Contextual effects in visual working memory reveal hierarchically structured memory representations

Timothy F. Brady
Department of Psychology, University of California, San Diego, La Jolla, CA, USA

George A. Alvarez
Department of Psychology, Harvard University, Cambridge, MA, USA

Influential slot and resource models of visual working memory make the assumption that items are stored in memory as independent units, and that there are no interactions between them. Consequently, these models predict that the number of items to be remembered (the set size) is the primary determinant of working memory performance, and therefore these models quantify memory capacity in terms of the number and quality of individual items that can be stored. Here we demonstrate that there is substantial variance in display difficulty within a single set size, suggesting that limits based on the number of individual items alone cannot explain working memory storage. We asked hundreds of participants to remember the same sets of displays, and discovered that participants were highly consistent in terms of which items and displays were hardest or easiest to remember. Although a simple grouping or chunking strategy could not explain this individual-display variability, a model with multiple, interacting levels of representation could explain some of the display-by-display differences. Specifically, a model that includes a hierarchical representation of items plus the mean and variance of sets of the colors on the display successfully accounts for some of the variability across displays. We conclude that working memory representations are composed only in part of individual, independent object representations, and that a major factor in how many items are remembered on a particular display is interitem representations such as perceptual grouping, ensemble, and texture representations.

Introduction

Working memory is the ability to hold information actively in mind, and to manipulate that information to perform a wide variety of cognitive tasks (Baddeley, 2000). This memory system constrains processing across many domains. For example, individual differences in working memory capacity predict differences in fluid intelligence, reading comprehension, and academic achievement (Alloway & Alloway, 2010; Daneman & Carpenter, 1980; Fukuda, Vogel, Mayr, & Awh, 2010; Oberauer, Schulze, Wilhelm, & Süß, 2005). Thus, understanding the architecture and limits of the working memory system is a fundamental goal for cognitive science, and many models have been developed to help explain the limits on mental storage capacity (Cowan, 2001; Miller, 1956; Miyake & Shah, 1999). In the domain of visual working memory, these models have grown particularly sophisticated and have been formalized in an attempt to derive measures of working memory capacity (Alvarez & Cavanagh, 2004; Bays, Catalao, & Husain, 2009; Cowan, 2001; Luck & Vogel, 1997; Wilken & Ma, 2004; Zhang & Luck, 2008).

Most models of visual working memory agree that capacity can be quantified in terms of the number of individual items stored and the precision with which those items are stored, although these models disagree on the nature of working memory resources and how they are allocated to individual items (for a review, see Brady, Konkle, & Alvarez, 2011; Suchow, Fougnie, Brady & Alvarez, 2014). For example, some models view memory as limited by resources that are continuously divisible and flexibly allocated to either objects or features (e.g., Alvarez & Cavanagh, 2004; Bays & Husain, 2008; Fougnie, Asplund, & Marois, 2010; Wilken & Ma, 2004), whereas other models view memory as limited by fixed slots that are constrained to represent a discrete number of objects (e.g., Awh, Barton, & Vogel, 2007; Cowan, 2001; Luck & Vogel, 1997; Miller, 1956; Zhang & Luck, 2008).

Due to this focus on individual item representations, the vast majority of working memory studies attempt to isolate memory for individual items by constructing

Citation: Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, *15*(15):6, 1–24, doi:10.1167/15.15.6.

doi: 10.1167/15.15.6 Received May 27, 2015; published November 17, 2015 ISSN 1534-7362 © 2015 ARVO

displays composed of simple stimuli that are randomly chosen and randomly positioned. Such working memory displays are, as best as possible, prevented from having any overarching structure, gist, or perceptual grouping cues, and analyses are done by averaging over all of the displays to attempt to remove these factors. Results from these studies are then typically modeled by assuming that each item is stored as an independent unit and that items do not influence one another's representation (Alvarez & Cavanagh, 2004; Bays et al., 2009; Luck & Vogel, 1997; Rouder et al., 2008; van den Berg, Shin, Chou, George, & Ma, 2012; Wilken & Ma, 2004; Zhang & Luck, 2008; although see Huang & Sekuler, 2010; Johnson, Spencer, Luck, & Schoner, 2009; Lin & Luck, 2009). As such, these models either implicitly or explicitly propose that working memory holds a list of unrelated items and that items representations do not affect each other.

