huang and Sekuler (2010) showed that visual short-term recall memory for the spatial frequency of a sinusoidal grating is

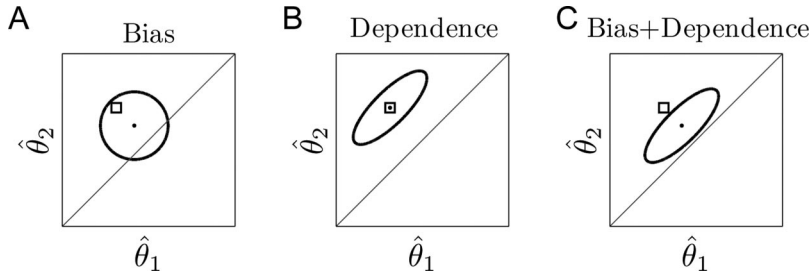

Figure 1. Schematic illustration of biases and dependencies that may arise in the joint estimates of two items based on their visual short-term memory representations. A. Biases manifest themselves as shifts of the distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ (represented here by a single contour) from the actual feature values of the items. In this example, the mean of $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ (represented by the black dot) is shifted away from the actual feature values of the items (θ_1, θ_2) (represented by the square) and toward the main diagonal, indicating that the estimates of the feature values of both items are biased toward the mean of θ_1 and θ_2. Note, however, that the distribution is spherical; hence, assuming a Gaussian distribution for simplicity, there are no dependencies between $\hat{\theta}_1$ and $\hat{\theta}_2$. B. Dependencies manifest themselves as statistical dependencies among $\{\hat{\theta}_i\}_{i=1}^N$. In this example, representations of the two items, $\hat{\theta}_1$ and $\hat{\theta}_2$, are correlated. Note, however, that there are no biases in the representations, as $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ is centered on (θ_1, θ_2). Also note that this example depicts only a simple form of dependence between $\hat{\theta}_1$ and $\hat{\theta}_2$, namely, second-order correlation. More complex or higher order dependencies between $\hat{\theta}_1$ and $\hat{\theta}_2$ are also possible. C. A hypothetical example where there are both biases and dependencies in the joint estimates of the two items.

biased toward both the nontarget gratings shown on the same trial and the average frequency of the gratings shown on all previous trials. On each trial of their Experiment 2, a subject was successively shown a pair of Gabor stimuli. One of the Gabors was then cued, and the subject reported the spatial frequency of the cued Gabor by adjusting the spatial frequency of a comparison Gabor using a computer mouse. The recall error (the difference between the actual spatial frequency of the target Gabor and the subject's reproduction) was measured on each trial. Over trials, this yielded an error distribution that reflected the precision of, and potential biases in, a subject's short-term memory for spatial frequency. It was found that there were two distinct biases influencing subjects' recall of the spatial frequencies of target Gabors: a bias toward the spatial frequency of the nontarget Gabor shown on the same trial, and a bias toward the average spatial frequency of stimuli shown on previous trials in the experiment. A similar bias toward mean spatial frequencies was observed in Experiment 9 of Wilken and Ma (2004). Again, these results indicate that the estimate of a target item t depends on the nontarget items presented on the same trial, as well as on items shown on previous trials, and not solely on the feature value of the target item itself [i.e., $p(\hat{\theta}_t|\{\theta_i\}_{i=1}^N) \neq p(\hat{\theta}_t|\theta_t)$].

Specifically (and assuming, for example, $N = 2$ items), the biases observed in Huang and Sekuler (2010) and in Wilken and Ma (2004) suggest that $p(\hat{\theta}_1|\theta_1,\theta_2)$ is biased (or shifted) toward θ_2 and, conversely, $p(\hat{\theta}_2|\theta_1,\theta_2)$ is biased toward θ_1 (see Figure 1A). In a later section (*Simulations*), we show that our PCT satisfies this property and explains the biases observed in Experiment 9 of Wilken and Ma (2004).

Dependencies in VSTM

The biases in VSTM reviewed in the previous subsection indicate that estimates of individual items depend not just on the actual feature values of their corresponding item but also on the feature values of other items presented in the display. Consequently, VSTM cannot be characterized using a simple joint probability distribution of the form $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N) = \Pi_{i=1}^N p(\hat{\theta}_i|\theta_i)$. However, these biases do not rule out slightly more complex joint probability models of the form $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N) = \Pi_{i=1}^N p(\hat{\theta}_i|\{\theta_j\}_{j=1}^N)$. Here, the estimates of individual items $\hat{\theta}_i$ depend on the feature values of all visual items, but these estimates are independent of each other given the feature values of all items. A series of elegant experiments by Jiang, Olson, and Chun (2000), however, ruled out this latter form of joint probability model as an accurate description of the organization of VSTM.

In each trial of their Experiment 1, Jiang, Olson, and Chun (2000) briefly presented a display consisting of colored squares. Following a blank interval, subjects were shown a test display. There were two test conditions. In the single probe condition, only one of the squares (called the target probe) reappeared, either with the same color as in the original display or with a different color. In the minimal color change condition, the target probe (again with the same color or with a different color) reappeared together with distracter squares, which always had the same colors as in the original display. In both conditions, the task was to decide whether a color change occurred in the target probe. It was found that subjects' performances were significantly better in the minimal color change condition than in the single probe condition. This result suggests that the color for the target square was not encoded independently of the colors of the distracter squares, because, otherwise, the absence or presence of the distracter squares would not have affected change detection performances for the target. In Experiment 2, the authors observed a similar result for location memory. Location memory for a target was better in the minimal change condition than in the single probe condition or in a maximal change condition in which all distracters were presented but at locations differing from their original locations.

These results are easy to understand in terms of a joint probability model for the estimates $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ (in what follows, we omit the dependence on $\{\theta_i\}_{i=1}^N$ for brevity of notation, but all distributions should be considered to be implicitly conditioned on $\{\theta_i\}_{i=1}^N$). Intuitively, the single probe condition taps into the marginal probability distribution of a subject's estimate of the target item $p(\hat{\theta}_t)$ where t indexes the target item, because in the single probe condition distracters are not shown to the subject during test, and thus he or she has to marginalize over his or her uncertainty regarding the feature values of the distracter items. In contrast, the minimal color change condition taps into the conditional probability distribution of the estimate of the target given that the estimates of the distracters are set to the actual feature values of their corresponding items (i.e., $p(\hat{\theta}_t|\hat{\theta}_{-t} = \theta_{-t})$, where $-t$ is the set of indices of the distracter items) because the actual distracters θ_{-t} are shown to the subject during test. If the target probe has high probability under these distributions, then the subject will be more likely to respond "no-change," whereas if it has low probability, then the subject will be more likely to respond "change." Importantly, if the items are represented independently in VSTM, the marginal and conditional distributions are the same [i.e., $p(\hat{\theta}_t) = p(\hat{\theta}_t|\hat{\theta}_{-t})$]. Because subjects' performances in the single probe and minimal color change conditions were different, subjects' marginal and conditional probability distributions must have been different (see Figure 1B for a graphical illustration of a simple form of dependence between estimates of different items). Hence, the results of Jiang, Olson, and Chun (2000) provide evidence against the independence assumption.

It is also easy to understand why subjects performed better in the minimal color change condition than in the single probe condition. The conditional distribution $p(\hat{\theta}_t|\hat{\theta}_{-t})$ is, in general, a lower variance distribution than the marginal distribution $p(\hat{\theta}_t)$. Although this is not exclusively true for the Gaussian distribution, it can be analytically proven in the Gaussian case. If $p(\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_N)$ is modeled as an N-dimensional multivariate Gaussian distribution:

$$[\hat{\theta}_t, \hat{\theta}_{-t}]^T \sim \mathcal{N}\left([a, b]^T, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}\right), \tag{3}$$

then the conditional distribution $p(\hat{\theta}_t|\hat{\theta}_{-t})$ has mean

$$a + CB^{-1}\left(\hat{\theta}_{-t} - b\right)$$

and variance $A - CB^{-1}C^T$, whereas the marginal distribution $p(\hat{\theta}_t)$ has mean a and variance A, which is always greater than $A - CB^{-1}C^T$.

Using three novel experiments, *Experiments* provides additional experimental evidence for dependencies in VSTM and discusses how these dependencies are accounted for by our PCT.

Hierarchical Encoding of Items in VSTM

The phenomena reviewed in the previous section restrict the range of suitable joint probability distributions for characterizing VSTM, but they do not completely determine the exact form of this distribution. In this section, we discuss a proposal for the joint distribution recently put forward by Brady and Alvarez (2011). We first review their experimental results and their hierarchical encoding models. The discussion of their work will help us motivate PCT as a natural generalization of hierarchical encoding models, as we believe that PCT addresses some of the shortcomings of their hierarchical modeling approach.

Brady and Alvarez (2011)

It has recently been argued that VSTM is organized hierarchically where items are simultaneously encoded at multiple levels of abstraction (Brady & Alvarez, 2011). At a fine scale, each item might be represented individually. Items might also be represented at a coarser scale through summary or ensemble statistics of the feature values of all items in a display. Brady and Alvarez (2011) formalized this idea using a hierarchical modeling approach. We show below that, in our framework, this corresponds to using a specific hierarchical form for the prior distribution over the feature values, $p(\theta_1, \ldots, \theta_N)$, in Equation 1.

Similar to the interitem similarity effect shown by Kahana and Sekuler (2002), Brady and Alvarez (2011) demonstrated that memory for individual items in a display is influenced by ensemble statistics of all presented items. In their Experiment 1, subjects were presented with blue, red and green circles of different sizes for a brief duration. Subjects were explicitly instructed to ignore the green circles but to remember the sizes of the red and blue circles. After a delay interval, a comparison circle appeared at the location of a red or blue circle in the original display. Subjects' task was to indicate the size of the original circle that was at that location, referred to as the target circle, by using the mouse to resize the comparison. The authors found that the reported size of the target circle was biased toward the average size of the circles having the same color as the target.

Brady and Alvarez (2011) hypothesized that the fact that color was a task-relevant feature in Experiment 1 (subjects had to remember only the red and blue circles and ignore the green circles) might have increased the salience of this feature, thereby inducing subjects to use a color-based encoding for the items. If so, then the observed bias toward the mean size of the same-colored circles should disappear when performing a task that is similar except that color is task-irrelevant. In their Experiment 2, the authors tested this prediction by removing the green circles from the display and presenting only red and blue circles in each trial. Subjects were asked to remember the sizes of all circles in the display. Therefore, color was no longer a task-relevant feature. Consistent with their hypothesis, the authors found that subjects did not show a bias toward the mean size of the same-colored circles. Instead, subjects' estimates showed a bias toward the mean size of all circles in a display. The results of Experiments 1 and 2 suggest that subjects employ a flexible strategy, encoding stimuli at different levels of abstraction in VSTM in different task contexts.

To explain the distinctive pattern of biases when color was a salient feature versus when it was not, Brady and Alvarez (2011) used a two-level hierarchical model to account for subjects' data in Experiment 2 and a three-level hierarchical model to account for data from Experiment 1. The two-level model assumes that subjects encode the items at two different levels of abstraction: the level of individual circles (individual encoding) and the ensemble mean and variance of the feature values (sizes) of all circles in a display. The model also assumes that the feature values of individual items are conditionally independent given the ensemble statistics (though they are still marginally dependent due to the shared ensemble statistics). Using our notation, their two-level model corresponds to choosing the following prior over the feature values of the items in Equation 1:

$$p(\theta_1, \theta_2, \ldots, \theta_N) = \int p(\theta_1, \ldots, \theta_N|\Phi)p(\Phi)d\Phi = \int p(\theta_1|\Phi)\ldots p(\theta_N|\Phi)p(\Phi)d\Phi, \quad (4)$$

where Φ denotes the ensemble statistics for the feature values of all items in a display. Since the color-based grouping of circles is not taken into account in this model, only a bias toward the overall mean size is predicted when estimating the sizes of individual circles (Gelman, Carlin, Stern, & Rubin, 2004, p. 117), in accord with the results of Experiment 2.

In the three-level model, the three levels were (a) the level of individual circles (individual encoding), (b) the group-level means of the sizes of the red circles and of the blue circles, and (c) the ensemble mean size of all circles in the display. Similar to Equation 4, the three-level model corresponds to using the following prior in Equation 1:

$$p(\theta_{r,1}, \ldots, \theta_{r,N_r}, \theta_{b,1}, \ldots, \theta_{b,N_b}) = \iiint p(\theta_{r,1}, \ldots, \theta_{r,N_r}|\Phi_r)p(\theta_{b,1}, \ldots, \theta_{b,N_b}|\Phi_b)p(\Phi_r, \Phi_b|\Phi) p(\Phi)d\Phi_r d\Phi_b d\Phi, \quad (5)$$

where $\theta_{r,i}$ and $\theta_{b,j}$ are the feature values of individual red and blue circles at the finest level, Φ_r and Φ_b are the summary statistics at the group or color level, Φ is the global ensemble statistics of all circles, and N_r and N_b are the number of red and blue circles, respectively. Since the color-based grouping of circles is explicitly incorporated into the model, biases toward group-level means and toward the global ensemble mean are predicted by this model, mostly consistent with the results of Experiment 1 (whether the authors observed a bias toward the global mean in addition to the group-level bias in Experiment 1 is unclear).

The hierarchical encoding framework of Brady and Alvarez (2011) provided an elegant way of accounting for biases in VSTM by assuming that subjects simultaneously encode items at multiple levels of abstraction in VSTM. As such, their framework furthers our understanding of the organization of VSTM. However, we believe that it also has important disadvantages that are illustrated by considering the specific models proposed by the authors. The two-level and three-level models allow for the representation of

items at multiple levels of abstraction, but there are at least two problems with the way they do so. First, the use of hierarchical models with different numbers of levels to account for different patterns of results observed in different experiments is ad hoc. In general, it is not clear what determines the number of levels that should be used for a given experiment or the appropriate "grain" of those levels. Second, the number of groups (the number of relevant colors) and the assignment of each circle to a group was explicitly specified in an a priori manner when formulating the three-level model (see Equation 5). Although it is easy to do so for the purposes of modeling Experiment 1, it is generally not clear how to define groups or how to assign items to groups in more naturalistic cases. Grouping of visual items is often highly ambiguous both with respect to the number of groups and the assignment of items to groups. A model that could automatically determine these properties and also take into account the uncertainty about them would provide significant explanatory power.

We illustrate these issues within the context of a hypothetical VSTM experiment. Consider an experiment in which subjects are asked to remember the horizontal locations of a number of briefly presented colored squares. A representative display from a single trial of such an experiment is shown in Figure 2 where the vertical locations of the squares are linearly spaced and fixed (represented by the six horizontal lines), whereas their horizontal locations are assigned randomly. Each of the squares in the display can be represented at multiple scales (or levels of abstraction) in VSTM. Consider, for instance, the square that lies on the fifth line from the top. At the finest scale, this square can be represented individually. At a slightly coarser scale, it can be encoded together with the square that lies on the sixth line from the top (whose horizontal location is closest to that of the fifth square). At a still coarser scale, it can be encoded together with the first, second and the sixth squares from the top, and at the coarsest scale, all squares can be represented together.

Once the possibility of encoding items at multiple scales in VSTM is established, three important questions arise: (a) How many scales should be used to represent a set of items, (b) how are the appropriate scales for the representation of items determined, and (c) how much weight should be given to representations at different scales? With respect to the example shown in Figure 2, the previous paragraph mentioned four possible scales for the representation of the fifth square from the top. But why these scales in particular? Instead of representing this square together with the sixth square at an intermediate scale, why not group it with the fourth square and encode them together, or why not introduce a different scale and represent the fifth square together with, say, the third and the fourth squares? Note that the hierarchical modeling approach of Brady and Alvarez (2011) cannot give satisfactory answers to these questions, as the specific scales and groups for representing an item are explicitly specified in advance, and not inferred by the theoretical model. In contrast, as discussed more extensively below, our proposed PCT provides answers to these questions.

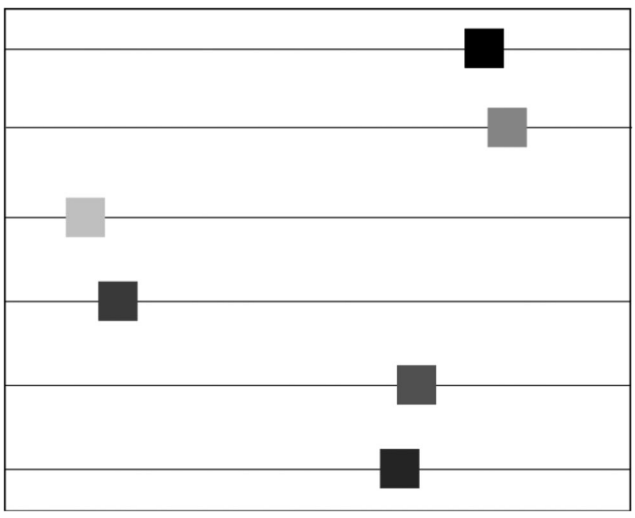

Figure 2. Single trial of a hypothetical visual short-term memory experiment in which subjects are asked to remember the horizontal locations of six colored squares.

PCT is based on describing a set of items in terms of a probability distribution over all possible partitions where each partition might have a different "granularity." Since PCT considers multiple possible partitions, it can represent an item at multiple scales simultaneously (see the next section). Through standard probabilistic (Bayesian) inference, PCT automatically determines the appropriate partitions for the particular set of items at hand and the probabilities or weights that should be allocated to each partition.

Probabilistic Clustering Theory (PCT)

Just as Brady and Alvarez's (2011) two- and three-level hierarchical models can be expressed as specific choices for the prior distribution over the feature values of items (i.e., $p(\theta_1, \ldots, \theta_N)$; Equations 4–5), PCT also corresponds to a specific choice for the prior distribution. In this section, we describe the properties of the prior assumed by PCT in an informal way. Mathematical details will be provided in a later section (*Models*).

Intuitively, the prior assumed by PCT imposes a probabilistic clustering structure on the feature values. According to PCT, an observer's internal model of the generative process for $\theta_1, \ldots, \theta_N$ assumes that these feature values are generated in clusters, even if the actual generative process does not involve any clusters, as is the case in all experiments considered in this article. In other words, the observer assumes that the world has a "clumpy" structure. In estimating the feature values of a set of items based on noisy observations of these feature values, the observer takes into account (or integrates over) his or her uncertainty about the clustering structure of the set of items. This uncertainty might concern both the number of clusters and the assignment of items into clusters.

More specifically, PCT assumes that VSTM automatically infers a probability distribution over all possible partitions of a set of items. Consider a set of three items with feature values denoted by θ_1, θ_2, and θ_3. There are five possible partitions of these items: (a) $\{\theta_1\}, \{\theta_2\}, \{\theta_3\}$ (each item belongs to its own group or cluster); (b) $\{\theta_1, \theta_2, \theta_3\}$ (all items belong to the same cluster); (c) $\{\theta_1, \theta_2\}, \{\theta_3\}$; (d) $\{\theta_1\}, \{\theta_2, \theta_3\}$; and (e) $\{\theta_1, \theta_3\}, \{\theta_2\}$. Based on the similarities among the items' feature values, VSTM infers a distribution over these five possibilities. If, for example, the items are highly similar, VSTM will tend to assign a large probability to the partition that places all items in the same cluster: $\{\theta_1, \theta_2, \theta_3\}$. If

items are highly dissimilar, a large probability will be assigned to the partition that places each item in its own cluster: $\{\theta_1\}$, $\{\theta_2\}$, $\{\theta_3\}$. And if θ_1 and θ_2 are somewhat similar (and somewhat dissimilar), but they are both highly dissimilar from θ_3, then moderate probabilities will be assigned to partitions $\{\theta_1, \theta_2\}$, $\{\theta_3\}$ and $\{\theta_1\}$, $\{\theta_2\}$, $\{\theta_3\}$.

Why might it be rational to group items? As we discuss in more detail in the *General Discussion*, grouping items reduces the variances of the estimates of the feature values of individual items, by effectively sharing information between the estimates of the feature values of items belonging to the same group. Although grouping also introduces biases in the estimates, the reduction in variance might outweigh the increase in bias to reduce the overall expected error of the estimates.

We note that if the number of clusters is constrained to be 1 (i.e., all items are necessarily grouped into a single cluster), then PCT reduces to the two-level hierarchical model (see Equation 4) of Brady and Alvarez (2011). Thus, the latter model can be seen as a special case of PCT, where the one-cluster partition, where all items are assigned to the same cluster, is given a probability of 1, and the other partitions are given a probability of 0. However, in the general case, PCT does not set any a priori bounds on the number of clusters but, instead, determines this automatically from the data. According to PCT, VSTM does not infer a single partition of visual items; rather, it infers a probability distribution over all possible partitions of these items. This enables the representation of items at multiple scales as in the hierarchical modeling approach of Brady and Alvarez (2011). However, unlike the Brady and Alvarez approach, the appropriate scales for the representation of items and their weights are determined automatically from the data (i.e., from the noisy observations of the feature values of items).

Importantly, PCT predicts that observers' VSTM representations will display biases and dependencies. PCT predicts biases in the estimates of feature values of items toward the cluster means of the clusters that they are assigned to. Because PCT does not infer a single clustering, but a probability distribution over many clusterings, in general, there will be biases at multiple scales (toward the means of all clusters that an item might be assigned to), the net effect of which will depend on the posterior probabilities, or the weights, of different clusterings.

PCT also predicts dependencies between the estimates of feature values of different items assigned to the same cluster. Two items that are never assigned to the same cluster will not share parameters and hence their estimates will be independent. In contrast, items assigned to the same cluster share parameters and thus their estimates are dependent. PCT predicts that the magnitude of the dependence between the estimates of the feature values of two items should increase with the similarity between the feature values of the items. This is because, according to PCT, the more similar two items are, the more likely they are to be assigned to the same cluster, hence to share parameters. In a later section (*Experiments*), we provide experimental evidence supporting this crucial prediction of PCT.

A more detailed discussion of the predictions of PCT regarding biases and dependencies in VSTM are given in *Models* below.

Relationship Between PCT and Previous Works on Hierarchical Encoding in Human Memory

Prior to Brady and Alvarez (2011), several other researchers argued that human memory is organized hierarchically, thereby accounting for categorical effects in memory (Hemmer & Steyvers, 2009a, 2009b; Huttenlocher, Hedges, & Duncan, 1991). Consistent with the predictions of a hierarchical Bayesian model, Hemmer and Steyvers (2009b) showed that in an episodic recall task, subjects' estimates were biased toward both the mean feature value of the presented object category (e.g., mean size of apples) and the mean feature value of the superordinate category (e.g., mean size of all fruits). Hemmer and Steyvers (2009a) extended this result to unfamiliar objects and developed a hierarchical Bayesian model that could account for the distinct pattern of biases observed for familiar and unfamiliar objects. For familiar objects, the model predicted, in accord with experimental results, that the recalled size of an object would be biased more toward the mean size of the objects of that kind. For unfamiliar objects, the recalled size was be biased toward the superordinate-level mean (e.g., mean size of fruits).

These biases are similar to the biases observed in VSTM experiments. PCT accounts for these analogous biases in VSTM, not as an effect of preexisting categories in memory but, rather, as an effect of encoding displayed items at multiple scales. To make the similarity to the work of Hemmer and Steyvers (2009b) clearer, one can argue that in VSTM experiments, subjects spontaneously form clusters or "categories" at multiple scales when shown multi-item displays. These categories influence short-term memories for visual items in a way that is similar to the way categories in long-term memory affect episodic recall performance of individual items.

Huttenlocher et al. (1991) demonstrated that people show systematic biases even in the simple task of estimating the spatial location of a single dot presented within a circle. Subjects showed a bias toward the centers of the four quadrants dividing the circle in their judgments of the angular locations of single dots presented for a brief duration. Huttenlocher et al. (1991) conceived of the four quadrants as categories and the centers of the quadrants as prototypical examples of those categories. Their model of how subjects estimated spatial locations was essentially the same as Brady and Alvarez's (2011) two-level hierarchical Bayesian model reviewed above, with "categories" or quadrants providing the higher level representations. Similar to the model of Hemmer and Steyvers (2009a), their model also assumed preexisting knowledge of categories (in this case, quadrants). Huttenlocher, Hedges, and Vevea (2000) extended these results by showing that similar biases were evident when subjects estimated other perceptual features of objects belonging to inductively defined categories and that the observed biases were influenced by properties of the distributions describing these categories.

Brady and Tenenbaum (2010) developed a discrete slot-based model of VSTM that encodes both high-order structure about a simple display of dot patterns and detailed information about specific dots. They formalized the concept of high-order structure in terms of a correlation parameter, called the "gist" parameter, in a Markov random field such that larger values of this parameter corresponded to images that tended to have similarly colored neighbors, whereas smaller values corresponded to images that

tended to have differently colored neighbors. In addition, the model assumed the encoding of detailed information about K specific dots in such a way that "outlier" dots (dots that did not conform to the overall gist of the image) were more likely to be encoded. Brady and Tenenbaum (2010) presented evidence suggesting that the performance of this "gist + exception" model of encoding correlated well with human performance on an image-by-image basis.

Brady and Tenenbaum's (2010) gist representation applies only to images sampled from a Markov random field, whereas the representations in PCT can be applied more generally. The individual encoding mechanism in their model is biased toward exceptions to the gist. It is interesting that PCT implements a similar bias because the further the feature value of an item is from feature values of other items in a display, the more likely that the item will be assigned to a cluster of its own, meaning that the item will be encoded individually.

The hierarchical models described in this section are significant because they advance our understanding of how hierarchical representations can account for biases and categorical effects observed in human memory. A fundamental difference between the previously proposed hierarchical models and PCT is that the hierarchical models assume prespecified and fixed levels of abstraction to represent items. Hemmer and Steyvers (2009a, 2009b) assumed two levels of abstraction: an object-based level (e.g., apples) and a categorical level (e.g., fruits). Huttenlocher et al. (1991) also assumed two fixed levels of abstraction: a fine-grained representation of the location of a dot and a more global, coarse-grained prototype representation based on prespecified prior knowledge. Brady and Tenenbaum's (2010) model represents a display at two fixed levels of abstraction, and Brady and Alvarez's (2011) hierarchical modeling framework assumes an appropriate prespecification of the levels of abstraction for the representation of items.

In contrast, an innovation of our approach is that it automatically determines multiple scales or levels of abstraction that are appropriate for the representation of an item in VSTM and how to weight the different scales without assuming prespecified and fixed levels of abstraction. It does so by inferring multiple partitions with different granularities that are appropriate for a given set of items along with the posterior probabilities of those partitions. Our focus is on VSTM, but we speculate that studies of other human memory systems would benefit from our approach which emphasizes the use of multiscale representations without prespecified, fixed hierarchies.

Models

This section describes the specific computational models that will be used in the following sections. Three models are described. The first model is a Dirichlet process mixture model (DPMM), also known as an infinite mixture model. The second model is a Bayesian finite mixture model (BFMM). Both models automatically infer posterior distributions over multiple partitions of a set of items. The only difference between these two models is that they make different assumptions about the maximum number of clusters that the data can be grouped into. The DPMM does not set any a priori limit on the maximum number of clusters, whereas a BFMM assumes that the data can be grouped into at most K clusters, for a finite K specified in advance. Because of this difference, the DPMM can infer a probability distribution over all possible clusterings or partitions of a set of items but a BFMM with too few clusters cannot. For this reason, we regard the DPMM as an exact implementation of the PCT, whereas the BFMM can be regarded as an approximation to the DPMM (where the quality of approximation will be determined by K). The last model we describe is the two-level hierarchical Bayesian model (HBM) proposed by Brady and Alvarez (2011). As discussed below, the HBM can be regarded as a special case of BFMMs where the number of components K is constrained to be 1. A graphical representation of the models considered in this article is shown in Figure 3.

Dirichlet Process Mixture Models (DPMMs)

DPMMs are commonly used in statistics and machine learning (Görür & Rasmussen, 2010; Neal, 2000). In cognitive science, they are gradually becoming popular in the study of perception and cognition where they have been used as normative models of word segmentation (Goldwater, Griffiths, & Johnson, 2009), causal learning (Gershman, Blei, & Niv, 2010) and categorization (Sanborn, Griffiths, & Navarro, 2010). Excellent introductions to these models can be found in Goldwater et al. (2009) and Navarro, Griffiths, Steyvers, and Lee (2006).

Here, we describe the application of the DPMM to the problem of encoding multiple items in VSTM. Consider a single trial of a hypothetical VSTM experiment in which an observer is asked to remember the feature values (e.g., horizontal locations of squares or orientations of Gabor gratings) of N items in a display. For the moment, we consider items defined by a single feature (e.g., position, orientation, color, shape). We denote the actual feature value of item i by θ_i. As laid out in an earlier section, we assume that the observer does not have access to the actual feature values of the items, but to noise-corrupted observations thereof, denoted by x_i. In addition, the observer's internal model of the generative process for θ_is assumes that these feature values are generated in clusters (even if the actual generative process does not involve any clusters). In estimating the feature values of a set of items based on the corresponding noisy observations, the observer integrates out its uncertainty about the clustering structure of the set of items. Mathematically, the full model can be specified as follows (see Figure 3):

$$G \sim DP(G_0, \alpha) \tag{6}$$

$$G_0(\mu_i, \tau_i) = \mathcal{U}(\mu_i;\ a,\ b)\mathcal{G}(\tau_i;\alpha_\tau, \beta_\tau) \tag{7}$$

$$\mu_i, \tau_i | G \sim G \tag{8}$$

$$\theta_i | \mu_i, \tau_i \sim \mathcal{N}(\theta_i; \mu_i, \tau_i) \tag{9}$$

$$x_i | \theta_i \sim \mathcal{N}(x_i; \theta_i, \tau_{obs}) \tag{10}$$

Here, $x_i \sim N(x_i;\ \theta_i,\ \tau_{obs})$ means that x_i is distributed according to a normal distribution with mean θ_i and precision τ_{obs}. The precision τ_{obs} is meant to capture the combined effects of both sensory and memory noise in generating noisy internal observations of the actual feature values. However, sensory noise is likely to be negligible compared with memory noise. We assume that τ_{obs} may depend on set size, but otherwise it is identical for all items in a given trial and across different trials with the same set size. Recent results suggest that introducing variability in τ_{obs} across trials and across items in a given trial can lead to better models of capacity

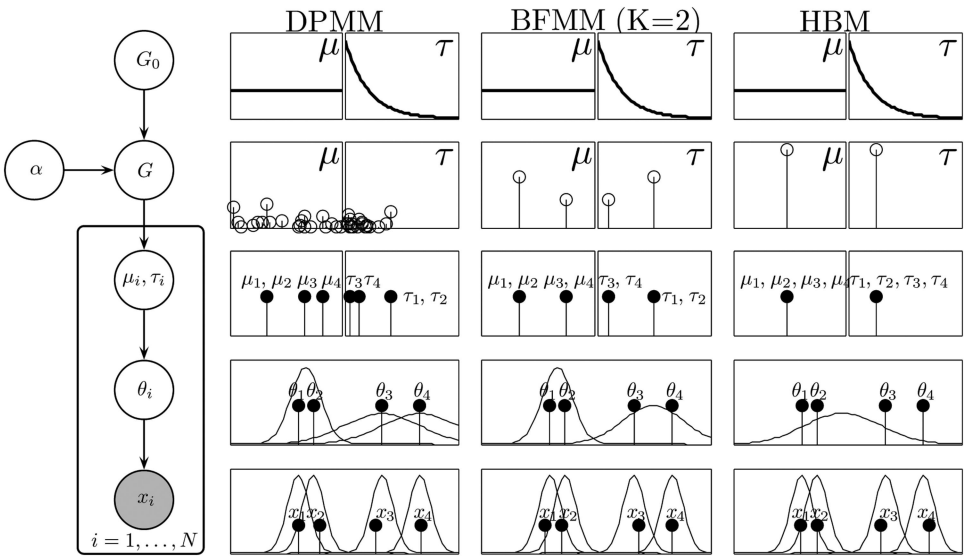

Figure 3. A graphical representation of the models considered in this article. All models have a common structure represented by the graphical model shown on the left. We illustrate this graphical model using plate notation, where the nodes inside the plate are meant to be replicated *N* items. The shaded node represents the observable variables (i.e., the noisy observations x_i). The other variables are latent or unobservable. The remaining plots illustrate the generative processes defining the models. Each row illustrates the generation of variables at the corresponding level in the graphical model on the left. The only difference between the models is in the variable *G*. For the Dirichlet process mixture model (DPMM), *G* is a discrete distribution with an infinite number of "atoms"; for the Bayesian finite mixture models (BFMMs), it is a discrete distribution with *K* atoms; and for the hierarchical Bayesian model (HBM), it is a single atom. In the example shown here, the DPMM uses three clusters to generate the four items represented by θ_is, the BFMM uses two clusters, and the HBM uses a single cluster. μ_i (mean) and τ_i (precision) represent the cluster parameters for item *i*. The distributions at the bottom two rows illustrate the distributions from which the variables at the corresponding levels were drawn.

limitations in VSTM (van den Berg, Shin, Chou, George, & Ma, 2012). We found that variability in τ_{obs} was not essential for accounting for the phenomena we consider in this article (i.e., biases and dependencies in VSTM); therefore, for simplicity, we decided to assume identical τ_{obs} across trials and across items. μ_i and τ_i represent the mean and the precision of the cluster to which item *i* belongs, and they are jointly distributed according to a countably infinite discrete distribution denoted by *G*. *G* is itself distributed according to a Dirichlet process with base distribution G_0 and concentration parameter α. The base distribution G_0 is the product of a uniform distribution for mean μ_i defined over the interval [*a, b*] and a gamma distribution for precision τ_i with scale parameter α_τ and shape parameter β_τ.

In a given trial, the observer's goal is to infer the feature values of the items, or the feature value of a single target item, given the noisy observations $\{x_i\}_{i=1}^N$. This problem can be formalized as the inference of the joint posterior distribution $p(\{\theta_i\}_{i=1}^N|\{x_i\}_{i=1}^N)$ if the feature values of all items are to be estimated, or the marginal posterior $p(\theta_t|\{x_i\}_{i=1}^N)$ if only the feature value of a single target item *t* is to be estimated. In the simulations reported below, these posterior distributions were computed using a Markov chain Monte Carlo (MCMC) sampling algorithm with auxiliary variables (Algorithm 8 of Neal, 2000; see Görür, 2007, and Appendix A for additional details). For recall tasks, we then use the mean of the marginal posterior distribution as the observer's estimate of the feature value of the target item in that trial. For old/new recognition tasks, on the other hand, the posterior distribution is transformed into probabilities of responding "old" or "new" to a given probe item (see the next section for details).

We now describe the clustering properties of the DPMM, given by Equations 6–9, in more detail. Let $\mathbf{p}$ denote a vector or "atom" in (μ, τ) space (i.e., $\mathbf{p}$ is a possible set of values for mean μ and precision τ). It can be shown that a single draw, *G,* from a Dirichlet process is a countably infinite discrete distribution over atoms $\mathbf{p}$ (Ferguson, 1973). That is, *G* is a weighted sum of an infinite number of discrete atoms $\{\mathbf{p}_k\}_{k=1}^\infty$:

$$G(\mathbf{p}) = \sum_{k=1}^{\infty} \pi_k \delta(\mathbf{p} = \mathbf{p}_k). \tag{11}$$

The base distribution of the Dirichlet process, G_0, is a prior distribution over the (μ, τ) space. G_0 determines the locations of the atoms because $\{\mathbf{p}_k\}_{k=1}^\infty$ are independent samples from G_0. The concentration parameter α to the Dirichlet process determines the weights of the atoms π_k in Equation 11. For small values of α, a small number of atoms are given large weights, and the rest are assigned very small weights. For large values of α, weights are distributed more broadly across atoms.

A clearer understanding of the concentration parameter α emerges when one considers the relationship between α and the clustering properties of the DPMM. Recall that the mean μ_i and the precision τ_i are the parameters of the cluster to which item *i* is assigned. When performing inference, μ_i and τ_i for different items *i* are assigned

identical values if these items are assigned to the same component. The concentration parameter α acts as a bias on this assignment process by influencing the probability that two items will be grouped together. Roughly, α controls the observer's tendency to group items. For small values of this parameter, the model is biased toward a small number of clusters or groups of similar items ("chunks"). For large values, it tends to assign each item to its own cluster.

As indicated above, the base distribution of the Dirichlet process G_0 is the product of a uniform distribution over μ_i and a gamma distribution over τ_i. Since we apply the model to small data sets (typically displays with two to eight items), it is important to use a relatively noninformative base distribution for μ_i. Otherwise its posterior distribution would be strongly affected by the base distribution, creating large biases due to the choice of this distribution alone. Consequently, we use a uniform base distribution over μ_i defined over a sufficiently large interval (thereby making the model nonconjugate; see Görür & Rasmussen, 2010). We set the range of the uniform distribution, $[a, b]$, to a sufficiently large interval that includes the minimum and maximum possible values for the relevant variable in each experiment considered below. For the parameters of the gamma distribution on τ_i, we put a $\mathcal{G}(1, 1)$ prior on scale parameter β_τ and set $\alpha_\tau = 1$. Last, we put a $\mathcal{G}(\alpha_c, 1)$ prior on the Dirichlet process concentration parameter α and treat α_c as a free parameter. For a given set size, this reduces the number of free parameters to just two, namely, α_c (a prior parameter for concentration parameter α) and τ_{obs} (memory precision). The same values of α_c and τ_{obs} were used for all trials of a simulated experiment.

We illustrate the working of the model with a simple example in Figure 4. Figure 4A shows three noisy observations, x_1, x_2, and x_3 (represented by the vertical lines), and the three marginal posteriors, $p(\theta_1|x_1, x_2, x_3)$, $p(\theta_2|x_1, x_2, x_3)$, and $p(\theta_3|x_1, x_2, x_3)$ (represented by the solid curves), for four different settings of the parameters. Figure 4B shows the posterior distributions over the number of clusters for each setting of the parameters in Figure 4A. The marginal posteriors in Figure 4A display biases (i.e., their means are not centered on the x_is). This is because an item is often grouped with one or both other items, shifting the marginal posteriors toward the x_is associated with those items. This is essentially the mechanism by which the DPMM accounts for the biases reviewed in *Biases and Dependencies in VSTM* above (also see *Simulations* below). Increasing α_c forces the model to use a larger number of clusters. This has the effect of reducing biases in the marginal posteriors, because each item is now more likely to be assigned to its own cluster. Increasing τ_{obs}, on the other hand, reduces the variance of the marginal posteriors and also reduces the biases, because when τ_{obs} is high, the model relies more heavily on the observations, x_i, in computing the posteriors over θ_is, and less on the prior over θ_is induced by the DPMM.