However, in contrast to these standard modeling assumptions, it is difficult if not impossible for any display to isolate individual item representations, as the visual system is designed to represent structured real-world scenes rather than simple, unrelated geometric shapes (Felsen & Dan, 2005; Simoncelli & Olshausen, 2001). Thus, even in simple displays, the visual system represents high-level texture and ensemble information (Alvarez, 2011; Brady & Alvarez, 2015; Freeman & Simoncelli, 2011; Haberman & Whitney, 2007; Portilla & Simoncelli, 2000; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012) and performs perceptual organization processes using simple Gestalt rules (Kubovy & Pomerantz, 1981; Palmer, 1999) and more complex integration mechanisms. These mechanisms take into account spatial frequency and combine multiple low-level features into higher order representations of texture and scene gist (Brady & Oliva, 2012; Oliva, 2005; Oliva & Schyns, 1997). If the visual system uses both individual item information *and* ensemble properties of the display—what we refer to as structured representations—even for the simplest displays, then it is unlikely that the number of individual items is the only determinant of performance on visual working memory tasks. Instead, different displays, even with the same number of items, will vary in their memory representations based on how the items combine into perceptual groups and/or ensemble representations.

Indeed, there is significant evidence that such ensemble and perceptual grouping effects affect working memory capacity estimates. This is consistent with the general fact that nearly all memory is strongly context-dependent (e.g., Godden & Baddeley, 1975; Howard & Kahana, 2002; Tulving & Thomson, 1973). Specifically in visual working memory, items appear to be encoded with respect to a spatial context (Jiang, Olson, & Chun, 2000), such that if the participants' task is to detect whether a particular item changed color, performance is worse if the other items in the display do not reappear or if they reappear with their relative spatial locations changed (see also Olson & Marshuetz, 2005; Vidal, Gauchou, Tallon-Baudry, & O'Regan, 2005). Items are also represented with a temporal context (e.g., Kahana, Zhou, Geller, & Sekuler, 2007; Nosofsky & Kantner, 2006; Viswanathan, Perl, Visscher, Kahana, & Sekuler, 2010), such that the general similarity of a set of items modulates memory for each particular item. Displays where objects group together into perceptual units also result in better visual working memory performance, as though each unit in the group was encoded more easily (Woodman, Vecera, & Luck, 2003; Xu, 2006; Xu & Chun, 2007). Similarly, visual working memory performance is improved when items appear more similar to one another (Lin & Luck, 2009; Viswanathan, Perl, Visscher, Kahana, & Sekuler, 2010; see also Johnson et al., 2009), perhaps because people encode items relative to each other (Lin & Luck, 2009).

In addition to effects of item similarity and perceptual grouping, participants are better able to recognize changes to displays if those changes alter the ensemble statistics of the display; For example, if a display is changed from mostly black squares to mostly white squares, participants notice this change more easily than a matched change that does not alter the global statistics (Brady & Tenenbaum, 2013; Victor & Conte, 2004; see also Alvarez & Oliva, 2009). This can have important impacts on working memory models: For example, changes that completely change the identity of an item are also likely to alter the global texture of the display, confounding estimates of how many items can be remembered with memory for spatial ensembles or texture information (Brady & Alvarez, 2015). Such ensemble representations also result in biases in memory for individual items, such that participants tend to report items as more similar to the other items on a trial than they really were (Brady & Alvarez, 2011; Huang & Sekuler, 2010; Orhan & Jacobs, 2013). Thus, there are a number of known influences of interitem representations on visual working memory, above and beyond the effects of memory for individual items. Yet, to date these effects have not been demonstrated and modeled in standard working memory tasks.

In the present study, we explore the visual system's tendency to make use of perceptual grouping, ensemble representations, and other forms of structured representations in a standard color working memory task. In doing so, we seek to illustrate and understand the impact that forming structured representations has on quantifying working memory capacity. Although much work has suggested that the assumption of independent object representations is too simplistic, it is nevertheless often assumed that the primary constraint in visual

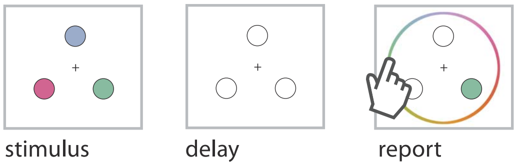

Figure 1. Participants performed a continuous report task, where they were briefly shown a set of colors and then had to report the exact color of items that were present at a cued location. They reported all three colors on each trial, with each location cued one at a time in a random order.

working memory is the number of individual items that can be stored, with only minor adjustments needed to account for how objects influence each other (e.g., Johnson et al., 2009; Lin & Luck, 2009, 2012). This assumption of independence is particularly prevalent in simple color memory experiments (e.g., Zhang & Luck, 2008), as the vast majority of perceptual grouping and ensemble effects have been demonstrated with either spatial memory or more complex stimuli (e.g., Brady & Alvarez, 2011, 2015; Brady & Tenenbaum, 2013; Xu, 2006), and the canonical color memory task (e.g., Luck & Vogel, 1997; Zhang & Luck, 2008) is often assumed to avoid many confounds and complications present in other visual working memory experiments (e.g., Lin & Luck, 2012). Thus, memory for color is an important test bed for understanding the role of context in individual item memory.

To explore the role of interitem effects and structured representations in visual working memory, we introduce a novel technique for probing the contents of visual working memory based on showing the same individual displays to hundreds of participants. This technique allows us to understand the contents of working memory for each item in each display. Intuitively, even though displays are randomly generated, some are likely to be easier or harder to remember. For example, a display might be easier to remember if all the left-most items are warm colors and all the right-most items are cold colors, compared to a display where the items are heterogeneous and intermixed. By examining the representation of each item on each display, we can determine the extent to which working memory performance is affected by interitem factors that vary across displays even when set size is constant.

We find that participants are highly consistent in which items and displays are hardest or easiest to remember and how precisely they are remembered. In addition, participants seem to represent ensemble information independent of their memory for individual items (e.g., even when participants are wrong about the colors of the items, they maintain information about the variance of the colors). A simple grouping or chunking strategy cannot explain this individual-display variability, but a model that includes a hierarchical representation of items plus the mean and variance of the colors on the display can account for some of the variability across displays. These findings demonstrate that a major factor in how many items are remembered on a particular display is interitem representations such as perceptual grouping, ensemble, and texture representations. We conclude that item-based models of visual working memory should be updated to capture the rich, structured nature of working memory representations.

Methods

Participants

For the main experiment, 300 participants were recruited on Amazon's Mechanical Turk. All participants were from the United States, were over 18 years old, and gave informed consent in accordance with the procedures and protocols approved by the Harvard Committee on the Use of Human Subjects. Turk users form a representative subset of adults in the United States (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011), and data from Turk users are known to closely match data from the lab on working memory tasks (Brady & Alvarez, 2011; Brady & Tenenbaum, 2013). All participants indicated they had normal or corrected-to-normal color vision. All participants were paid approximately 50 cents for several minutes of their time.

Procedure

Participants were given instructions on the memory task, and then consented to participate in the study. They were then asked to confirm they had normal or corrected-to-normal color vision and to confirm the entire experimental display was visible on their computer monitor. After they did so, the main experiment began. On each trial, participants saw three colored circles, presented a fixed distance from the center of the screen, as in Figure 1. The colors were present for 1000 ms and then disappeared for 1000 ms. Then an item was cued and participants used a color wheel on the screen to select the color the cued item had been, entering their response by clicking the mouse button (see Figure 1). Participants were next cued to another item and asked to report that item's color, until all three items had been tested. Items were tested in a random order for each participant.

Critically, each of our participants saw the exact same 48 displays, with the order of displays randomized across participants. These displays were generated by independently choosing a random color for each location, in line with previous experiments examining working memory capacity (Zhang & Luck, 2008). In contrast to previous work, however, testing all participants on the same displays enabled us to determine what participants represent about particular items and particular displays. The colors' spatial locations were randomized across participants, thus enabling us to focus on the contribution of the particular set of colors to performance independent of their locations. This ensured that a propensity to encode objects from particular spatial locations would not cause some colors in a particular display to be represented more often or more accurately than other colors.

Stimuli

We used the same color space as used in Zhang and Luck (2008). In particular, we used a fixed luminance circle ($L^* = 70$) through $L^*a^*b^*$ color space with a radius of 60 centered on the point $a^* = 20$, $b^* = 38$. We used Adobe Flash to present the stimuli, as this software uses the color profile participants have chosen for their monitor. Because of the nature of experiments run on the Internet, we could not ensure participants' monitors were properly color calibrated. However, variations in the color calibration of participants' monitors can only artificially decrease the consistency between participants. Since our main result is the remarkable consistency across particular participants on the same displays, it is unlikely that color calibration was a major concern.