The DPMM also predicts dependencies between the estimates of feature values of different items encoded in VSTM. These dependencies arise in the model when different items are grouped into the same normal component. Specifically, the representations corresponding to these items become dependent due to the shared parameters of the normal component. Importantly, the model further predicts that the dependency between the memory representations of two items should decrease with the distance between their feature values. Intuitively, this is because the probability that two items will be assigned to the same component decreases with the distance between their feature values. In the extreme case, if two items are highly dissimilar, they will never be assigned to the same component by the model and there should be no dependency between the representations of those items. Conversely, if two items are highly similar and, thus, consistently assigned to the same component, there should be a high degree of dependency between their representations, the exact magnitude of which de-

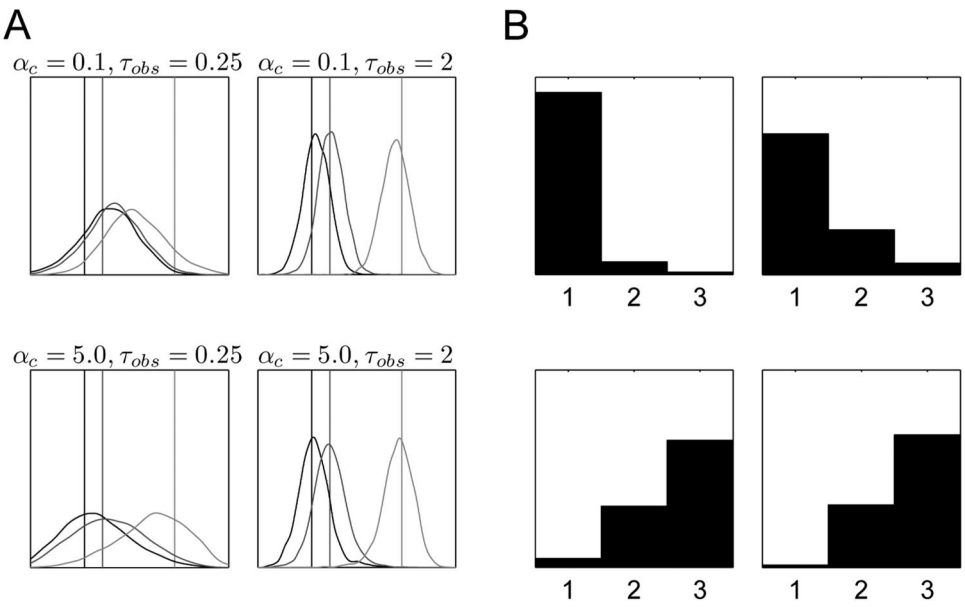

Figure 4. A. Three noisy observations—x_1, x_2, and x_3 (vertical lines)—and the three marginal posteriors—$p(\theta_1|x_1, x_2, x_3)$, $p(\theta_2|x_1, x_2, x_3)$, and $p(\theta_3|x_1, x_2, x_3)$ (solid curves)—for four different settings of the parameters, α_c and τ_{obs}. B. Posterior distributions over the number of clusters for each of the corresponding subplots in A.

pends on other factors, such as the precision of the component that they are both assigned to. In *Experiments* below, we present experimental evidence supporting this prediction.

Figure 5 illustrates this prediction of the model with a simple example. The leftmost plot in Figure 5 shows three items with feature values θ_1, θ_2 and θ_3 (vertical lines) and the marginal distributions of the estimates [i.e., $p(\hat{\theta}_i|\theta_1,\theta_2,\theta_3)$] (solid curves) computed over 1,000 simulated presentations of the same set of feature values. The remaining plots show each of the three two-dimensional marginals of the estimates [i.e., $p(\hat{\theta}_i,\hat{\theta}_j|\theta_1,\theta_2,\theta_3)$]. (Note that the distributions shown here are different from the ones shown in Figure 4A. The distributions in Figure 4A depict $p(\theta_i|x_1, x_2, x_3)$ for a specific set of noisy observations x_1, x_2, x_3. For the data plotted in Figure 5, noisy observations are integrated out. The relationship between the distributions depicted in Figures 4A and 5 is as follows: The posterior mean of the distribution depicted in Figure 4A would correspond to a single point in Figure 5 [also see Equation 2].) The two-dimensional marginals of the estimates in Figure 5 show that the estimates are correlated and the correlations decrease with the difference between the actual feature values of the items. The biases in the estimates are also apparent in this figure. Note, for example, that the means of the two-dimensional marginals (represented by the circles) are closer to the diagonal than the actual feature values of the items (represented by the crosses), indicating that the estimates are biased toward the mean of the feature values.

Multivariate extension. We also consider a multivariate version of the DPMM presented above, where items are now defined not by a single feature dimension, but by multiple feature dimensions. In this case, the univariate normal components are replaced by multivariate normal components. In detail, the multivariate DPMM is defined by the following equations (see Görür & Rasmussen, 2010):

$$G \sim DP(G_0, \alpha) \tag{12}$$

$$G_0(\mu_i, \Sigma_i) = \mathcal{U}(\mu_i;\mathbf{a}, \mathbf{b})\mathcal{IW}(\Sigma_i;\Psi, \kappa) \tag{13}$$

$$\mu_i, \Sigma_i|G \sim G \tag{14}$$

$$\theta_i|\mu_i, \Sigma_i \sim \mathcal{N}\left(\theta_i;\mu_i, \Sigma_i\right) \tag{15}$$

$$x_i|\theta_i \sim \mathcal{N}(x_i;\theta_i, \Sigma_{obs}), \tag{16}$$

where θ_i, μ_i and x_i are now d-dimensional vectors, $\mathcal{N}(\theta_i; \mu_i, \Sigma_i)$ is a multivariate normal distribution with mean μ_i and covariance matrix Σ_i. Σ_{obs} is the common covariance matrix of the noisy observations, x_i. We assume Σ_{obs} to be a diagonal matrix in accordance with recent findings that recall errors are largely independent across different stimulus dimensions in VSTM (Bays, Wu, & Husain, 2011; Fougnie & Alvarez, 2011). The uniform base distribution for μ_i, $\mathcal{U}(\mu_i;$ **a, b**), is defined over a d-dimensional hypercube. Similar to the univariate case, we set the region over which the uniform base distribution for μ_i is defined to a large volume that includes the minimum and maximum possible values of each component of μ_i. The base distribution for Σ_i is an inverse-Wishart distribution with inverse scale parameter Ψ and degrees-of-freedom parameter κ. We place a vague inverse-Wishart prior on Ψ and treat the degrees-of-freedom parameter κ as a free parameter. The concentration parameter α is given a $\mathcal{G}(1, 1)$ prior. As in the univariate case, posterior inference is performed via an MCMC algorithm with auxiliary variables (Algorithm 8 in Neal, 2000).

Bayesian Finite Mixture Models (BFMMs)

In a finite mixture model, each θ_i is assumed to be generated by one of K Gaussian components, where K is a fixed, finite positive integer. Formally, a Bayesian finite mixture of Gaussians is very similar to the DPMM introduced above (Equations 6–10). The only difference between the DPMM and a BFMM comes from the discrete distribution G over the component parameters (see Figure 3). In the DPMM, G is distributed according to a Dirichlet process with base distribution G_0 and concentration parameter α and can be expressed as a weighted sum of an infinite number of discrete atoms, where atoms represent component parameters (see Equation 11). In a finite mixture model, on the other hand, G is a weighted sum of a finite number of atoms only (reflecting the assumption that the data were generated by a fixed, finite number of components):

$$G(\mathbf{p}) = \sum_{k=1}^{K} \pi_k \delta(\mathbf{p} = \mathbf{p}_k). \tag{17}$$

As in the DPMM, the atoms (i.e., the component parameters) are drawn independently from a base distribution G_0. The component weights π, on the other hand, are drawn from a symmetric Dirichlet prior with concentration parameters α/K:

$$\pi \sim \text{Dirichlet}\,(\alpha/K, \ldots, \alpha/K), \tag{18}$$

whereas the weights π in the DPMM are distributed according to what is known as a GEM (or stick-breaking) process with concen-

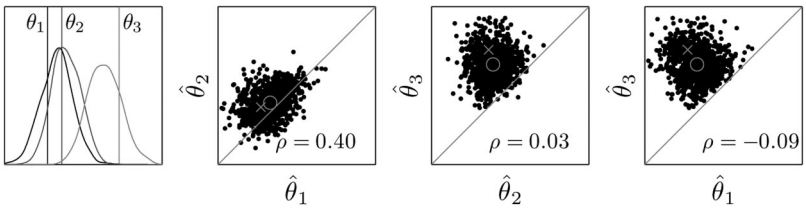

Figure 5. The leftmost plot depicts three items with feature values θ_1, θ_2, and θ_3 (vertical lines) and the marginal distributions of the estimates (i.e., $p(\hat{\theta}_i|\theta_1,\theta_2,\theta_3)$; solid curves) computed over 1,000 simulated presentations of the same set of feature values. The remaining plots show each of the three two-dimensional marginals of the estimates (i.e., $p(\hat{\theta}_i, \hat{\theta}_j|\theta_1,\theta_2,\theta_3)$). The means of the marginals are represented by the circles and the actual feature values of the items are represented by the crosses. The numbers inside the plots are the correlation coefficients between the estimates of each pair of items.

tration parameter α ($\pi \sim \text{GEM}(\alpha)$). The close similarity between the DPMM and the BFMM is not accidental. Indeed, it can be shown that the DPMM is mathematically equivalent to a BFMM in the limit $K \rightarrow \infty$ (Rasmussen, 2000).

We use the same base distribution and hyper-priors for the BFMM as for the DPMM. Specifically, for the base distribution, we use $G_0(\mu_i, \tau_i) = \mathcal{U}(\mu_i; a, b)\mathcal{G}(\tau_i; \alpha_\tau, \beta_\tau)$ (we put a $\mathcal{G}(1, 1)$ prior over β_τ and set $\alpha_\tau = 1$). We put a $\mathcal{G}(\alpha_c, 1)$ prior over the precision parameter α of the BFMM and treat α_c as a free parameter. Thus, the DPMM and BFMM have the same number of free parameters. As with the DPMM, it is also straightforward to extend the BFMM to multivariate components. In what follows, we consider BFMMs with $K = 2$ and $K = 4$ components.

BFMMs predict qualitatively similar biases and dependencies in VSTM as DPMMs, using essentially the same mechanisms. However, the quantitative details of the biases and dependencies predicted by a BFMM might depend on K. In general, for larger and larger K, the predictions of a BFMM will be more and more similar to the predictions of a DPMM. Indeed, a popular algorithm for performing efficient approximate inference in the DPMM truncates the infinite sum in Equation 11 at a finite but sufficiently large level, thus making the model identical to a BFMM (Ishwaran & James, 2001).

Hierarchical Bayesian Model (HBM)

In the following sections, we also consider the two-level hierarchical Bayesian model (HBM) used in Brady and Alvarez (2011). The model assumes the following generative process for a single trial of a VSTM experiment (Brady & Alvarez, 2011):

$$\mu, \tau \sim \mathcal{U}(\mu; a, b)\mathcal{G}(\tau; \alpha_\tau, \beta_\tau) \tag{19}$$

$$\theta_i | \mu, \tau \sim \mathcal{N}(\theta_i; \mu, \tau) \;\; i = 1, \ldots, N \tag{20}$$

$$x_i | \theta_i \sim \mathcal{N}(x_i; \theta_i, \tau_{obs}) \; i = 1, \ldots, N. \tag{21}$$

As in the DPMM and BFMMs, $\{\theta_i\}_{i=1}^N$ (as well as the ensemble statistics μ and τ) are treated as latent or unobserved variables that the observer does not have access to. Instead, the observer only has access to noisy observations $\{x_i\}_{i=1}^N$ each generated from a corresponding Gaussian distribution with mean θ_i and some constant variance representing the memory noise (or the combined effect of sensory and memory noise). Given these noisy observations $\{x_i\}_{i=1}^N$, the observer then infers the joint posterior distribution over $\{\theta_i\}_{i=1}^N$. The group-level mean μ is given a uniform prior over a sufficiently large range. To make the two-level HBM truly a special case of BFMM (with $K = 1$), we use the same prior over the group-level precision τ as in the BFMM, namely, a $\mathcal{G}(\alpha_\tau, \beta_\tau)$ prior where β_τ is, in turn, given a $\mathcal{G}(1, 1)$ prior and α_τ is set to 1. Brady and Alvarez (2011) use a different prior over τ (they use a uniform prior over the group-level standard deviation $1/\sqrt{\tau}$), but we found that this difference did not significantly affect the simulation results reported below. The individual memory precision τ_{obs} is treated as the only free parameter of the model.

As the specification of the model in Equations 19–21 makes clear, the two-level HBM can be regarded as a special case of BFMMs where the number of components K is constrained to be 1 (see also Figure 3).

Simulations

This section studies the DPMM, two versions of the BFMM, and the HBM in the context of three experiments from the visual short-term recall and recognition memory literatures. Our focus will be on the DPMM. Although a BFMM with a sufficiently large K performs as well as a DPMM, we focus on the DPMM because it is an exact implementation of PCT and, as discussed at the end of this section, has conceptually appealing properties not shared by BFMMs (e.g., not setting an a priori limit on the number of clusters into which a set of items can be grouped). We model experimental results from two short-term recall tasks (Brady & Alvarez, 2011; Wilken & Ma, 2004) and a short-term recognition task (Viswanathan, Perl, Visscher, Kahana, & Sekuler, 2010). We also quantitatively compare the fits of the DPMM with those of the BFMMs and the two-level HBM, using the Bayesian information criterion (BIC) measure (Schwarz, 1978). Bayesian model comparison depends on the calculation of the marginal log-likelihood of the data under different models. BIC provides only an approximation to the marginal log-likelihood of the data under a given model. We opted for the BIC measure primarily due to computational considerations (it was relatively easy to compute the BIC values given the optimization procedure we adopted in our simulations; see Appendix B). Given the similarity of the structures of the models compared in this article (see Figure 3), it is difficult to see how BIC would unfairly favor one model over the others.

Biases in VSTM: Wilken and Ma (2004)

As briefly mentioned before, Wilken and Ma (2004) found that subjects displayed systematic biases in their judgments in a VSTM experiment that used spatial frequency as the relevant feature. In each trial of their Experiment 9, subjects briefly viewed a number of Gabor stimuli with different spatial frequencies randomly drawn from 16 frequency values uniformly spaced between four and eight cycles/degree. Different set sizes used in the experiment were $N = 2, 4, 6, 8$. After a delay interval, one of the N Gabors, called the target Gabor, was cued, and subjects adjusted the frequency of a comparison Gabor using the arrow keys to indicate their estimate of the frequency of the target Gabor in the original display. Wilken and Ma (2004) found that subjects tended to overestimate the spatial frequencies of low frequency Gabors, but tended to underestimate the spatial frequencies of high frequency Gabors (i.e., subjects showed a bias toward the mean spatial frequency in their judgments). These authors also showed that the magnitude of this bias depended on the set size with smaller set sizes leading to smaller biases (see Figure 8 in Wilken & Ma, 2004; also reproduced in Figure 6 here).

We sought to determine whether the DPMM could explain the biases observed by Wilken and Ma (2004). We first generated a data set according to the procedure described above. For each simulated trial, we randomly selected N spatial frequency values from 16 frequencies uniformly spaced between four and eight cycles/degree. We then generated noisy observations of each of the N items from Gaussian distributions with mean equal to the true spatial frequency of the item and precision τ_{obs}. For each set size N, 4,000 such trials were simulated. We then ran the univariate version of the DPMM on noisy observations from these simulated trials. In each trial, we used the mean of the marginal posterior over the target spatial frequency as the model's response:

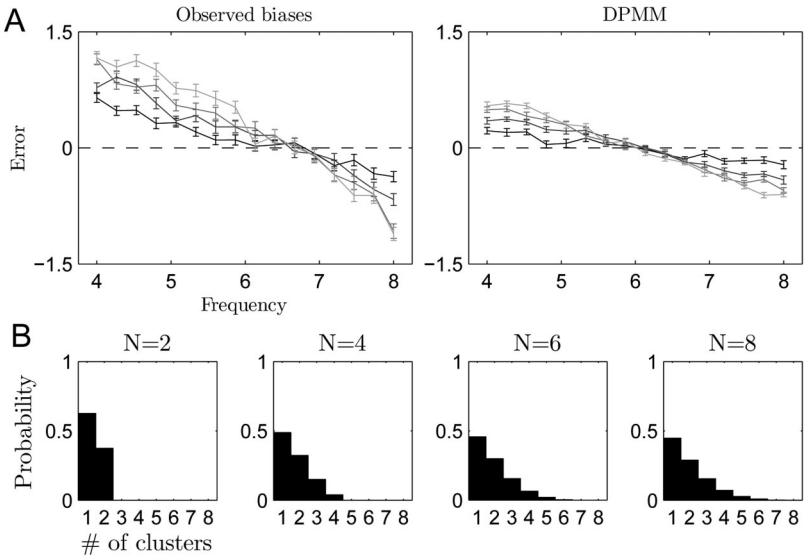

Figure 6. A. The observed biases (from Wilken & Ma, 2004) and biases predicted by a nonoptimized Dirichlet process mixture model (DPMM) with $\alpha_c = 1$ and $\tau_{obs} = 1$ for four different set sizes. Lighter colors represent larger set sizes. Error bars represent ± 1 *SEM* across subjects. B. Posterior distributions over the number of clusters inferred by the DPMM averaged over all trials for different set sizes.

$\hat{\theta}_t = E[\theta_t | \{x_i\}_{i=1}^N]$. We confirmed that using the posterior mode instead of the posterior mean yielded similar results.

To demonstrate that the ability of the DPMM to qualitatively explain the pattern of biases observed by Wilken and Ma (2004) did not critically depend on the optimization of the free parameters, we first arbitrarily set $\alpha_c = 1$ and $\tau_{obs} = 1$ for all set sizes and did not optimize these free parameters. Figure 6 illustrates the behavior of the DPMM with this fixed setting of the parameters. The DPMM was able to capture the two main qualitative patterns in the observed biases: a linear relationship between the bias and frequency of the target Gabor for all set sizes, and an increase in the magnitude of the bias with set size. We emphasize that the DPMM was able to explain the latter phenomenon without having to use different parameter values for different set sizes.

Model comparison. For the purposes of model comparison, we calculated the maximum likelihood (ML) estimates of the free parameters of the models introduced in the previous section: the DPMM, BFMMs with $K = 2$ and $K = 4$ components, and the HBM. An alternative approach would be to put noninformative or vague priors over these parameters and perform Bayesian inference to compute their posteriors. Computational infeasibility prevented us from taking this approach. The free parameters of the models were optimized via simple grid searches to find the parameter values that maximized the log-likelihood given the mean observed biases of 15 subjects (details of the model evaluation and optimization procedures are provided in Appendix B). When fitting the DPMM and BFMMs to observed biases, τ_{obs} was allowed to vary across different set sizes, but α_c was fixed across different set sizes. Although fixing both free parameters across different set sizes produced biases that increased with set size (consistent with the biases observed in the experimental data), the differences between biases for different set sizes were less dramatic for the models than in the experimental data (see, for example, Figure 6 where τ_{obs} was fixed at 1 for all set sizes). Allowing τ_{obs} to vary across different set sizes helped the models achieve better fits to the observed biases. This is consistent with existing hypotheses about relationships between task demands and precision of representations. Wilken and Ma (2004) and Bays and Husain (2008) reported a monotonic decline with set size in the precision with which individual items can be encoded. When we fit the DPMM to data from Wilken and Ma (2004), we found that the best fits were obtained if we allowed τ_{obs} (which controls the precision of memory noise) to vary across set sizes such that the precision of memory noise monotonically decreased with set size. For the BFMMs and the HBM, the memory precision parameter τ_{obs} was allowed to vary across different set sizes in a similar manner. Model fits were compared using the BIC measure (see Appendix B for details).

Results. Overall, all four models were able to capture the linear relationship between the bias and target frequency for all set sizes and the increase in the magnitude of the bias with set size. Table 1 documents the BIC values of the models relative to the BIC value of the DPMM. The two-level HBM was slightly favored over the other models due to its smaller number of parameters. However, the differences between the BIC scores of different models were small. For the DPMM (as well as for BFMMs) the posterior distributions were dominated by partitions with small numbers of clusters (typically one or two clusters), suggesting that subjects tended to group items into a small number of clusters.

All four models account for the biases by assuming that subjects spontaneously encode a given display at multiple scales. However, an alternative explanation of the observed biases would be that subjects might simply have a bias toward reporting the overall mean of the range of presented frequencies and that this bias increases with set size. We believe that this latter explanation is unlikely for two reasons. First, in a similar experiment, Huang and

Table 1
BIC Values of the BFMMs ($K = 2$ and $K = 4$) and the Two-Level HBM Relative to the BIC Value of the DPMM on Three Previous Studies That Reported Biases in VSTM

Study	BFMM ($K = 2$)	BFMM ($K = 4$)	HBM
Wilken & Ma (2004)	0.0951	0.0644	−2.8290
Viswanathan et al. (2010)	61.3027	1.3165	177.3836
Brady & Alvarez (2011)	0.2810	0.1692	1.7824

Note. BIC = Bayesian information criterion; BFMM = Bayesian finite mixture model; HBM = hierarchical Bayesian model; DPMM = Dirichlet process mixture model; VSTM = visual short-term memory. Negative values indicate better fits than the DPMM, positive values worse fits. Smaller values indicate better fits.