Control experiments

In addition to the main experiment, an additional 1,200 participants ran in three control experiments: 300 at Set Size 1; 300 at Set Size 6; 300 in a replication of the main experiment (Set Size 3) with the displays rotated 180° in color space (e.g., red became green); and 300 in a replication of the main experiment (Set Size 3) where only a single item was tested on each display. At Set Size 6, only half of the items were tested on each display, and which items were tested was counterbalanced across participants such that all six items were tested on each display equally often across participants.

Data quality

In general, we find that participants on Mechanical Turk enjoy our color memory task, and perform as well as participants we run in the laboratory. This is consistent with the fact that Mechnical Turk users have been known to outperform participants in the lab on difficult cognitive tasks (Goodman, Cryder, & Cheema, 2013). We excluded no subjects from our analyses and yet found strong consistency between participants. In the current experiments, our participants on Mechanical Turk had capacities approximately the same as the participants from Zhang and Luck (2008), even for difficult trials where we would expect the largest difference (e.g., at Set Size 6, $p_{mem} = 0.41$ in Zhang & Luck, 2008, and $p_{mem} = 0.42$ in our participants).

Modeling methods

Overview

Error data from the continuous report task can be fit using several formal models of working memory, which we can then compare using standard model comparison techniques. On each trial, participants reported a single color for each item, and we calculated the circular distance (in degrees on the color wheel) between the color they reported and the actual color of the item. When combined across multiple participants, this procedure results in a histogram of errors that appears to consist of roughly a normal distribution centered near the correct response, along with a collection of responses far from the correct response, effectively uniformly distributed over the color wheel (see Figure 2).

We can estimate parameters of these error distributions to compare several working memory models, which we group into three classes: item-based, chunk-based, and hierarchical. Item-based and chunk-based models are essentially the same, focusing on the number of independent units (items or chunks) that can be stored, except that multiple items are sometimes treated as a single unit in the chunk-based model. In contrast, the hierarchical model posits multiple, interacting levels of representation (e.g., individual items and clusters of individual items). Here we describe how variants of each model class can be formalized in terms of parameters that can be estimated from the observed error distributions.

Item-based models

The standard item-based model

Item-based models assume that participants' color reports are entirely item based, such that: (a) If the cued target item was remembered, but there was some noise or uncertainty in its exact color, then the error distribution would be a normal distribution centered around zero, with wider distributions indicating noisier representations. (b) If the cued item was not remembered, then the observer would guess randomly, resulting in a uniform distribution of errors. Standard item-based models assume that the overall error distribution is a mixture of these two types of responses (Zhang & Luck, 2008). Thus, item-based models can be formalized using a mixture model that combines a von Mises distribution (a circular normal distribution) to capture the remembered-item responses, and a uniform distribution to capture the more disparate responses (Zhang & Luck, 2008).

This standard item-based model has three parameters: p_{mem}, the proportion of responses that were target-related (i.e., the proportion that come from the von Mises distribution rather than the uniform distribution); a *bias* term indicating whether participants' target-related responses were, overall, shifted clockwise or counterclockwise relative to the correct response; and a standard deviation parameter, *SD*, indicating the fidelity of participants' target-related responses (see Appendix for further details). Item-based models interpret the p_{mem} parameter in terms of item capacity (the number of items represented), and the *SD* parameter as the precision with which participants represent items in memory. While this may be an incorrect theoretical interpretation of these parameters, they can nevertheless provide a reasonable summary of the response distribution even if they are not properly interpreted in terms of capacity and precision alone. Details of the model specification and model fitting procedures can be found in the Appendix.

We can fit such an item-based model to either all the data across all the displays, as is standard, giving a three-parameter model, or we can fit a separate value of *bias*, *SD*, and p_{mem} for each item on each of our 48 displays separately, giving a 432 parameter model (48 displays × 3 items/display × 3 parameters/item). To fit the model we used the *MLE* function of the MemToolbox (memtoolbox.org; Suchow, Brady, Fougnie, & Alvarez, 2013), which relies upon MATLAB's standard function minimization techniques, with several starting points for the optimization used to avoid local minima.

Item-based, swap model

In addition to a standard mixture model, we also consider a "swap model," as described by Bays et al. (2009). This model is similar to the standard mixture model of Zhang and Luck (2008) but with the addition of another parameter to capture the possibility that participants sometimes report the wrong item from a display. Thus, in addition to considering the target color, this model also considers the *m* distractor colors present on the same display, and the likelihood of reporting these values incorrectly, $p_{distractor}$. Details of the model specification and model fitting procedures can be found in the Appendix.