Sekuler (2010) teased apart the contributions to the observed biases of the overall mean of the frequencies presented to the subject on previous trials versus the frequencies presented on the current trial and found that both make significant contributions to the observed biases, suggesting that the biases cannot be completely attributed to a general bias toward reporting the overall mean frequency. Second, as discussed below, in a carefully controlled experiment, Brady and Alvarez (2011) showed that subjects displayed a specific bias toward the mean size of the same-colored circles presented on the same trial as a target circle that, because of the way their experiment was designed, could not be attributed to a general bias toward the overall mean size of the circles of a given color.

Inter-Item Similarity Effect: Viswanathan et al. (2010)

As discussed above, Kahana and Sekuler (2002) showed that interitem similarity between stimuli influences subjects' judgments in a standard old/new recognition task. Assuming fixed probe-item similarities, they found that a smaller interitem similarity (i.e., a less homogeneous set of stimuli) increases the likelihood that subjects will judge a probe to be an old or familiar item.

The interitem similarity effect has been replicated in several other studies. Here, we consider a study by Viswanathan et al. (2010). The design of their experiment was, for our purposes, equivalent to the design of the experiments in Kahana and Sekuler (2002) described in a previous section (*Biases and Dependencies in VSTM*). On each trial, a subject viewed three Gabor gratings, referred to as study gratings, followed by a probe grating. The subject then judged whether the spatial frequency of the probe was "old" (the same as the frequency of one of the study gratings) or "new" (a novel frequency).

The experiment used both medium and high homogeneity conditions. Representative trials for these two conditions are schematically depicted in the left and middle plots of Figure 7. In this figure, the spatial frequencies (in just-noticeable-difference or JND units) of study gratings are represented by solid vertical lines (at 1, 4, and 8 JND in the medium homogeneity condition, and at 3, 4, and 8 JND in the high homogeneity condition), and the frequencies

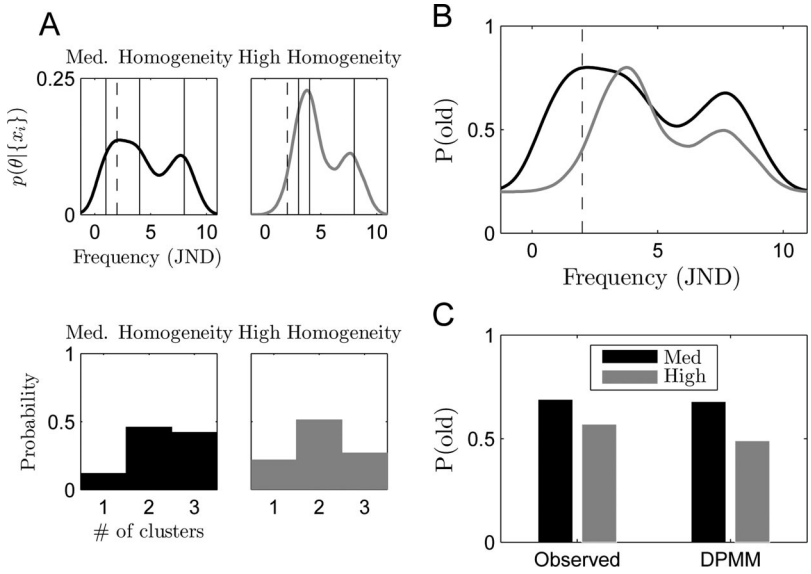

Figure 7. Predictions of a nonoptimized Dirichlet process mixture model (DPMM) with parameters set to $\alpha_c = 1$ and $\tau_{obs} = 1$. A. Representative trials from the medium- and high-homogeneity conditions of Viswanathan et al. (2010). For purposes of illustration, the noisy observations were set to the actual spatial frequencies of the study gratings in the two conditions (i.e., $x_i = \theta_i$). These observations are indicated by solid vertical lines, and the frequency of the lure probe at 2 just-noticeable-difference (JND) units is represented by the dashed line. The solid black and gray curves show the combined posterior densities $p(\theta|\{x_i\}_{i=1}^N)$ in the medium- and high-homogeneity conditions, respectively. Lower panel shows the posterior distributions over the number of clusters in the two conditions. B. Probabilities of responding "old" in the two conditions as a function of probe frequency. C. The observed and predicted probabilities of responding "old" to the probe at 2 JND units in the two conditions. The model prediction was estimated over 1,800 simulated trials.

of probe gratings are represented by dashed vertical lines (at 2 JND in both conditions). In this figure, the probe is a "new" item (or a lure) in both conditions. The experiment was designed so that the individual probe-study item similarities were identical in the two conditions, meaning that the only difference between these conditions was the interitem similarity of the study items, with the high homogeneity condition having a higher interitem similarity than the medium homogeneity condition. The interitem similarity effect refers to the finding that subjects had a significantly higher probability of responding "old" in the medium homogeneity condition than in the high homogeneity condition (mean $P(old) = 0.69$ vs. mean $P(old) = 0.57$).

Since the task in Viswanathan et al. (2010) is an old/new recognition task, we cannot use the mean of the marginal posterior of the target item to simulate the model's responses. Unlike other tasks considered in this section which are recall tasks, there is no single target item in an old/new recognition task. Consequently, we constructed a combined posterior density from the marginal posteriors over each θ_i by marginalizing over the indices of the items and assuming each item was equally likely:

$$p\left(\theta|\{x_i\}_{i=1}^N\right) \propto \Sigma_{j=1}^N p\left(\theta = \theta_j|\{x_i\}_{i=1}^N\right).$$

This density can be roughly thought of as a nonparametric density estimate of the spatial frequencies presented on this trial based on the noisy observations of the frequencies, and it quantifies the posterior density of θ being the value of any one of the study items.

To model the experimental results, $p(\theta|\{x_i\}_{i=1}^N)$ needs to be transformed into a probability of responding "old." We did this by mapping $p(\theta|\{x_i\}_{i=1}^N)$ between 0.2 and 0.8 so that the minimum probability of responding "old" was 0.2 and the maximum probability of responding "old" was 0.8. We did not map the probabilities between 0 and 1 because subjects tend to have relatively high probabilities of responding "old" even for very dissimilar probes in these types of experiments, and relatively low probabilities of responding "old" even for perfect matches between probes and study items. The specific values of 0.2 and 0.8 were chosen based on a similar previous study by Kahana et al. (2007).[1] Finally, the model's response was drawn from a Bernoulli distribution with success probability equal to the probability of responding "old." One thousand eight hundred trials each of medium and high homogeneity conditions were simulated.

We first demonstrate the behavior of a nonoptimized DPMM with the free parameters set to $\alpha_c = 1$ and $\tau_{obs} = 1$. Figure 7A shows the combined posterior densities $p(\theta|\{x_i\}_{i=1}^N)$ in a single representative trial of the medium and high homogeneity conditions (solid black and gray curves, respectively). For purposes of illustration, the noisy observations were set to the actual feature values of the items in these examples (i.e., $x_i = \theta_i$). The nonoptimized DPMM was able to reproduce the interitem similarity effect without fitting its parameters to the observed data. This can be seen in Figure 7A by noting that the black curve, representing the combined posterior density in the medium homogeneity condition intersects the dashed vertical line at a higher point than the gray curve, representing the combined posterior density in the medium homogeneity condition. Figure 7B shows the probabilities of responding "old" in the two conditions as a function of probe frequency, which were obtained simply by normalizing the combined posterior densities shown in Figure 7A between 0.2 and 0.8. Although the nonoptimized DPMM with $\alpha_c = 1$ and $\tau_{obs} = 1$ explained the interitem similarity effect, it did not provide an excellent quantitative fit to the experimentally observed probabilities of "old" responses in the two conditions (Figure 7; $P(old) = 0.68$ vs. $P(old) = 0.49$ in the medium and high homogeneity conditions, respectively, compared with the observed mean probabilities of $P(old) = 0.69$ and $P(old) = 0.57$ for the respective conditions in the actual experiment).

Intuitively, the reason that the DPMM successfully accounts for the interitem similarity effect is that the posterior distributions over the number of clusters have significant masses at one- and two-cluster partitions for both medium and high homogeneity conditions (see the bottom row in Figure 7A). In one-cluster partitions, all items are grouped into a single cluster, and in two-cluster partitions, the leftmost two items are typically grouped into a single cluster and the rightmost item is assigned to its own cluster. Relative to the medium homogeneity condition, the spatial frequencies of items in the high homogeneity condition have a lower variance. Therefore, a single cluster fit to the spatial frequencies of all items in a trial has a lower variance in the high homogeneity condition. Similarly, in two-cluster partitions, the cluster containing the leftmost two items has a lower variance in the high homogeneity condition. This, combined with the fact that the probe at 2 *JND* is closer to the means of these clusters in the medium than in the high homogeneity condition, makes the probe more similar to the study items in the medium homogeneity condition.

Model comparison. We applied the univariate version of the DPMM, BFMMs and the HBM to simulated trials of the medium and high homogeneity conditions, and computed the ML estimates of the free parameters by searching for the parameter values that maximized the log-likelihood given the experimentally observed probabilities of "old" responses in the two conditions under each model (details of the optimization procedure are provided in Appendix B). Models fits were again compared using the BIC measure.

Results. BIC values of the models are given in Table 1. Qualitatively, all four models were successful at capturing the main interitem similarity effect. The BIC values for the DPMM and the BFMM with four components were similar. For these models, the posterior distributions over the number of clusters were dominated by three-cluster partitions. However, one- and two-cluster partitions also had significant probabilities. The BIC values for the HBM and the BFMM with two components indicated a significantly worse fit for these models.

[1] We tried several different ways of transforming $p(\theta|\{x_i\}_{i=1}^N)$ into a probability of responding "old": using different values for the minimum and maximum probabilities of responding "old," normalizing the combined posterior densities separately for the medium and high homogeneity conditions, as well as normalizing them together (i.e., using the same $max(p(\theta|\{x_i\}_{i=1}^N))$ value in normalizing the combined posterior densities in both cases). Although these manipulations in general affected the model's quantitative fit, the ability of the model to qualitatively explain the interitem similarity effect, as well as the relative order of the quantitative fits of different models, were not sensitive to the specific choice of the transformation method.

Encoding at Multiple Levels of Abstraction: Brady and Alvarez (2011)

We now show that the multivariate version of the DPMM accounts for the results of both Experiments 1 and 2 in Brady and Alvarez (2011). Recall from our earlier discussion that in Experiment 1, subjects were presented with blue, red and green circles of different sizes. Subjects were instructed to ignore the green circles. After a brief delay, a comparison circle appeared at the location of a red or blue circle in the original display. Subjects' task was to indicate the size of the original circle that was at that location, referred to as the target circle, by using the mouse to resize the comparison. It was found that the reported size of the target circle was biased toward the average size of the circles having the same color as the target. Experiment 2 was identical to Experiment 1 except that displays did not include green circles. Brady and Alvarez hypothesized that color is an irrelevant feature in this case, and thus subjects would not use a color-based encoding scheme. Consistent with this hypothesis, subjects did not show a bias toward the mean size of the same-colored circles in their size estimates. Instead, subjects' estimates showed a bias toward the mean size of all circles in a display.

Brady and Alvarez (2011) accounted for their results using two-level (to account for the results of Experiment 2) and three-level (to account for the results of Experiment 1) hierarchical models (see Equations 4–5, respectively). Our goal here is to show that the same DPMM explains the results of both experiments in a way that does not require the modeler to stipulate different numbers of levels of abstraction for the encoding of items in the two experiments.

In our simulations, we represented circles as points in a two-dimensional feature space defined by color and size. Following Brady and Alvarez (2011), we assumed that the removal of the green circles in Experiment 2 reduced the salience or weight of the color dimension, thereby shrinking distances along the color dimension in the two-dimensional space. This is illustrated in the top row of Figure 8A where blue and red circles are separated along the color dimension for Experiment 1 (color is a relevant feature) but not for Experiment 2 (color is an irrelevant feature). We used arbitrary numerical values for red and blue (red = 125, blue = 25 in Experiment 1; red = 75, blue = 75 in Experiment 2). Simulations with different numerical values confirmed that our results did not depend on the choice of specific numerical values for the colors. As long as the salience of the color dimension was sufficiently reduced in Experiment 2, we were always able to get similar results, albeit with different parameter values. Sizes of the blue and red circles in each simulated trial were generated in accordance with the procedures described in Brady and Alvarez (2011). Noisy observations of the sizes and colors of the circles were then generated from multivariate Gaussian distributions with mean equal to the true feature values of the circle and covariance matrix $\Sigma_{obs,}$ which we assumed to be a diagonal matrix with equal variances along the diagonal, denoted by σ^2_{obs}. The models were then applied to these noisy observations in each simulated trial.[2] For all models, in each simulated trial, the mean of the marginal posterior of the target item, $\hat{\theta}_t = E[\theta_t|\{x_i\}_{i=1}^N]$, was computed and the size dimension of this estimate was taken to be the model's response in that trial.

Model comparison. As in previous examples, we determined the ML estimates of the parameters of the four models using grid searches (see Appendix B for details) and compared the model fits using the BIC measure. The multivariate DPMM and BFMMs have two free parameters, κ and σ^2_{obs}. The two-level HBM has a single free parameter, σ^2_{obs}.

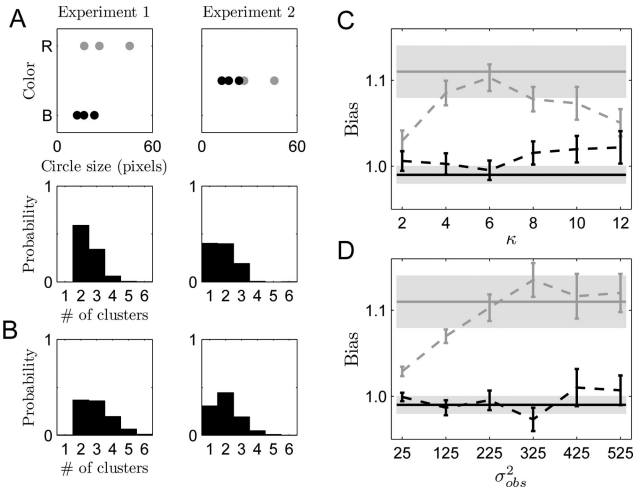

Figure 8. Simulation results for the optimized multivariate Dirichlet process mixture model (DPMM) with κ and σ^2_{obs} set to their maximum-likelihood estimates (κ = 6, σ^2_{obs} = 225). A. Left and right columns show simulation results for representative trials from Experiments 1 and 2, respectively. Top row shows the visual stimuli used in each trial in color-size space. Bottom row shows the posterior distributions over the number of clusters for the corresponding trials. B. Posterior probabilities over the number of clusters averaged over all trials in the two simulated experiments. C. Predicted biases for the DPMM in Experiment 1 (gray) and Experiment 2 (black) of Brady and Alvarez (2011) as a function of κ and σ^2_{obs} (D). Horizontal lines represent the observed mean biases of subjects in these experiments, and gray areas represent ±1 *SEM* around the mean biases. The dashed gray and black lines represent the mean predicted biases for the DPMM in Experiment 1 and in Experiment 2, respectively, and the error bars correspond to ±1 *SEM* around the mean predicted biases over 25 simulations of both experiments, with 50 simulated trials in each experiment. In C, σ^2_{obs} was fixed at 225 and κ was varied, whereas in D, κ was fixed at 6 and σ^2_{obs} was varied.

Results. BIC values of the models are provided in Table 1. The DPMM and BFMMs were able to qualitatively capture the differing pattern of biases observed in Experiment 1 and Experiment 2 of Brady and Alvarez (2011). The DPMM and the BFMM with $K = 4$ components yielded similar BIC scores, whereas the BFMM with $K = 2$ components provided a slightly worse fit (as evidenced by its higher BIC score). This was because the BFMM with $K = 2$ components tended to overestimate the bias toward the mean size of the same-colored circles in Experiment 1, as it almost always favored two-cluster, color-based partitions where the red circles are assigned to one cluster and the blue circles to a separate cluster. The DPMM and the BFMM with four components, on the

[2] To control for potential artifacts, Brady and Alvarez (2011) designed their experiments to use matched pairs of trials. Subjects' biases were computed based on their responses to these pairs. In our simulations, we followed the same procedures. For brevity, we do not explain these procedures here. The interested reader is referred to Brady and Alvarez (2011).

other hand, frequently favored more fine-grained partitions with a larger number of clusters, which had the effect of reducing the magnitude of biases predicted by these models. Predictably, the two-level HBM only generated a bias toward the mean size of all circles in a display and was unable to explain the color-based bias observed in Experiment 1. As discussed earlier, Brady and Alvarez (2011) used a three-level HBM to explain the bias toward the mean size of the same-colored circles observed in Experiment 1 and a two-level HBM to explain the bias toward the mean size of all circles observed in Experiment 2. Our results, however, demonstrate that it is not necessary to postulate hierarchical models with different numbers of levels of abstraction to account for the observed pattern of biases in the two experiments. We applied the same DPMM and BFMMs to both experiments. Despite changes in experimental conditions across experiments, these models succeeded at automatically determining the appropriate scales for representing stimuli in each experiment as well as the weights that should be allocated to each scale.

The left and right columns of Figure 8A show simulation results for the optimized DPMM for a representative trial from Experiment 1 and for the corresponding trial from Experiment 2 (where the sizes of the circles remained the same and only the numerical values associated with the colors changed). Also shown are the posterior distributions over the number of clusters for these two trials. For Experiment 1 in which color is a salient dimension, color-based partitions where red circles are grouped into a single cluster and blue circles into another cluster, are highly probable (note the high probability of two-cluster partitions in the posterior distribution), thereby producing a bias toward the mean size of the same-colored circles. For Experiment 2 in which color is no longer salient, color-based partitions are not probable under the model. Single-cluster partitions where all stimuli are grouped into a single cluster are highly probable, producing a bias toward the overall mean size of all circles, consistent with the results from Experiment 2 of Brady and Alvarez (2011). Although two-cluster partitions are also highly probable, most of the likely two-cluster partitions are size-based, not color-based, partitions. Figure 8B shows the posterior probabilities over the number of clusters averaged over all trials in the two simulated experiments. In Experiment 1, as in the representative trial shown in Figure 8A, two-cluster partitions are the most probable, and the overwhelming majority of these partitions are color-based partitions where blue circles are grouped into one cluster and red circles are grouped into another. In Experiment 2, one-cluster partitions in which all items are grouped together has significant probability (unlike Experiment 1). Again, although two-cluster partitions are also highly probable, most of these two-cluster partitions are size-based. We emphasize that the same values of the parameters κ *and* σ^2_{obs} were used for both experiments. Therefore, the models applied to the two simulated experiments were identical.

To demonstrate that the ability of the DPMM to explain the specific pattern of biases observed in Brady and Alvarez (2011) did not critically depend on the optimization of the free parameters κ and σ^2_{obs} to fit the particular numerical values of these biases observed in their experiments, we varied each one of these two free parameters over a broad range of values while keeping the other parameter fixed and computed the model's predicted biases in Experiment 1 and Experiment 2. Figures 8C and 8D show the predicted biases in Experiment 1 and Experiment 2 as a function of κ and σ^2_{obs}, respectively. Note that with the specific bias measure used in Brady and Alvarez (2011), a bias value of 1 indicates that there is no bias toward the mean size of the same-colored circles, whereas a bias value significantly greater than 1 indicates a bias toward the mean size of the same-colored circles. Figure 8C demonstrates that the ability of the DPMM to qualitatively explain the pattern of biases observed in Brady and Alvarez (2011) is not very sensitive to the choice of κ. Except for very large or very small κ values, the model was able to capture the experimentally observed pattern of biases. The model preferred a larger number of clusters for larger κ values and thus underestimated the bias in Experiment 1. Figure 8D shows the predicted biases as a function of σ^2_{obs} while κ was fixed at 6. For larger values of σ^2_{obs}, the predicted biases were more variable. For small σ^2_{obs} values, on the other hand, the model again tended to underestimate the bias in Experiment 1, as it relied more heavily on the noisy observations and less on the prior induced by the DPMM.