Chunk-based models

Chunk-based models posit that groups of similar items will be chunked together, and stored as a single unit in memory (e.g., upon noticing two of the items are a similar green, only a single green will be stored in memory with a tag indicating it goes with both locations). There are no formal models of chunking in the literature that we could apply to the current data, and there are a wide variety of possible formalizations of such a model. Rather than formalize a particular chunking model, we conducted a number of analyses that test specific predictions of the chunk-based account (see the Results section). These predictions were qualitative predictions that are either (a) necessarily for any chunk-based model and/or (b) impossible to account for using any version of a chunk-based model. This allows us to examine this entire class of models at once without explicitly formalizing a chunk-based model to fit to the error histograms.

Structured representation model

Finally, we consider a model based on hierarchically structured representations, which posits that working memory stores multiple, integrated levels of representation. Specifically, it assumes that memory stores individual item information and information about clusters of items, and that these levels of representation are nonindependent: The distribution of individual items constrains the likely clusters in the display, and the set of likely clusters in the display influences the representation of each individual item (as in Brady & Alvarez, 2011; Orhan & Jacobs, 2013). Critically, this model, like the standard item-based model, only has three parameters: the probability of retaining individual item information, the precision of individual item representations, and the strength of clustering between items. A detailed specification of this model and model-fitting procedures are described in the Appendix.

Results

Variance between displays is high at Set Size 3

The results showed a striking degree of variability in error histograms across individual displays and individual items (see Figures 2 and 3), even within displays at Set Size 3 (Figure 3; see Supplemental Material for