Discussion of Simulation Results

In summary, the simulation results provide evidence in favor of our proposed Probabilistic Clustering Theory (PCT). The DPMM—the exact implementation of PCT—provided good fits, both qualitatively and quantitatively, to the experimental data of Wilken and Ma (2004), Viswanathan et al. (2010), and Brady and Alvarez (2011). Unsurprisingly, the two-level HBM was unable to explain the main result of Brady and Alvarez (2011), and it also yielded significantly worse quantitative fits than other models to the results from Viswanathan et al. (2010). To account for the different pattern of results in their two experiments, Brady and Alvarez (2011) used HBMs with different numbers of levels for the two experiments (three levels for Experiment 1 and two levels for Experiment 2), where the number of levels for each model was specified in advance. In contrast, the DPMM and the BFMMs succeeded at accounting for the different pattern of results in the two experiments without any changes in the structure or parameters of the models across the experiments and without any specification of the appropriate levels of abstraction in advance.

The BFMM—an approximate implementation of PCT—with four components yielded similar results to the DPMM. The BFMM with two components yielded a significantly worse fit than the DPMM and the BFMM with four components in the simulation of the Viswanathan et al. (2010) study, and it overestimated the magnitude of the bias toward the mean size of the same-colored circles in Experiment 1 of Brady and Alvarez (2011). This demonstrates the danger of setting an a priori limit on the number of components and, hence, provides indirect support for a more flexible nonparametric approach such as the DPMM that automatically determines the appropriate number of components in each specific case without assuming an a priori limit on this number. Because of people's tendency to spontaneously group items in a display (Woodman, Vecera, & Luck, 2003; also the studies reviewed in this section) and the relatively small set sizes used in VSTM experiments, a mixture model would typically not need more than four to five components to account for the contents of people's memories for simple multi-item displays, thereby explaining the success of the BFMM with four components in modeling the data considered in this section. However, in some

cases, subjects might not have a strong tendency to group items in VSTM (which corresponds to a large α_c in the DPMM). Requiring that the items be grouped into four components, as the BFMM with four components does, might lead to the prediction of biases and dependencies in the estimates of the feature values of items that do not fit such "nongrouping" subjects' responses well enough. In any case, it is conceptually more appropriate to avoid making unwarranted a priori assumptions about the maximum number of components that would be needed for modeling different subjects' data in VSTM experiments (or equivalently about the grouping tendency of different subjects). Rather, it is preferable to automatically determine this from the data itself.

Experiments

As discussed above, our main goal in this article is to characterize the organization of VSTM in terms of the joint probability distribution of the estimates of the feature values of multiple items,

$$p\left(\left\{\hat{\theta}_i\right\}_{i=1}^{N} \middle| \left\{\theta_i\right\}_{i=1}^{N}\right).$$

This joint distribution allows for the possibility of rich and highly complex dependencies between the feature values of the visual items and their estimates, as well as among the estimates themselves. It also replaces informal notions of dependence or independence in encoding multiple items that are prevalent in the VSTM literature with well-defined notions of statistical dependence or independence. In earlier sections, we showed how various behavioral phenomena in the VSTM literature constrain the range of appropriate forms for this joint distribution. We also proposed a specific theory of the organization of VSTM, probabilistic clustering theory (PCT), which states that the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^{N}|\{\theta_i\}_{i=1}^{N})$ should be characterized in terms of a probability distribution over all possible clusterings or partitions of the items (see *Models*). An advantage of this theory is that it postulates that VSTM represents items at multiple granularities or scales without the need for a prespecified, fixed hierarchy. An implementation of the theory, the DPMM, was shown in the previous section to account for a number of findings in previous research on VSTM.

In this section, we take an empirical approach and try to determine the properties of the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^{N}|\{\theta_i\}_{i=1}^{N})$ directly. We report the results of three novel experiments specifically designed to uncover potential dependencies between the actual feature values of different items and their estimates, as well as dependencies among the estimates themselves. These novel tasks rely on the idea of probing subjects' estimates of the feature values of all presented items in each trial, rather than probing their estimates of the feature value of a single target item, as is customary in standard VSTM tasks. We also compare subjects' experimentally determined joint probability distributions with the joint distributions predicted by the models considered in this article (DPMM, BFMMs, and the two-level HBM).

Experiment 1

We first designed a VSTM recall experiment where subjects were asked to remember the horizontal locations of a number of briefly presented squares. We then asked subjects to report their estimates of the horizontal locations of *all* presented squares. This contrasts with previous approaches where only the feature value of a single target item is probed in each trial. Over trials, this procedure allowed us to uncover potential dependencies between joint estimates of the feature values of different items.

Method.

Procedure. Subjects were seated 57 cm from a CRT monitor with a screen resolution of 1,280 × 1,024 pixels and a refresh rate of 85 Hz. Each trial began with the display of a fixation cross at a random location within an approximately 12° × 16° region of the screen for 1 s. In separate experiments, subjects were then presented with $N = 2$ (Experiment 1A) or $N = 3$ (Experiment 1B) colored squares (1.4° × 1.4°) on uniformly spaced dark and thin horizontal lines for 100 ms (see Figure 9). After a delay interval of 1 s (during which the horizontal lines remained visible, but not the squares), a probe screen was presented. Initially, the probe screen contained only the horizontal lines. Subjects used the computer mouse to indicate their estimate of the horizontal location of each of the colored squares presented on that trial. Unlike other VSTM experiments, in which subjects are asked to report their estimate of the feature value of a single probed item from a display, our task required subjects to indicate the feature values of all displayed items. This procedure allowed us to study the dependencies between VSTM representations of all items. Subjects were allowed to adjust their location estimates as many times as they wished. When they were satisfied with their estimates, they proceeded to the next trial by pressing the space bar. Figure 9 shows the sequence of events on a single trial of the experiment.

We used different combinations of horizontal locations: $\Theta = \{\theta_i\}_{i=1}^{N}$ for the squares. We call each combination Θ a particular "display configuration." To cover a diverse range of possible display configurations, we first defined a regular grid in the N-dimensional configuration space (where each dimension represents the horizontal location of a different square), and then added a small amount of Gaussian jitter ($SD \approx 0.5°$) to each θ_i in each grid node. In Experiment 1A ($N = 2$), the grid was 6 × 6, yielding a total of 36 different

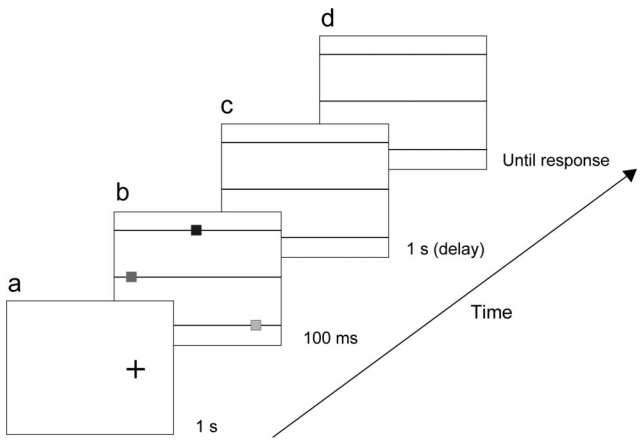

Figure 9. The sequence of events in a single trial of Experiment 1B: (a) a small fixation cross is presented at a random location for 1 s; (b) the target configuration is flashed briefly (100 ms); (c) a delay interval of 1 s follows the target configuration; (d) the probe display (initially containing only the dark horizontal lines) remains on until the subject indicates the horizontal locations of all items using the computer mouse.

configurations. In Experiment 1B ($N = 3$), the grid was $3 \times 3 \times 3$, yielding a total of 27 different display configurations.

To uncover potential dependencies in $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$, we presented the same configuration Θ multiple times, and collected a subject's estimates each time. The estimates of the feature values of all items collected in each presentation of a particular display configuration can be thought of as a single sample from the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ for that particular configuration. In Experiment 1A, each of the 36 display configurations was presented 24 times (yielding a total of 864 trials) with the presentation of different configurations randomly interleaved during the course of the experiment. In Experiment 1B, each of the 27 display configurations was presented 26 times (yielding a total of 702 trials). All subjects participating in the same experiment saw the same set of display configurations. The correlation between $\hat{\theta}_i$ and $\hat{\theta}_j$ for each pair of items *i, j* in each configuration was estimated by calculating the correlation coefficient between the subject's estimates of θ_i and their estimates of θ_j over all presentations of that particular configuration.

Participants. Eight naive subjects participated in Experiment 1A, and 11 naive subjects participated in Experiment 1B. Subjects were undergraduate or graduate students at the University of Rochester. All subjects reported normal or corrected-to-normal vision, and they were compensated at a rate of $10 per hour for their time. Subjects completed the experiment in two sessions.

Model comparison. For each subject, we determined the ML estimates of the free parameters of each of the four models (DPMM, BFMMs with two and four components, and the two-level HBM) by searching for the parameter values that maximized the subject's trial-by-trial responses under the distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ induced by the model. In other words, each setting of the free parameters of a model yields a slightly different distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$. Since these distribution do not have an analytic form, they were computed by sampling, as in previous cases in this article (see Equation 2). The likelihood of the subject's trial-by-trial responses was then computed under these distributions (note that the distributions are dependent on the display configuration $\{\theta_i\}_{i=1}^N$). The parameter values maximizing this likelihood were chosen as the ML estimates of the parameters. The models were then compared using the BIC measure as usual (see Appendix B for additional details).

Results. The joint distributions $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ estimated from subjects' responses displayed notable dependencies between estimates of different items. In particular, there were positive pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ for different items i and j and the magnitude of these correlations decreased with the distance $|\theta_i - \theta_j|$ between the actual feature values of the corresponding items.

The leftmost plot in Figure 10A shows the results for a representative subject (RD) in Experiment 1A. In this figure, the crosses represent the 36 stimulus configurations $\Theta = \{\theta_1,\theta_2\}$ presented to the subject, the dots represent the subject's mean estimates for each configuration [i.e., the mean of the joint distribution $p(\hat{\theta}_1,\hat{\theta}_2,|\theta_1,\theta_2)$ for each configuration (θ_1, θ_2)], and the contours represent the shapes of bivariate Gaussian distributions fitted to the subject's responses for each configuration (i.e., contours of bivariate Gaussian approximations to the joint distributions $p(\hat{\theta}_1,\hat{\theta}_2,|\theta_1,\theta_2)$ for all configurations (θ_1, θ_2); red contours show cases where $\hat{\theta}_1$ and $\hat{\theta}_2$ were significantly and positively correlated, whereas blue contours show cases where $\hat{\theta}_1$ and $\hat{\theta}_2$ were significantly and negatively correlated). For stimulus configurations where the two items have similar horizontal locations, the subject's responses tended to be correlated (note the predominance of red contours near the main diagonal representing $\theta_1 = \theta_2$), whereas for configurations where the two items have dissimilar horizontal locations, this tendency was gradually reduced. This pattern was apparent in most of the subjects. Figure 10B (left) shows, for subject RD, the correlations between $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of the distance $|\theta_1 - \theta_2|$ between the actual horizontal locations of the two items. In this plot, the 36 stimulus configurations (θ_1, θ_2) presented to the subject were divided into 6 equal-length bins based on $|\theta_1 - \theta_2|$, and the mean correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$ as well as the standard error of the mean were calculated for each bin. Figure 11A (left) shows the correlations as a function of $|\theta_1 - \theta_2|$ for combined data from all eight subjects in Experiment 1A. We also performed linear regressions of the correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$ on the distance $|\theta_1 - \theta_2|$ for each subject separately as well as for combined data from all subjects. For seven of the eight subjects in the experiment with $N = 2$ (as well as for combined data from all subjects), the linear regression was significant, i.e., the 95% confidence interval for the slope excluded zero ($p < .05$), and the slope was negative, suggesting that correlations decreased with the distance $|\theta_1 - \theta_2|$.

Another notable pattern in subjects' responses was the bias toward the mean horizontal locations for many of the configurations presented to them (this can be seen in Figure 10A by noting that the mean estimates of the subject are closer to the main diagonal than the actual stimulus configuration for many configurations). This bias is consistent with similar biases previously reported in the literature for other feature dimensions as reviewed earlier in this article (see the subsection titled *Biases in VSTM*).

Results for Experiment 1B ($N = 3$) were qualitatively similar. There were again positive pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ for different items i and j, and the magnitude of these correlations decreased with the distance between the actual horizontal locations of the corresponding items. Figure 11A (right) shows the pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ as a function of the distance between the actual horizontal locations of the two items for combined data from all 11 subjects in Experiment 1B. In this figure, all $\{\theta_i, \theta_j\}$ pairs for all stimulus configurations ($27 \times 3 = 81$ pairs in total) were divided into three equal-length bins based on the distance $|\theta_i - \theta_j|$ between the horizontal locations of items i and j. As in the analysis of data from Experiment 1A, we also performed linear regressions of the correlation between $\hat{\theta}_i$ and $\hat{\theta}_j$ on the distance $|\theta_i - \theta_j|$ for each subject separately as well as for combined data from all subjects. For all 11 subjects in the experiment with $N = 3$ (as well as for combined data from all subjects), the linear regression was significant ($p < .05$), and the slope was negative, suggesting that correlations decreased with the distance $|\theta_i - \theta_j|$.

Overall, subjects exhibited a smaller number of significant $\{\hat{\theta}_1,\hat{\theta}_3\}$ correlations than $\{\hat{\theta}_1,\hat{\theta}_2\}$, or $\{\hat{\theta}_2,\hat{\theta}_3\}$ correlations. This is probably due to the fact that the $\{\theta_1, \theta_3\}$ pair had a larger vertical distance than the other pairs. We also carried out a

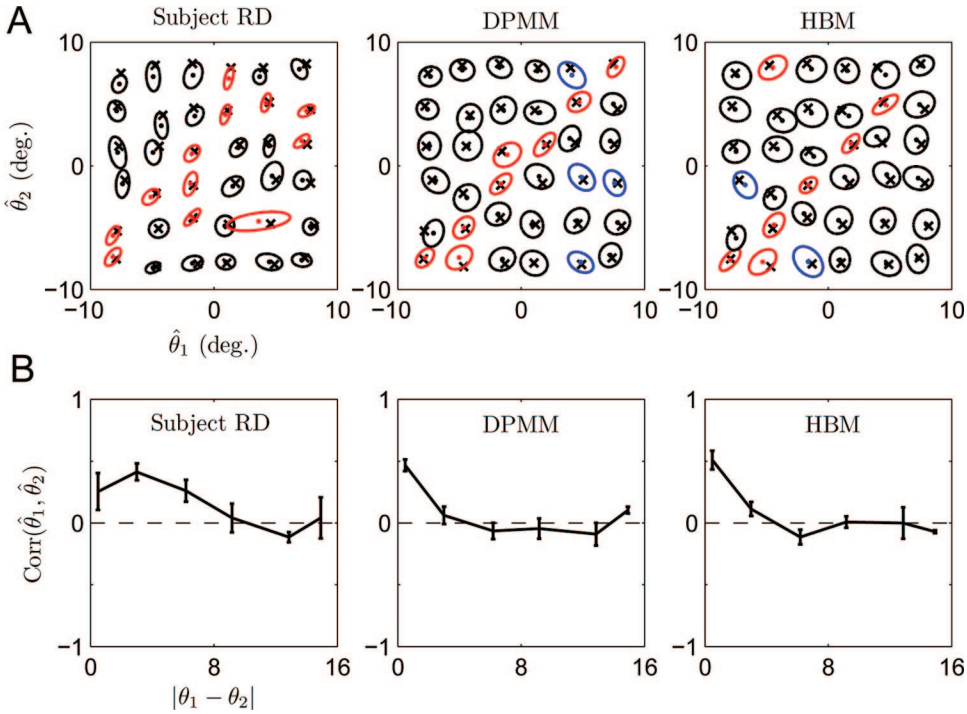

Figure 10. A. Results from a representative subject (RD) in Experiment 1A (left). Crosses represent the 36 stimulus configurations presented to the subject, dots represent the subject's mean estimates for each configuration and contours represent single-level contours of bivariate Gaussian distributions fitted to the subject's responses for each configuration (red contours show cases where the subject's estimates of the horizontal locations of the two presented items were significantly and positively correlated, whereas blue contours show cases where the subject's estimates were significantly and negatively correlated). The middle and the right panels show the predictions of the best fitting Dirichlet process mixture model (DPMM) and the best fitting hierarchical Bayesian model (HBM), respectively (colors and line styles same as in subject's data). B. Correlations between $\hat{\theta}_1$ and $\hat{\theta}_2$ as a function of the distance $|\theta_1 - \theta_2|$ for subject RD (left) and the corresponding predictions from the best fitting DPMM (middle) and the best fitting HBM (right). Error bars represent standard errors of the mean.

similar analysis using the two-dimensional locations of the items (considering both horizontal and vertical locations of the items in computing the Euclidean distance between two items rather than considering only their horizontal locations). This analysis yielded qualitatively similar results. As in Experiment 1A, subjects also displayed a bias toward the mean horizontal locations in their estimates.

We also looked at the standard deviation of subjects' estimates, $\mathrm{Std}(\hat{\theta}_i)$, as a function of the standard deviation of the actual feature values of items in a configuration, $\mathrm{Std}(\{\theta_i\}_{i=1}^N)$ (Figure 11B). This function had an inverse U shape both in Experiment 1A and in Experiment 1B. Subjects' estimates had small standard deviations for configurations with homogeneous or low-variance feature values. $\mathrm{Std}(\hat{\theta}_i)$ increased for configurations with more heterogeneous or high-variance feature values but decreased again for the most heterogeneous or highest-variance configurations. We note that a similar increase in the variance of encoding for individual items with increased heterogeneity of the actual feature values of items has been recently reported in Sims, Jacobs, and Knill (2012). The decrease in $\mathrm{Std}(\hat{\theta}_i)$ for the most heterogeneous configurations is attributable to edge effects, because the most heterogeneous configurations (i.e., configurations for which $\mathrm{Std}(\{\theta_i\}_{i=1}^N)$ has the highest value) are configurations where one of the items is close to the left or the right edge of the screen and the other item or items are close to the opposite edge. For these configurations, subjects' estimates had low variance, because their estimates were constrained on one side by the edge of the screen.

Interestingly, the DPMM and, to a lesser extent, the other models were able to qualitatively explain this inverse U-shaped relationship between $\mathrm{Std}(\{\theta_i\}_{i=1}^N)$ and $\mathrm{Std}(\hat{\theta}_i)$. Intuitively, this can be understood as follows. By grouping items together, the DPMM reduces the variance of encoding for individual items, because information about an individual item can be gained from other items in the same group. This reduction in the variance of encoding for individual items is largest for the most homogeneous configurations, because the group has the lowest possible variance in such configurations, increasing the amount of information one can get about an individual item from other items in the same group. For more heterogeneous groups, one can get less information about an individual item from other items. For the most heterogeneous groups, since we used a uniform distribution over the horizontal length of the screen as the base distribution over μ, the edge effect mentioned above acts to reduce the variance in the model's estimates. We note

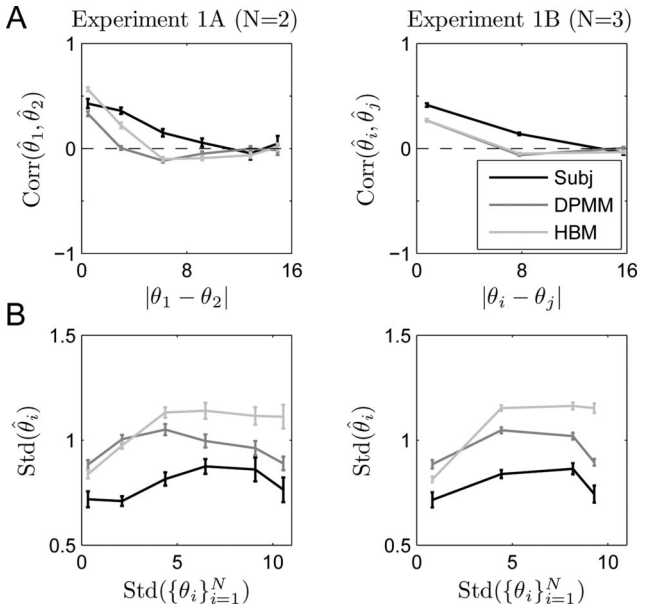

Figure 11. A. Pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ as a function of the distance $|\theta_i - \theta_j|$ between the horizontal locations of pairs of items for combined data from all eight subjects in Experiment 1A (left) and for combined data from all 11 subjects in Experiment 1B (right). (B) Standard deviations of $\hat{\theta}_i$ as a function of the standard deviation of the actual feature values of items for combined data from all eight subjects in Experiment 1A (left) and for combined data from all 11 subjects in Experiment 1B (right). Also shown are the predictions of the best fitting Dirichlet process mixture models (DPMMs) and the best fitting hierarchical Bayesian models (HBMs). Error bars represent standard errors of the mean.

that the models were only fit to the subjects' trial-by-trial responses and were not optimized to display this inverse U-shaped relationship between $\text{Std}(\{\theta_i\}_{i=1}^N)$ and $\text{Std}(\hat{\theta}_i)$.