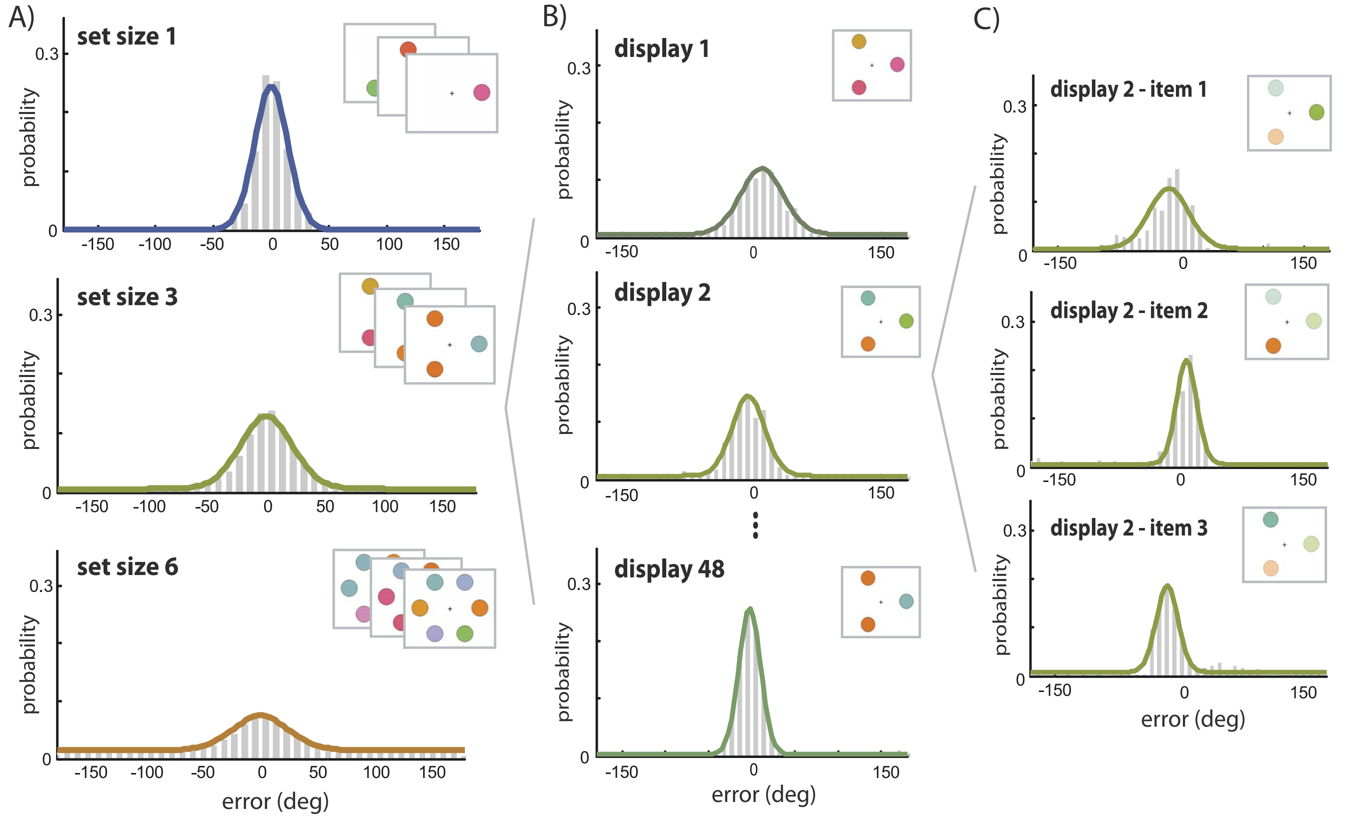

Figure 2. (A) Error histograms across all displays at Set Sizes 1, 3, and 6, and item-based model fits to these histograms. Participants' responses are modeled as a mixture of a circular normal distribution (corresponding to the remembered items) and a uniform distribution (corresponding to guesses or other nontarget related responses). (B) Data and model fits for a representative set of three of the 48 individual displays we tested at Set Size 3. Different displays have different standard deviations (*SD*), and different rates of target memory (p_{mem}), in contrast to what we'd expect if participants represented each item equally well and without any interactions among items, in which case all Set Size 3 displays should look similar. (C) Fits to the individual items within the second display (the item in full color corresponds to the shown data). Individual items on individual displays have highly reliable but distinct representations, also contrasting with the prediction of slot and resource-based models.

graphs of all items in all displays). To emphasize the magnitude of variability at Set Size 3, it is useful to compare the variability across set sizes (Figure 2, left column) to the variability within Set Size 3 displays (Figure 2, center column). For example, notice that one of the better remembered Set Size 3 displays (Figure 2, center/bottom) shows more accurate representations than the overall Set Size 1 distribution (Figure 2, left/top). In general, this variability is important because models based on individual items, like slot or resource models, cannot account for this variability because they assume items are independently represented and so predict no interactions between them.

We can also fit a standard mixture model (e.g., with a *bias*, p_{mem}, and *SD*) to the error distributions across displays at each set size, and separately for each individual item in each individual display. Fitting each item in each display independently with such an item-based model reveals that different displays have different standard deviations (*SD*) and probabilities of reporting a target-related response (p_{mem}), and that these differences across displays are reliable (see Figure 3). Split-half correlations provide estimates of the reliabilities of *bias*, *SD*, and p_{mem} for each display as $r =$ 0.98, 0.87, and 0.86, respectively. The variance between different individual displays at Set Size 3 is thus both large and reliable across different participants. Some items show higher or lower precision than the average item, in a way that is consistent across participants. And some items show higher or lower probabilities of being remembered than other items, also in a way that is quite consistent across participants.

High variance between displays at Set Size 3 rejects standard item-based models

One way to formalize the variation across individual displays is by comparing a standard item-based model fit to the data across all displays, which assumes all displays have the same bias, capacity (p_{mem}), and

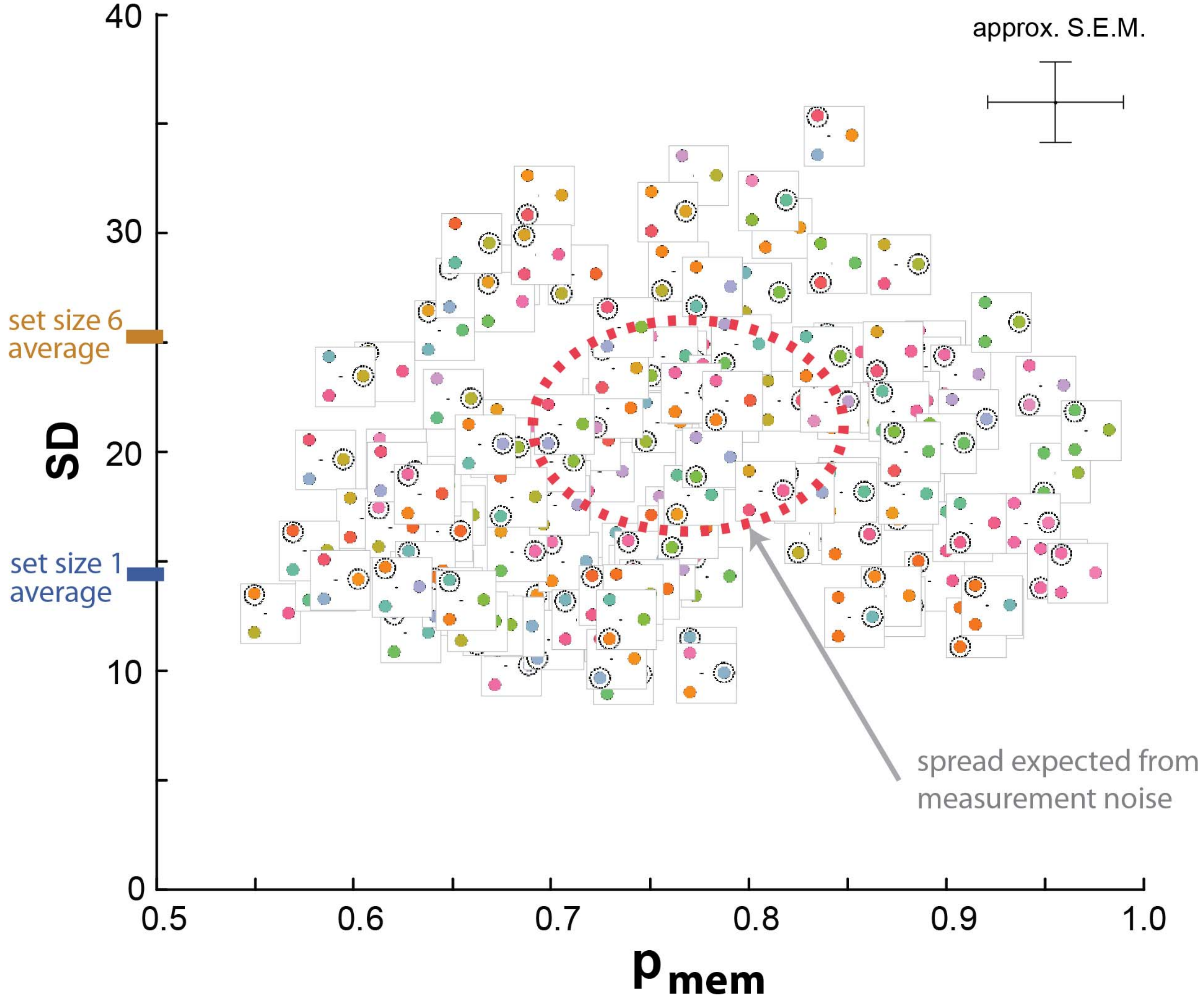

Figure 3. Variance in participants' memory for individual items. Each individual item is plotted as a separate point (48 displays × 3 items/display = 144 items plotted). The tested item for each point is shown with a dashed black circle. The large spread—well beyond that expected from measurement noise—demonstrates that different items on different displays are remembered with reliably different precisions (*SD*, *y*-axis) and are reported correctly with different frequencies (p_{mem}, *x*-axis). The average size of the error bars for an individual item is shown in the top right corner of the graph. The visible spread indicates that within Set Size 3, the p_{mem} values for individual items ranged from nearly 0.5 to nearly 1.0, and the *SD*s ranged from below 10° to nearly 35°. This is a difference in capacity estimates of nearly 50% for different items, and, even more importantly, the range of precisions present within Set Size 3 is larger than the difference in mean precision across set sizes (mean *SD* at Set Size 1 was 14.1°, at Set Size 6, it was 25.7°, while within Set Size 3 alone, *SD* estimates range from 9.9° to 33.6°).

fidelity (*SD*) to a model that fits each item independently and thus assumes all items in all displays are entirely independent, with each item having a separate bias, p_{mem}, and *SD*. Contrasting these models is a test of the idea that items are not equal and interchangeable. In particular, it tests whether it is better to assume that all Set Size 3 displays have the same *bias*, *SD*, and p_{mem} or to assume that there is no relationship at all between the *bias*, *SD*, and p_{mem} for one Set Size 3 display and another.