The models overestimated the standard deviations of subjects' estimates (Figure 11B). This is because subjects' estimates displayed significant biases. To be able to explain these biases, τ_{obs} had to be sufficiently small. But such small τ_{obs} values led to feature estimates with higher variances than observed in the data. Larger τ_{obs} values led to relatively unbiased estimates and did not fit subjects' estimates well enough. On the other hand, very small τ_{obs} values led to estimates with much higher variances than observed in the data and did not fit subjects' estimates well either.

How did the models do in explaining the biases and dependencies in subjects' estimates? Qualitatively, the models were able to capture the major aspects of the biases and dependencies observed in the empirical joint distributions $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ determined from subjects' responses. In particular, as shown in Figures 10A and 10B for the DPMM (middle column) and the HBM (right column), respectively, the models were able to generate correlated estimates for the horizontal locations of different items when their horizontal locations were very similar. However, a common failure of all four models was that they underestimated the magnitude of pairwise dependencies between estimates of different items. The correlations predicted by the models decayed faster than the actual correlations observed in subjects' responses. Note that the models were fit to the trial-by-trial responses of the subjects and not directly to the observed pattern of correlations or biases.

The models were, in principle, capable of generating correlations that resembled the correlations observed in subjects' responses (i.e., correlations that decayed more slowly). However, this required using fewer clusters and/or noisier memory observations in the case of the DPMM (a smaller α_c and/or a smaller τ_{obs}) or noisier memory observations in the case of the HBM (a smaller τ_{obs}) than warranted by subjects' trial-by-trial responses.

In a similar vein, the models also underestimated the magnitude of subjects' biases. Again, to be able to generate biases that matched subjects' biases in their magnitude, the models had to use noisier representations than warranted by subjects' trial-by-trial responses.

In Experiment 1A ($N = 2$), the DPMM had the best BIC score for six out of eight subjects and the HBM had the best BIC score for the remaining two subjects. The HBM tended to provide better fits for subjects with broader error distributions. However, the differences between the BIC scores of the four models were not very large and all four models made qualitatively similar predictions. In Experiment 1B ($N = 3$), the DPMM had the best BIC score for seven out of 11 subjects, the BFMM with four components had the best BIC score for two subjects, and the HBM had the best BIC score for the remaining two subjects. Again, the differences in the BIC scores of these three models were not very large. In contrast, the BFMM with two components provided consistently worse fits than the other models. The reason for this is illustrated in Figure 12, which shows the posterior distributions over the number of clusters inferred by the DPMM and the BFMM with two and four components for two representative subjects in Experiments 1A and 1B, respectively. The posterior distributions for the DPMM and the BFMM with four components give substantial probability to three-cluster partitions (where each item is represented individually with no grouping). Because the BFMM with two components could use, at most, two clusters, it was not able to account for these "nongrouping" partitions, and consequently could not fit subjects' responses well. A similar problem arises in the HBM, but in BIC calculations, the HBM benefits from having one less parameter than the other models.

Discussion. Our main contribution in Experiment 1 was that we designed a novel experimental task allowing us to empirically determine the form of the joint probability distribution

$$p\left(\left\{\hat{\theta}_i\right\}_{i=1}^N \middle| \left\{\theta_i\right\}_{i=1}^N\right)$$

characterizing a subject's estimates of the feature values of multiple items based on VSTM. The basic idea in our task is to ask subjects to report their estimates of the feature values (horizontal locations) of all presented items in each trial, rather than asking them to report the feature value of a single target item as is customary in more standard VSTM tasks. The estimates of the feature values of multiple items collected in each trial can be thought of as a single sample from the joint distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$. Over several presentations of the same configuration, this procedure allows us to determine the dependencies between the estimates of the feature values of different items.

Using this novel paradigm with horizontal location as the relevant feature dimension, we found a hitherto unrecognized form of dependence between estimates of the feature values of different items, namely, the existence of positive pairwise correlations be-

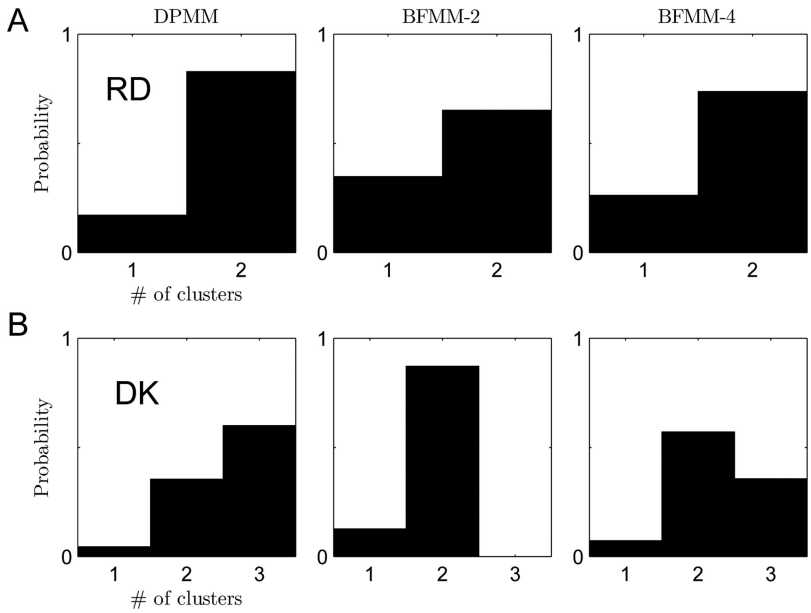

Figure 12. Posterior distributions over the number of clusters inferred by the Dirichlet process mixture model (DPMM) and the Bayesian finite mixture model (BFMM) with two and four components for two representative subjects in Experiment 1A (subject RD; A) and in Experiment 1B (subject DK; B), respectively.

tween $\hat{\theta}_i$ and $\hat{\theta}_j$ for different items *i, j* that decrease with the distance $|\theta_i - \theta_j|$ between the feature values of the corresponding items. In addition, consistent with previously reported biases in VSTM for spatial frequency (Wilken & Ma, 2004; Huang & Sekuler, 2010) and size (Brady & Alvarez, 2011), we found biases toward mean horizontal locations in subjects' estimates.

Among the four models we consider in this article, the Bayesian finite mixture model with two components (BFMM-2) can be ruled out because it yielded poor fits in Experiment 1B due to its inability to account for three-cluster partitions where each item is assigned to its own cluster. This shows the importance of not setting an a priori limit on the number of components, and hence favors a flexible nonparametric approach such as the DPMM that automatically infers the granularity of groups or clusters that is appropriate for each subject in an experiment. In our experiments, we used at most three items, therefore in a sense, one does not need more than three components for modeling the representations of the items. Consequently, it could be argued that the Bayesian finite mixture model with four components (BFMM-4) can be safely used in place of the nonparametric DPMM. However, in experiments that use larger set sizes, some subjects might not have a strong tendency to group items in VSTM. In such cases, requiring that the items be grouped into four components might lead to the overestimation of biases and/or dependencies in such "nongrouping" subjects' responses. Therefore, it would be safest to avoid making any a priori assumptions about the maximum number of components that would be needed to characterize different subjects' behaviors in a VSTM experiment, or equivalently about the grouping tendency of different subjects. This is the approach taken by the nonparametric DPMM.

Although the remaining three models (DPMM, BFMM-4 and HBM) were able to generate pairwise correlations between $\hat{\theta}_i$ and $\hat{\theta}_j$ for different items *i* and *j*, as observed in our subjects' responses, the correlations predicted by these models decayed significantly faster than the observed correlations. Similarly, these models also underestimated the magnitude of the biases in subjects' responses. These discrepancies suggest that further modeling efforts are needed to better capture these aspects of the experimental data.

Experiment 2

A potential concern about Experiment 1 is that subjects were asked to indicate their responses using the computer mouse, and thus the experimental task had a significant motor component. Therefore, the correlations observed between the estimates of the horizontal locations of different items might have been caused by subjects' hand movements, rather than by their VSTM representations of the items. Consider, for instance, a hypothetical trial where two squares have similar horizontal locations. Suppose that when the probe screen comes up, a subject first indicates his or her estimate *h* of the horizontal location of the upper square using the computer mouse. To minimize his or her vertical hand movement, the subject might simply move the cursor with a straight downward movement and indicate a horizontal location very similar to *h* for the lower square. This minimum vertical movement strategy would generate correlated estimates for the horizontal locations of different squares when their actual horizontal locations are similar, exactly as observed in Experiment 1.

To rule out this possibility, we designed a change-detection variant of Experiment 1 that minimizes the motor component of the task. The basic idea behind this change-detection task is schematically illustrated in Figure 13. Figure 13A shows a target configuration (i.e., a specific setting of the feature values of two

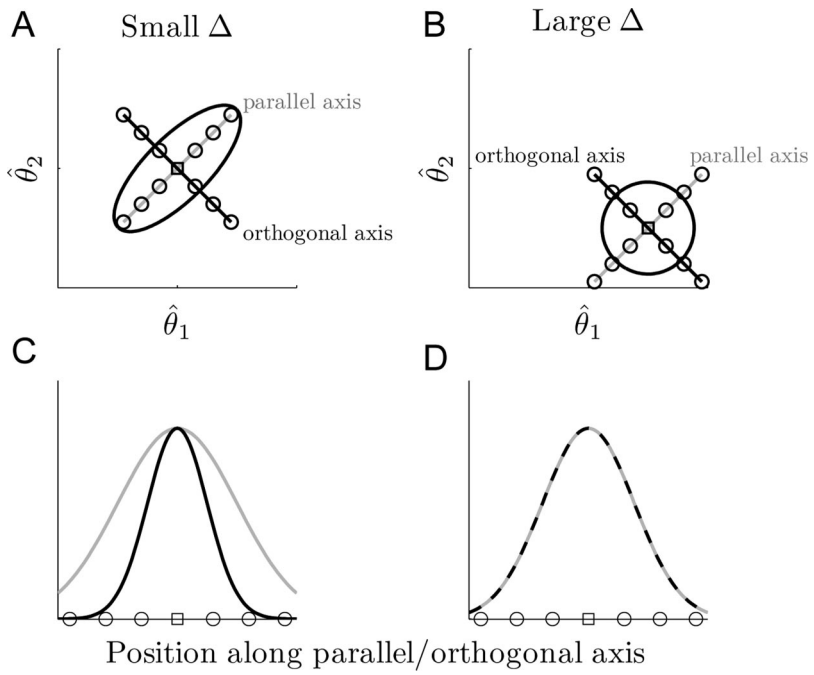

Figure 13. A and B. Target and lure configurations are represented by open squares and circles, respectively. Lure configurations are defined along two perpendicular axes (parallel and orthogonal axes, represented by the gray and black lines, respectively). The contour lines represent the shapes of the joint distributions $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1,\theta_2)$ describing the contents of a subject's memory of the target configurations. In A, items in the target configuration have similar feature values (small Δ), whereas in B, they have different feature values (large Δ). C and D. Unnormalized cross-sections of the joint distributions along the parallel (gray line) and orthogonal axes (black line). If the joint distribution is positively correlated for small Δ values, but uncorrelated for large Δ values, then for small Δ, lure configurations along the parallel axis should have higher probability under the joint distribution than equidistant lure configurations along the orthogonal axis (C). In contrast, when Δ is large, lure configurations along the parallel axis and equidistant configurations along the orthogonal axis should have more or less the same probability under the joint distribution (D).

items) represented by the black square and a number of "lure" configurations represented by the black circles. The lure configurations are equally spaced along two axes referred to as the parallel axis (because it is parallel to the main diagonal) represented by the gray line, and the orthogonal axis (because it is orthogonal to the main diagonal) represented by the black line. In Figure 13A, the two items in the target configuration have similar feature values (the black square representing the target configuration is close to the main diagonal), and therefore the difference

$$\Delta = \theta_1 - \theta_2$$

between the feature values of the two items in the target configuration is small. Figure 13B is similar to Figure 13A, except that the two items in the target configuration in Figure 13B have dissimilar feature values, and therefore Δ is large. The ellipses in Figures 13A and 13B indicate the predicted shapes of the distributions $p(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2)$ characterizing subjects' estimates of the feature values of the two items in the target configurations (these can be compared to the contours in Figure 10). Results from Experiment 1 suggest that $\hat{\theta}_1$ and $\hat{\theta}_2$ are positively correlated when $\Delta = \theta_1 - \theta_2$ is small (as in Figure 13A) and that this correlation decreases with Δ and ultimately vanishes when Δ becomes too large (as in Figure 13B).

If this is indeed the case, then when Δ is small, the lure configurations along the parallel axis should be judged to be more "similar" to, and therefore be more confusable with, the target configuration than equidistant (using the Euclidean metric) lure configurations along the orthogonal axis. This is illustrated in Figure 13C where the (unnormalized) cross-sections of $p(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2)$ along the parallel and orthogonal axes are shown by gray and black lines, respectively. The lure configurations along the parallel axis have higher probability under $p(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2)$ than equidistant lure configurations along the orthogonal axis. Therefore, in a same/different change-detection task, when the target configuration is presented successively with a lure configuration, the lure configurations along the parallel axis should elicit a higher probability of "same" responses than equidistant lure configurations along the orthogonal axis.

In contrast, when Δ is large, representations $\hat{\theta}_1$ and $\hat{\theta}_2$ should be less correlated, and ultimately uncorrelated as Δ becomes sufficiently large (as in Figure 13C). In this case, lure configurations along the parallel axis and equidistant configurations along the orthogonal axis should have more or less the same probability under $p(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2)$, as shown in Figure 13D. Thus, they should be equally confusable with the target configuration, and therefore

elicit more or less the same proportion of "same" responses when successively presented with the target configuration in a change-detection task. In Experiment 2, we tested these predictions using a simple change-detection task.

Method.

Procedure. Subjects were seated 57 cm from a CRT monitor with a screen resolution of 1,280 × 1,024 pixels and a refresh rate of 85 Hz. Subjects' heads were stabilized with the help of a chin rest. The sequence of events in a single trial of the experiment is shown in Figure 14. In each trial, a small fixation cross was presented at the center of the display for 1 s. As in Experiment 1A, subjects were then presented with a target configuration consisting of two differently colored squares (1.4° × 1.4°) placed on two dark and thin horizontal lines (spaced approximately 10° apart) for 100 ms. Subjects were asked to remember the horizontal locations of the two squares in the target configuration. We used five different target configurations with $\Delta = -5.6°, -2.8°, 0°, 2.8°, 5.6°$. Target configurations were presented at random on each trial either completely on the right side or completely on the left side of the fixation cross (as in Bays & Husain, 2008). Therefore, the stimuli were always presented in the visual periphery. A small amount of jitter (uniform between −0.7° and 0.7°) was added to the horizontal locations of both squares in each trial so that the subjects never saw the exact same target configuration twice (the same jitter was used for both squares, so that the difference Δ remained the same).

After a 1-s delay interval (during which only the horizontal lines remained visible), a probe configuration was presented that remained on the display until the subject responded. On half of the trials, the probe configuration was exactly the same as the target configuration. On the other half of the trials, the probe configuration was a lure configuration. The lure configurations were generated as schematically illustrated in Figure 13. Specifically, we first defined two orthogonal axes passing through the target configuration in the two-dimensional space of (θ_1, θ_2) where θ_1 is the horizontal location of the first square and θ_2 is the horizontal location of the second square. The parallel axis is defined to be the line passing through the target configuration that is parallel to the main diagonal (the line $\theta_1 = \theta_2$) and the orthogonal axis is the line passing through the target configuration that is orthogonal to the main diagonal. Six lure configurations were generated on each of these two axes (these are represented by the open circles in Figures 13A and 13B and the target configuration is represented by the open square). On the parallel axis, the horizontal locations of the squares in the six lure configurations differed from the horizontal locations of the squares in the target configuration by (−1.32°, −1.32°), (−0.88°, −0.88°), (−0.44°, −0.44°), (0.44°, 0.44°), (0.88°, 0.88°), or (1.32°, 1.32°). As a matter of notation, these six lure configurations along the parallel axis are given the numeric labels −3, −2, −1, 1, 2, 3, respectively (the target configuration was given the label 0). On the orthogonal axis, the horizontal locations of the squares in the six lure configurations differed from the horizontal locations of the squares in the target configuration by (−1.32°, 1.32°), (−0.88°, 0.88°), (−0.44°, 0.44°), (0.44°, −0.44°), (0.88°, −0.88°), or (1.32°, −1.32°). Similarly, these six lure configurations along the orthogonal axis are given the numeric labels −3, −2, −1, 1, 2, 3, so that the lure configurations along the parallel and orthogonal axes with the same numeric label are equidistant to the target configuration.

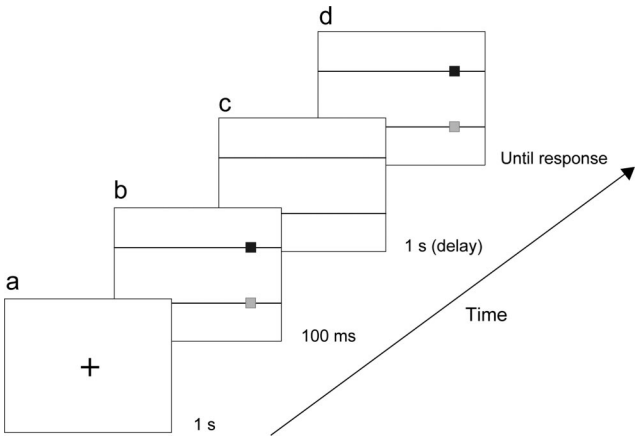

Figure 14. The sequence of events in a single trial of Experiment 2: (a) a small fixation cross is presented at the center of the screen for 1 s, (b) the target configuration is flashed briefly, (c) a delay interval of 1 s follows the target configuration, (d) the probe configuration is presented until the subject responds either "same" or "different" (in this example, the probe configuration is the same as the target configuration).

Subjects were asked to remember the horizontal locations of the squares in the target configuration and to detect any changes in these horizontal locations when the probe configuration came on. Subjects responded by pressing one of two designated keys: "f" for "same" (or no change) and "j" for "different" (or change). Subjects were given auditory feedback after each trial. In addition, written feedback was presented on the screen after every 30 trials. Each subject completed a total of 600 trials (120 trials for each of the five target configurations; half of these 120 trials were "same" trials and the remaining 60 "lure" trials were equally divided between the 12 lure configurations).

Participants. Twenty naive subjects participated in the experiment. Subjects were undergraduate students at the University of Rochester. All subjects reported normal or corrected-to-normal vision, and they were compensated at a rate of $10 per hour for their time. Subjects completed the experiment in a single session.

Results. The average accuracy in Experiment 2 was 68.7% correct. Figure 15 shows the probability of "same" responses as a function of the position along the parallel/orthogonal axis (recall that "0" denotes the target configuration itself, i.e., no-change trials, whereas the other numeric labels represent different lure configurations) for the five different target configurations with different $\Delta = \theta_1 - \theta_2$ values (the curves shown in the figure were called "mnemometric functions' in Zhou, Kahana, & Sekuler, 2004). When the magnitude of Δ was small, the proportion of "same" responses to a lure configuration along the parallel axis was higher than the proportion of "same" responses to an equidistant lure configuration along the orthogonal axis. This difference between the proportion of "same" responses to lure configurations along the parallel axis versus the proportion of "same" responses to equidistant lure configurations along the orthogonal axis was largest for the smallest Δ (Δ = 0°) and became smaller and smaller as the magnitude of Δ increased. As argued earlier in this subsection, this pattern of results is exactly as one would predict if VSTM representations of the two items in the target configurations were correlated and the correlations decreased with the distance be-

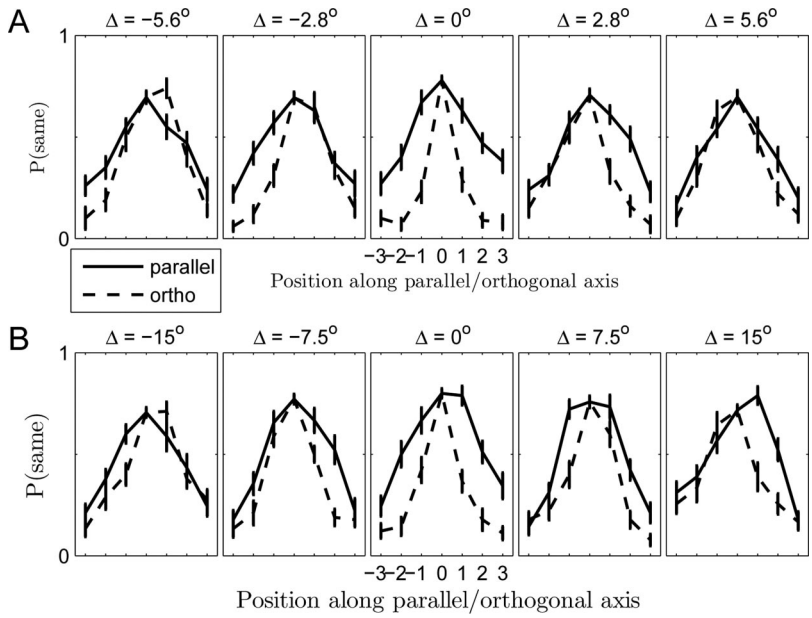

Figure 15. Probability of "same" responses as a function of the position along the parallel/orthogonal axis for the five different target configurations with different Δ values in Experiment 2 (location change detection; A) and Experiment 3 (orientation change detection; B). "0" indicates the target configuration itself (i.e., no-change trials), whereas the other numeric labels represent different lure configurations along the parallel (solid line) or the orthogonal axis (dashed line). Error bars represent standard errors of the mean across subjects.

tween the horizontal locations of the items (compare, e.g., Figure 13 with Figure 15A).

To quantify these correlations and to be able to compare the results of Experiment 2 with the results of Experiment 1 more directly, we developed a Bayesian model of subjects' same/different responses in the change-detection task of Experiment 2. This Bayesian model can be considered to be a two-dimensional generalization of the Bayesian psychometric functions introduced in Kuss, Jäkel, and Wichmann (2005). The model assumes a bivariate Gaussian distribution with mean $\mu = [\mu_1, \mu_2]$ and covariance matrix $\sum = [\sigma_1^2, \rho\sigma_1\sigma_2; \rho\sigma_1\sigma_2, \sigma_2^2]$ for the shape of the joint probability distribution $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ for a given target configuration, (θ_1, θ_2). In the one-dimensional case, this is similar to using a Gaussian cumulative distribution function for modeling psychometric functions. When a subject is presented with a probe configuration (l_1, l_2), he or she makes a same/different response based on the probability of the probe configuration under $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ [i.e., $p(\hat{\theta}_1 = l_1, \hat{\theta}_2 = l_2|\theta_1,\theta_2)$]. Intuitively, probe configurations with high probability under $p(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2)$ are likely to elicit more "same" responses or, in other words, they are more likely to be confused with the target configuration. Mathematically, we handled the mapping from $p(\hat{\theta}_1 = l_1,\theta_2 = l_2|\theta_1,\theta_2)$ to probabilities of same/different responses as follows. We linearly mapped the probabilities of probe configurations relative to the maximum attainable probability (i.e., $\hat{p}(\hat{\theta}_1 = l_1, \hat{\theta}_2 = l_2|\theta_1, \theta_2) = \frac{p(\hat{\theta}_1=l_1,\hat{\theta}_2=l_2|\theta_1,\theta_2)}{max(p(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2))}$; note that for a Gaussian distribution, this is equal to $\exp(-0.5*(l-\mu)^T\Sigma^{-1}(l-\mu))$, which is the unnormalized exponential part of the definition of the Gaussian density) to go from a lower bound b_l, which represents the minimum probability of "same" responses to an upper bound b_u, which represents the maximum probability of "same" responses (we performed Bayesian inference over these variables based on individual subjects' responses):

$$p_l = (b_u - b_l) * \hat{p}(\hat{\theta}_1 = l_1, \hat{\theta}_2 = l_2 | \theta_1, \theta_2) + b_l. \quad (22)$$

This is a reasonable choice because, as is evident from Figure 15, subjects do not respond "same" with a probability of 1, even for the target configuration itself, and do not respond "same" with a probability of 0, even for very dissimilar lure configurations. A "same" response in a given trial is modeled as a Bernoulli distributed variable with a success probability of p_l. Finally, all variables of interest in the model (μ_1, μ_2, ρ, σ_1, σ_2, b_l, b_u) are given suitable priors and their posteriors were computed via MCMC sampling based on individual subjects' same/different responses (additional details are provided in Appendix C). We note that this Bayesian analysis was performed for each subject separately.

We are specifically interested in the posterior distribution of the variable ρ for different target configurations because it represents the correlation coefficient of the underlying joint probability model $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1, \theta_2)$ that characterizes a subject's estimates of the feature values of the items in the target configuration. Given the posterior distribution of ρ, we computed its maximum a posteriori (MAP) estimate for each target configuration.

Figure 16B shows the means and the standard errors (across subjects) of the MAP estimates of ρ for the five target configurations. Confirming the qualitative observations we made from Fig-

ure 15A, it shows that for target configurations with a small |Δ|, ρ is high, and it gradually decreases for target configurations with larger |Δ|. The linear regression of the MAP estimates of ρ on |Δ| was significant, and the slope was negative, suggesting that ρ decreased with |Δ|($b = -0.08$, $t(98) = -6.55$, $p < .01$).

Figure 16A shows, for a representative subject in Experiment 2 (Subject MG), the means and shapes of the underlying distributions $p(\hat{\theta}_1,\hat{\theta}_2|\theta_1,\theta_2)$ for the five target configurations, with the parameters of these distributions set to their MAP estimates. The five target configurations used in Experiment 2 are represented by the black crosses.

Although estimates of the correlations and variances of the joint distributions obtained in Experiment 2 are roughly comparable to estimates of the corresponding variables in Experiment 1, biases toward the mean horizontal locations in Experiment 2, although consistent with the direction of the biases in Experiment 1, were, in general, weaker than the biases obtained in Experiment 1. In this regard, it is interesting to note that most previous reports of biases toward mean feature values in VSTM used recall tasks rather than change-detection or recognition type tasks (Wilken & Ma, 2004; Huang & Sekuler, 2010; Brady & Alvarez, 2011). Change-detection experiments with more target configurations would be needed to determine whether recall tasks (as in Experiment 1) induce additional biases over and above those observed in change-detection tasks (as in Experiment 2).

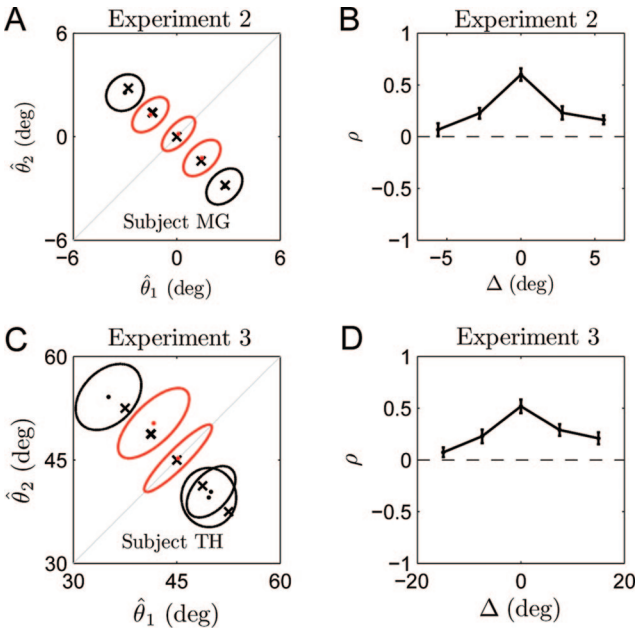

Figure 16. A. Results for a representative subject (Subject MG): For the five target configurations used in Experiment 2, the means and shapes of the distributions $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1,\theta_2)$ with the parameters of these distributions set to their maximum a posteriori (MAP) estimates. The five target configurations used in Experiment 2 are represented by the black crosses. Red contours represent cases where the 95% credible interval for ρ excludes 0. B. The means and the standard errors (across subjects) of the MAP estimates of ρ for the five target configurations in Experiment 2. C and D. Similar to A and B, but results are shown for Experiment 3. In C, results are shown for a representative subject in Experiment 3 (Subject TH). For better visualization, a different scaling was chosen to create the ellipses in C than in A. deg = degrees.

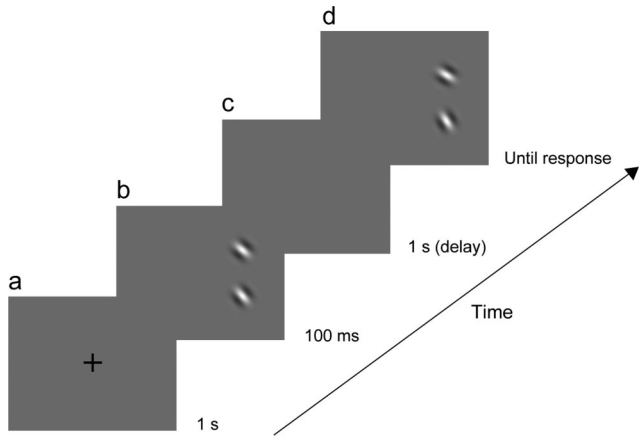

Figure 17. The sequence of events in a single trial of Experiment 3: (a) a small fixation cross is presented at the center of the screen for 1 s; (b) the target display, consisting of two oriented Gabor gratings, is flashed briefly; (c) a delay interval of 1 s follows the target configuration; (d) the probe display is presented until the subject responds either "same" or "different" (in this example, the probe display is different from the target display).

Discussion. Using a change-detection task, Experiment 2 replicated the main experimental finding of Experiment 1, namely, the existence of pairwise correlations between estimates of the horizontal locations of different items that decrease with the distance between the actual horizontal locations of the corresponding items. Because the change-detection task used in Experiment 2 has a minimal motor component, the results of Experiment 2 rule out possible motor explanations of the observed correlations.

Furthermore, because subjects make a single decision in each trial of Experiment 2, these results also rule out possible explanations of the observed correlations in terms of sequential reporting of the estimates of the feature values of items (the recall task used in Experiment 1 involved sequential reporting).

Experiment 3

Experiments 1 and 2 used horizontal location as the relevant feature dimension to be remembered. An interesting question is to what extent the results obtained in these experiments generalize to other feature dimensions. In particular, do the specific pattern of correlations observed between the representations of different items in VSTM extend to feature dimensions other than horizontal location as the feature-independent nature of our models would suggest? To address this question, Experiment 3 used orientations of Gabor gratings instead of horizontal locations of squares as the stimulus feature to be remembered. Otherwise, the experimental design was the same as in Experiment 2.

Method.

Procedure. Subjects were seated 57 cm from a CRT monitor with a screen resolution of 1,280 × 1,024 pixels and a refresh rate of 85 Hz. Subjects' heads were stabilized with the help of a chin rest. The sequence of events in a single trial of Experiment 3 is shown in Figure 17. In each trial, a small fixation cross was presented at the center of the display for 1 s. Subjects were then

presented with a target configuration consisting of two oriented Gabor gratings for 100 ms. The standard deviation of the isotropic Gaussian envelope of the gratings was 0.8°. The spatial frequency of the gratings was 0.9 cycles/degree. The two gratings were vertically separated by a distance of about 6° (center-to-center), and their horizontal locations were the same. Subjects were asked to remember the orientations of the gratings in the target configuration. We used five different target configurations with $\Delta = -15°, -7.5°, 0°, 7.5°, 15°$. Specifically, the target configurations used were (37.5°, 52.5°), (41.25°, 48.75°), (45°, 45°), (48.75°, 41.25°) and (37.5°, 52.5°). As in Experiment 2, a moderate amount of jitter (uniform between −30° and 30°) was added to the orientations of both gratings in each trial so that the subjects never saw the exact same target configuration twice. Again, the same jitter was used for both gratings, so that the difference Δ remained the same. On different trials, target configurations were presented either on the right side of the fixation cross or on the left side of the fixation cross, thus the stimuli were always viewed peripherally. The presentation side was determined randomly in each trial.

After a 1-s delay interval, a probe configuration was presented that remained on the screen until the subject responded. On half of the trials, the probe configuration was exactly the same as the target configuration. On the other half of the trials, the probe configuration was a lure configuration. The lure configurations were generated as in Experiment 2. For a given target configuration, six lure configurations each were generated on the parallel and orthogonal axes. On the parallel axis, orientations of the gratings in the six lure configurations differed from the orientations of the gratings in the target configuration by (−18°, −18°), (−12°, −12°), (−6°, −6°), (6°, 6°), (12°, 12°), or (18°, 18°). These six lure configurations along the parallel axis were given the numeric labels −3, −2, −1, 1, 2, 3, respectively (the target configuration itself was given the label 0). On the orthogonal axis, orientations of the gratings in the six lure configurations differed from the orientations of the gratings in the target configuration by (−18°, 18°), (−12°, 12°), (−6°, 6°), (6°, −6°), (12°, −12°), or (18°, −18°). Similarly, these six lure configurations along the orthogonal axis were also given the numeric labels −3, −2, −1, 1, 2, 3, respectively, so that the configurations along the parallel and orthogonal axes with the same numeric label are equidistant to the target configuration.

Subjects were asked to remember the orientations of the gratings in the target configuration and detect any changes in these orientations when the probe configuration came on. Subjects responded by pressing one of two designated keys: "f" for "same" (or no change) and "j" for "different" (or change). Subjects were given auditory feedback after each trial. In addition, written feedback was presented on the screen after every 30 trials. Each subject completed a total of 600 trials (120 trials for each of the four target configurations; half of these 120 trials were "same" trials and the remaining 60 "lure" trials were equally divided between the 12 lure configurations).

Participants. Eighteen naive subjects participated in the experiment. Subjects were undergraduate students at the University of Rochester. All subjects reported normal or corrected-to-normal vision, and they were compensated at a rate of $10 per hour for their time. Subjects completed the experiment in a single session.

Results. The average accuracy of the subjects was 68.8% correct in Experiment 3. Analysis methods used here are the same as in Experiment 2 above. Figure 15B shows the probability of "same" responses as a function of the position along the parallel/orthogonal axis for the five target configurations used in Experiment 3. Figure 16D shows the means and the standard errors (across subjects) of the MAP estimates of ρ for the target configurations. As in Experiment 2, ρ tended to be high for target configurations with a small $|\Delta|$ and gradually decreased for target configurations with larger $|\Delta|$. The linear regression of the MAP estimates of ρ on $|\Delta|$ was significant, and the slope was negative, suggesting that ρ decreased with $|\Delta|$($b = -0.02$, $t(88) = -5.08$, $p < .01$).

Figure 16C shows, for a representative subject in Experiment 3 (Subject TH), the means and shapes of the underlying distributions $p(\hat{\theta}_1, \hat{\theta}_2|\theta_1,\theta_2)$ for the five target configurations (represented by the black crosses), with the parameters of these distributions set to their MAP estimates. We did not find a consistent bias toward mean orientations in subjects' responses. Note, for instance, for the particular subject shown in Figure 16C, the means of the underlying distributions (black and red dots) were closer to the main diagonal than the target configurations (black crosses) in only one out of four cases, the opposite pattern was apparent in the remaining three cases (note that the fifth configuration was on the main diagonal). This could either be due to the limited number of target configurations we tested or due to the particular feature dimension used in this study (i.e., orientation). Further experiments using a larger number of target configurations are needed to resolve this issue.

Discussion. Experiment 3 confirmed the specific pattern of correlations observed in Experiments 1 and 2 for a different feature dimension, namely, orientation. Correlations between the estimates of the orientations of different items in a target configuration tended to be high (and credibly different from 0) for small orientation differences and gradually decreased for larger differences.

General Discussion

There is now extensive experimental evidence suggesting that the content of a visual memory for even a simple display encoded in VSTM can be very complex. VSTM uses organizational processes that make the representation of an item dependent on the representations of other items as well as on the actual features of the displayed items (Brady & Alvarez, 2011; Brady et al., 2011; Huang & Sekuler, 2010; Jiang et al., 2000; Kahana & Sekuler, 2002). In other words, the way we remember an individual item might depend not only on the actual properties of that item but also on the properties of other items simultaneously presented with that item and on how we remember those other items as well.

To account for these dependencies, we proposed PCT, a theory of the organization of VSTM. PCT states that VSTM infers probability distributions over partitions or clusterings of visual items. Probabilistic clustering of items gives rise to dependencies in the joint representation of multiple items in VSTM. Representations of items belonging to the same cluster share parameters, and thus are dependent. Representations of items belonging to different clusters do not share parameters, and thus are independent. Importantly, VSTM does not determine a single partition. Rather, it determines a probability distribution over all possible partitions. This property allows it to represent items at multiple granularities or scales simultaneously. Because PCT hypothesizes that VSTM makes use of multiple scales, it can account for experimental data that has previously motivated hierarchical models of VSTM. In fact, as we

showed in the section on *Models*, a two-level hierarchical model proposed by Brady and Alvarez (2011) can be seen as a special case of PCT with a single cluster (i.e., where all items are always assigned to a single cluster). In the general case, however, PCT does not set any a priori bounds on the number of clusters, but rather automatically infers the appropriate distribution over the number of clusters to use and the scales of those clusters from the feature values of a given set of items. Because PCT hypothesizes that VSTM automatically determines which particular scales to use, it overcomes many of the shortcomings of hierarchical models with prespecified, fixed structures (e.g., see *Hierarchical Encoding of Items in VSTM*).

We explored several possible implementations of PCT: an exact implementation based on the Dirichlet process mixture model (DPMM) and approximate implementations based on Bayesian finite mixture models (BFMMs) with different numbers of components, including a model using a single component that we have shown to be equivalent to the two-level hierarchical model of Brady and Alvarez (2011). We consider the DPMM to be an exact implementation of PCT in the sense that it does not assume any a priori limit on the number of components to be used and considers all possible partitions of the given set of items. In practice, however, the DPMM and BFMMs with a sufficiently large number of clusters make indistinguishable predictions. Because of the small set sizes used in VSTM experiments and subjects' tendency to group items in a display (Woodman et al., 2003; and the studies reviewed in this article), "sufficiently large" can be as small as four to five components, as demonstrated by the quite good performance of the BFMM with four components in fitting the data we modeled in this article.

Through computer simulations, we demonstrated that PCT explains a number of biases and dependencies in VSTM representations previously reported in the literature. Through novel experiments, we evaluated a crucial prediction of PCT, namely, that dependencies between estimates of the feature values of different items based on their VSTM representations should decrease with the distance between the actual feature values of the corresponding items. This prediction was qualitatively confirmed in a series of experiments specifically designed to measure such dependencies. However, quantitatively, the observed dependencies decayed more slowly than predicted by all models considered in this article, suggesting that further improvements in the models or better alternative models are needed to more accurately characterize dependencies in subjects' estimates.

DPMM as a Rational Model of Encoding in VSTM Under Capacity Constraints

An important question concerns the status of PCT in terms of Marr's (1982) levels of analysis. Consider, for instance, the DPMM (our comments apply equally to a BFMM with a sufficiently large K). Do we put forward this model as addressing a computational-level analysis of the problem of encoding in VSTM, or rather as a representational/algorithmic level description of the organizational processes in VSTM? Before answering this question, we acknowledge that the boundary between these two levels can often be fuzzy. Indeed, some researchers have discussed "rational process models," which are specifically intended to blur the distinction between the levels (Sanborn et al., 2010). Nonetheless, we believe that the DPMM is best viewed as addressing a computational-level analysis of the problem of encoding in VSTM.

We think that it is possible to consider the DPMM as a rational model (Anderson, 1990) of encoding in VSTM under some capacity constraints. In what follows, we informally describe one possible way in which the DPMM can be construed as an optimal model of encoding in VSTM. Clearly, when faced with the problem of reporting the feature value of a target item in a briefly presented multi-item display, without any constraints on the capacity of encoding in VSTM, the optimal strategy is to encode all items' features with infinite precision. However, this optimal strategy is not attainable by people due to capacity limitations in VSTM (Bays & Husain, 2008; Luck & Vogel, 1997; Wilken & Ma, 2004; Zhang & Luck, 2008). Given these capacity limitations, there may be different ways of defining optimality. One possible way is to define it as the minimization of the expected squared error of a subject's estimates over the trials of a recall task. It is well-known that this expected error can be decomposed into terms representing the bias and variance of the estimator. If the estimator has low bias but high variance, the subject might tolerate a small increase in the bias term in exchange for a larger decrease in the estimator's variance to decrease the expected error. A (two-level) hierarchical model performs precisely this type of a trade-off to reduce the expected error of the estimates. More specifically, it does so by sharing information between the estimates of different items. This reduces the variance of the estimates of individual items, but introduces biases in these estimates. However, sharing information between the estimates of all items indiscriminately might not be the best strategy because the introduced biases might be too large and/or the reduction in variance too small due to the heterogeneity or dissimilarity of the estimates of different items. A still better strategy would be to share information selectively among the estimates of different groups of items. If, for example, only estimates of groups of highly similar items share information, the introduced biases would be minimal because the estimates of the feature values of items in a group would already be similar to each other, and the reduction in variance would be significant because the group would be relatively homogeneous (low variance), meaning that grouping reduces the variance of individual estimates maximally. This is the strategy implemented by the DPMM and BFMMs.

The optimal way to trade-off the estimator bias against its variance to minimize the expected error may depend on the memory noise (i.e., τ_{obs} in the models we considered in this article). For example, when the individual estimates are highly noisy (small τ_{obs}), the benefits of variance reduction through grouping may outweigh biases introduced in this way, hence the optimal strategy may be to prefer grouping whenever possible. On the other hand, for a large τ_{obs}, the gains from variance reduction might not be as large as the biases introduced through grouping. In this case, a preference against grouping might be optimal. In the DPMM and BFMMs, the parameter α_c controls such preferences for or against grouping.

A number of researchers have previously used the DPMM or closely related variants as rational models for categorization or category learning problems (Anderson, 1990, 1991; Griffiths, Sanborn, Canini, & Navarro, 2008; Sanborn et al., 2010). Thus, considering the DPMM as a rational model of encoding information about multiple items in VSTM under certain capacity limitations parallels an analogous approach to the study of human categorization.

A Novel Form of Dependence Between Representations of Multiple Items in VSTM

The *Experiments* section discussed the results of three novel experiments that we designed to empirically determine properties of the joint probability distribution, $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$, for small set sizes. This led to the discovery of a previously unrecognized form of dependence between estimates of the feature values of different items in VSTM: There were strong positive pairwise correlations between representations of different items when the feature values of the corresponding items were similar, and the magnitude of these correlations gradually decreased with the distance between their feature values. The existence of these correlations were confirmed for two different feature dimensions: horizontal location (Experiments 1 and 2) and orientation (Experiment 3). Future studies will need to test for similar correlations using other feature dimensions. It is entirely possible that for some feature dimensions (e.g., for more categorical features), correlations might not exist.

In a similar vein, Experiments 1 and 2 used a location change-detection task, whereas Experiment 3 used an orientation change-detection task. It would be interesting to examine the interactions between the effects of spatial proximity and feature similarity (e.g., orientation similarity) when measuring response correlations by, for instance, parametrically changing both of these factors. For example, it is conceivable that for very large spatial distances between two gratings, correlations between the estimates of their orientations might be reduced or might vanish altogether.

Another important question concerns the dynamics or time-course of these correlations. Do they arise during encoding or maintenance? What would their temporal profile look like over the course of the delay period? For example, initially in the delay period, the representations of the items might be less correlated and correlations might gradually increase or, alternatively, correlations might originate during the encoding phase and might be more or less stable thereafter. These questions can be addressed by parametrically varying the presentation duration (cf. Bays, Gorgoraptis, Wee, Marshall, & Husain, 2011) and the length of the delay interval.

Finally, our experiments used small set sizes and considered only pairwise correlations between representations of different items in VSTM. For larger set sizes, and especially for more naturalistic stimuli, the joint distribution characterizing the estimates of the feature values of multiple items will almost certainly involve more complex, higher order, and more interesting dependencies than just pairwise correlations. The extension of our experimental methods for uncovering possible dependencies in such cases in a feasible way would be very valuable.

Extension of Our Probabilistic Framework to More Naturalistic Stimuli

How can our probabilistic framework be extended to modeling the organization of VSTM for more naturalistic stimuli? It may be that characterizing VSTM through the joint probability distribution $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N)$ would not be a good starting point when considering naturalistic stimuli because these stimuli often cannot be described in terms of a number of simple feature values. That is, they typically exhibit a much richer structure. In recent years, there has been considerable progress in probabilistic modeling of natural scenes (Fei-Fei & Perona, 2005; Sivic, Russell, Efros, Zisserman, & Freeman, 2005; Sudderth, 2006) in computer vision. In particular, nonparametric hierarchical Bayesian models (such as extensions of DPMMs) of the structure of natural scenes have been successfully applied to challenging recognition and classification tasks (Sudderth, 2006). Assuming that these models provide a reasonably good description of the structure of natural scenes, one possible approach to modeling the content of a subject's visual memory for a natural scene might be to introduce capacity constraints in such models, similar to the capacity constraints in the models discussed in this article, such as constraints on the precision of the observable nodes in such models, where the observable nodes might simply be noisy observations of the lowest level features, just as in the models we considered in this article (although these lowest level features might be more complex than the ones we used in this article). In future work, it would be interesting to compare the performances of suitably constrained extensions of DPMMs with the performances of humans in VSTM experiments using natural scenes.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429. doi:10.1037/0033-295X.98.3.409

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage and reallocation of visual working memory. *Journal of Vision, 11*(10), Article no. 6. doi:10.1167/11.10.6

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321,* 851–854. doi: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. doi:10.1016/j.neuropsychologia.2010.12.023

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. doi:10.1177/0956797610397956

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and towards structured representations. *Journal of Vision, 11*(5), Article no. 4. doi:10.1167/11.5.4

Brady, T. F., & Tenenbaum, J. B. (2010). Encoding higher-order structure in visual working memory: A probabilistic model. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 411–416). Austin, TX: Cognitive Science Society.

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

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In C. Schmid, S. Soatto, & C. Tomasi *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA: IEEE Computer Society.

Ferguson, T. S. (1973). Bayesian analysis of some nonparametric problems. *Annals of Statistics, 1,* 209–230. doi:10.1214/aos/1176342360

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), Article no. 3. doi:10.1167/11.12.3

Fukuda, K., Awh, E., & Vogel, E. K. (2010). Discrete capacity limits in visual working memory. *Current Opinion in Neurobiology, 20,* 177–182. doi:10.1016/j.conb.2010.03.005

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton, FL: CRC Press.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117,* 197–209. doi:10.1037/a0017808

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112,* 21–54. doi:10.1016/j.cognition.2009.03.008

Görür, D. (2007). *Nonparametric Bayesian discrete latent variable models for unsupervised learning* (Unpublished doctoral dissertation). Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Görür, D., & Rasmussen, C. E. (2010). Dirichlet process Gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology, 25,* 653–664. doi:10.1007/s11390-010-9355-8

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for rational models of cognition,* 303–328. Oxford, England: Oxford University Press.

Hemmer, P., & Steyvers, M. (2009a). A Bayesian account of reconstructive memory. *Topics in Cognitive Science, 1,* 189–202. doi:10.1111/j.1756-8765.2008.01010.x

Hemmer, P., & Steyvers, M. (2009b). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review, 16,* 80–87. doi:10.3758/PBR.16.1.80

Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision, 10*(2), Article no. 24. doi:10.1167/10.2.24

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in establishing spatial location. *Psychological Review, 98,* 352–376. doi:10.1037/0033-295X.98.3.352

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

Ishwaran, J., & James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association, 96,* 161–173. doi:10.1198/016214501750332758

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. doi:10.1037/0278-7393.26.3.683

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research, 42,* 2177–2192. doi:10.1016/S0042-6989(02)00118-9

Kahana, M. J., Zhou, F., Geller, M. K., & Sekuler, R. (2007). Lure similarity affects visual episodic recognition: Detailed tests of a noisy exemplar model. *Memory & Cognition, 35,* 1222–1232. doi:10.3758/BF03193596

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision, 5*(5), Article no. 8. doi:10.1167/5.5.8

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS: A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337. doi:10.1023/A:1008929526011

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Freeman.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology, 50,* 101–122. doi:10.1016/j.jmp.2005.11.006

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics, 9,* 249–265.

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software, 35,* 1–81.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems, 12,* 554–560.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 5975–5979. doi:10.1073/pnas.0711295105

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117,* 1144–1167. doi:10.1037/a0020511

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review, 119,* 807–830. doi:10.1037/a0029856

Sivic, J., Russell, B., Efros, A. A., Zisserman, A., & Freeman, W. (2005). Discovering objects and their location in images. In *International Conference on Computer Vision (ICCV 2005)* (pp. 370–377). Los Alamitos, CA: IEEE Computer Society.

Sudderth, E. B. (2006). Graphical models for visual object recognition and tracking (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 8780–8785. doi:10.1073/pnas.1117465109

Vidal, J. R., Gauchou, H. L., Tallon-Baudry, C., & O'Regan, J. K. (2005). Relational information in visual-short term memory: The structural gist. *Journal of Vision, 5*(3), Article no. 8. doi:10.1167/5.3.8

Visscher, K. M., Kaplan, E., Kahana, M. J., & Sekuler, R. (2007). Auditory short-term memory behaves like visual short-term memory. *PLoS Biology, 5*(3), e56. doi:10.1371/journal.pbio.0050056

Viswanathan, S., Perl, D. R., Visscher, K. M., Kahana, M. J., & Sekuler, R. (2010). Homogeneity computation: How inter-item similarity in visual short term memory alters recognition. *Psychonomic Bulletin & Review, 17,* 59–65. doi:10.3758/PBR.17.1.59

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), Article no. 11. doi:10.1167/4.12.11

Woodman, G. F., Vecera, S. P., & Luck, S. J. (2003). Perceptual organization influences visual working memory. *Psychonomic Bulletin & Review, 10,* 80–87. doi:10.3758/BF03196470

Yotsumoto, Y., Kahana, M. J., Wilson, H. R., & Sekuler, R. (2007). Recognition memory for realistic synthetic faces. *Memory & Cognition, 35,* 1233–1244. doi:10.3758/BF03193597

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

Zhou, F., Kahana, M. J., & Sekuler, R. (2004). Short-term episodic memory for visual textures: A roving probe gathers some memory. *Psychological Science, 15,* 112–118. doi:10.1111/j.0963-7214.2004.01502007.x

Appendix A

Posterior Inference

For the DPMM, posterior inference was performed using Algorithm 8 of Neal (2000). An excellent description of this algorithm can also be found in Görür (2007). For the BFMMs, we used a Gibbs sampling algorithm for posterior inference (Algorithm 2.1 in Sudderth, 2006). Finally, posterior inference in the HBM was performed in WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), modifying the WinBUGS code provided in the supplemental material of Brady and Alvarez (2011). For all models, in recall tasks, we used the posterior means as the models' estimates of the feature values of different items. As described in more detail in the main text, the procedure for the simulation of the Viswanathan et al. (2010) was slightly different, because this study involved a recognition memory task, unlike the other studies modeled here, which were all recall tasks. The source code used in the simulations reported in this article is available upon request from the authors.

Appendix B

Maximum-Likelihood Estimation of the Parameters and Model Comparison

In our simulations, we used grid searches to find the values of the parameters (for the DPMM and the BFMMs, α_c and τ_{obs} in the univariate case and κ and σ^2_{obs} in the multivariate case; for the HBM, τ_{obs}) that maximized an approximation to the likelihood given the observed data. The grid range over which we searched was determined through trial-and-error in each case.

In the simulation of the experiment by Wilken and Ma (2004), the data were the average observed errors in 64 conditions (four set sizes $\times$ 16 bins; see the top-left plot in Figure 6). Each model effectively produces a different likelihood distribution $p_M(biases|FP)$ as a function of its free parameters denoted by FP, where M could be DPMM, BFMM–2, BFMM–4, or HBM and biases denotes the collection of variables representing the average biases in each of the 64 conditions. These likelihood distributions do not have an analytic form; therefore, we computed them by sampling. Specifically, for each model, we simulated the experiment of Wilken and Ma (2004) 25 times and, for each of the 25 runs of the simulated experiment, collected the average biases predicted by the model in each of the 64 conditions. Furthermore, because trials in each simulated experiment are independent of each other, the distribution $p_M(biases|FP)$ can be factorized as $p_M(bias_1|FP)\, p_M(bias_2|FP) \ldots p_M(bias_{64}|FP)$ with $bias_i$ representing the average bias in the ith condition. We then approximated each $p_M(bias_i|FP)$ using a nonparametric kernel density estimate generated from the 25 collected samples of $bias_i$. The log-likelihood of a specific setting of the parameters FP given the observed data is then calculated as $\Sigma_{i=1}^{64} \log p_M(bias_i = obs\_bias\_i|FP)$, where obs_bias_i denotes the observed average bias in the ith condition. This procedure was repeated for each setting of the free parameters FP over the grid. The parameter values that maximized the estimated log-likelihood were chosen as the maximum likelihood (ML) estimates of the parameters. For the DPMM and the BFMMs, the parameter τ_{obs} was allowed to vary across set sizes, whereas α_c was fixed across set sizes. Due to the relatively high dimension of the search space in this particular simulation (five free parameters for the DPMM and the BFMMs and four free parameters for the HBM), grid searches were conducted in a greedy fashion. The uniform base distribution for μ was defined over the interval [0, 12]. The exact values of the endpoints of this interval did not affect the simulation results for this particular experiment, or for other experiments, as long as the interval was large enough to include the minimum and maximum possible values of the relevant feature in an experiment.

In the simulation of the experiment by Viswanathan et al. (2010), the data were the total number of observed "old" (or "yes") responses in 1,800 trials each of the medium- and high-homogeneity conditions. The likelihood in each case was modeled as a binomial distribution with $n = 1{,}800$ (number of trials) and the model's predicted success probability p, which was a function of the parameters of the model. The parameter values that maximized the log-likelihood, $\log(\text{Binomial}(k = obs\_med\_k;\ n,\ p_{med}(FP))) + \log(\text{Binomial}(k = obs\_high\_k;\ n,\ p_{high}(FP))$, where obs_med_k and obs_high_k are the total number of observed "old" responses in the medium and high homogeneity conditions, respectively, and $p_{med}(FP)$ and $p_{high}(FP)$ are the model's predicted success probabilities in the two conditions), were chosen as the maximum likelihood (ML) estimates of the parameters. The uniform base distribution for μ was defined over the interval [0, 11].

In the simulation of the experiments by Brady and Alvarez (2011), the data were the observed mean biases in Experiment 1 and Experiment 2. As in the simulation of the Wilken and Ma (2004) experiment, each model produces a different distribution over the mean biases in the simulated Experiments 1 and 2, $p_M(bias_1|FP)$ and $p_M(bias_2|FP)$, as a function of the free parameters. Since these distributions do not have an analytic form, we drew 25 samples from these distributions by simulating each experiment 25 times and computing the predicted mean bias in each case. We then approximated $p_M(bias_1|FP)$ and $p_M(bias_2|FP)$ with a nonparametric kernel density estimate using the 25 collected samples.

(*Appendices continue*)

The parameter values that maximized the estimated log-likelihood, $\log(p_M(bias_1 = obs\_mean\_bias_1|FP)) + \log(p_M(bias_2 = obs\_mean\_bias_1|FP))$, where $obs\_mean\_bias_1$ and $obs\_mean\_bias_2$ denote the observed mean biases in Experiment 1 and Experiment 2 of Brady and Alvarez (2011) were chosen as the maximum likelihood (ML) estimates of the parameters. For the multivariate DPMM and BFMMs, the same model was applied to both experiments. As explained in the main text, we placed a vague inverse-Wishart prior on Ψ (the inverse scale parameter of the base distribution for Σ). Specifically, the inverse scale parameter of the inverse-Wishart prior on Ψ was $3000I$, where I is the identity matrix, and its degrees-of-freedom parameter was 2 to maximize the variability (or vagueness) of the prior. The uniform base distribution for μ was defined over the interval [–50, 200] for both color and size dimensions.

In the application of the models to data from our own experiments, for each subject, the data were trial-by-trial responses of the subject. In the case of the DPMM and the BFMMs, for each configuration $\{\theta_i\}_{i=1}^N$, the models predicted a joint distribution, $p(\{\hat{\theta}_i\}_{i=1}^N|\{\theta_i\}_{i=1}^N,\alpha_c,\tau_{obs})$, over the estimates of the items as a function of the free parameters α_c and τ_{obs}. This distribution was computed by sampling using Equation 2. For each configuration, 50 samples were drawn from the joint distribution of the estimates. The log-likelihood of parameters given the subject's responses for that particular configuration was then computed by evaluating each response under a bivariate Gaussian approximation to the joint distribution constructed from those 50 samples, taking its logarithm and summing over all responses. This was repeated for all stimulus configurations and the total log-likelihood was computed by summing over all configurations. The parameter values that maximized the estimated total log-likelihood were chosen as the maximum likelihood (ML) estimates of the parameters. The uniform base distribution for μ was defined over the interval [–10, 10].

We used the Bayesian information criterion (BIC) to compare the model fits (Schwarz, 1978). BIC scores were computed according to

$$BIC = -2\log L + k \log n, \tag{B1}$$

where L denotes the maximum likelihood of the model (i.e., the likelihood value achieved when ML estimates of the parameters are used), k is the number of parameters and n is the number of data points. BIC scores reported in the article are relative to the BIC score for the DPMM.

Appendix C

Details of the Bayesian Model of Same/Different Responses in Experiments 2 and 3

In the application of the Bayesian model to experimental data from Experiment 2, the following priors were used: Each of μ_1 and μ_2 were given Gaussian priors with means centered on the actual feature values (i.e., horizontal locations) of the two items in the target configurations and with precision 0.01. The correlation coefficient of the underlying bivariate Gaussian distribution, ρ, was given a uniform prior over the interval [–1, 1]. The standard deviations along each of the two dimensions, σ_1 and σ_2, were given uniform priors over the interval [0.1, 1.3]. The lower bound on the probability of "same" responses, b_l, was given a uniform prior over [0, 0.5], and the upper bound on the probability of "same" responses, b_u, was given a uniform prior over [0.5, 1.0].

In the application of the Bayesian model to experimental data from Experiment 3, the following priors were used: Each of μ_1 and μ_2 were given Gaussian priors with means centered on the actual feature values (i.e., orientations) of the two items in the target configurations and with precision 0.01. The correlation coefficient of the underlying bivariate Gaussian distribution, ρ, was given a uniform prior over the interval [–1, 1]. The standard deviations along each of the two dimensions, σ_1 and σ_2, were given uniform priors over the interval [1.0, 18.0]. The lower bound on the probability of "same" responses, b_l, was given a uniform prior over [0, 0.5], and the upper bound on the probability of "same" responses, b_u, was given a uniform prior over [0.5, 1.0].

We implemented the model in Python using the PyMC package (Patil, Huard, & Fonnesbeck, 2010). For each of the two experiments, 11,000 samples were drawn from the posterior distributions of the variables. The first 1,000 samples were discarded as burn-in, and the remaining 10,000 samples were "thinned" down to every 10th sample to reduce dependencies between consecutive samples. This reduced the number of samples to 1,000. The results reported in this article are based on these 1,000 samples. Proper convergence and mixing were monitored and confirmed both visually and through a battery of diagnostics provided by the PyMC package.

Received August 1, 2011
Revision received November 5, 2012
Accepted November 8, 2012 ■