We compare these two models with formal model comparison techniques, rather than simply look at the goodness-of-fit (e.g., r^2), because these two models have vastly different numbers of free parameters, and models with more free parameters will generally fit better than models with fewer parameters. Model comparison techniques like Bayesian information criteria (BIC) or Akaike information criteria (AIC) penalize models for complexity to prevent such overfitting. In this case, even the model comparison technique that most

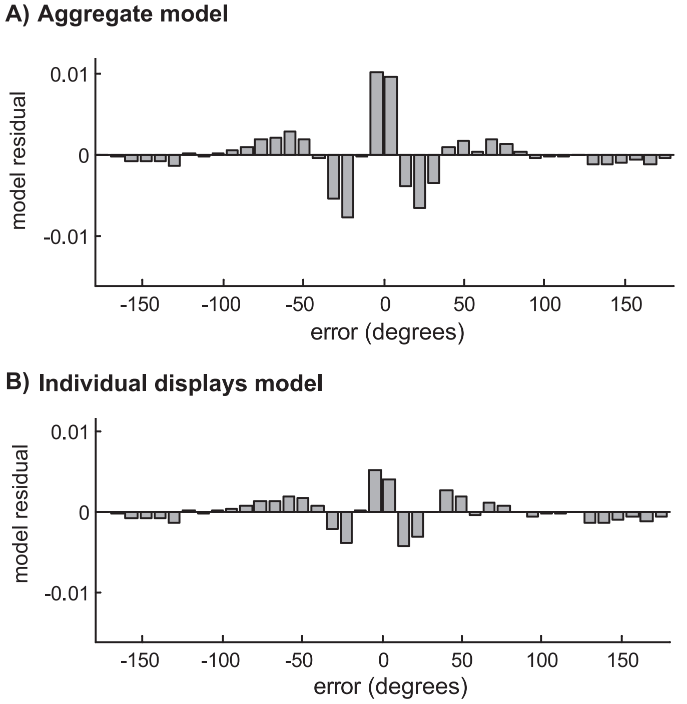

Figure 4. Residuals from the (A) aggregate model fit across all displays, and (B) separate models fit to each display. In both cases the data is "peakier" than the model fit; however, this is significantly attenuated when taking into account the reliable differences between individual displays.

punishes models for having more parameters (BIC, which penalizes complex models heavily) indicates that the data provide more evidence for a 432 parameter model that treats each individual item as totally independent (48 displays × 3 items/display × 3 parameters/item) than the traditional three-parameter model that treats all items on all displays as interchangeable. Furthermore, this difference is very large (BIC difference: 6,449; a difference of 10 or more is considered strong evidence). This effect is even stronger when using AIC, which penalizes complex models less but is more widely used (AIC difference: 10,170). Bootstrapping on these model comparison values reveals that this difference is not driven by a small number of participants but is consistent across participants (*SEM* on BIC preference: ±291; $t[299] = 24.141$, $p < 0.001$). Thus, the assumptions that all displays at the same set size share a set of parameters and are interchangeable appears to be untenable. In other words, a set of parameters describing each individual display separately (despite the complexity of this model) is preferred over an item-based model that assumes a single capacity and precision of representation constrains performance in all displays.

To demonstrate that these model comparison techniques are correctly penalizing the more complex model, we can examine what happens if we shuffle the data—that is, if we randomly reassign participants' errors to different displays than they originally came from. This shuffling procedure should cause all displays to share the same parameters. In fact, after shuffling, model comparison techniques strongly prefer the model with only three parameters rather than the model that treats each item as requiring its own parameters (BIC difference: 4,118 in favor of the simpler model), suggesting such techniques correctly penalize the more complex model. This validity demonstrates that the true, unshuffled data provide strong evidence against treating all displays interchangeably.

Variance across items within a display could be accounted for in existing item-based models, because these models can posit of trade-offs in which only a subset of items are encoded or given more resources (e.g., participants might systemically choose to encode blue items more than red, which, in fact, there is some evidence participants do: Morey, 2011). However, even collapsing responses across all three items in a display reveals a large amount of variance between displays (see Figure 2 for some examples), with capacity estimates ranging from 1.8 to 2.9 and *SD*s ranging from 12.8° to 32.5°. This demonstrates the failure of any model that treats displays as a whole as exchangeable, even if items within each display are not exchangeable. For example, resource models that assumes a trade-off between which items get more resources within a single display, but a fixed pool of resources across displays (Bays & Husain, 2008) or slot models that assume a fixed number of slots must be used to represent each display (Zhang & Luck, 2008) cannot account for this reliable variance between displays as a whole.

High variance between displays at Set Size 3 accounts for a significant fraction of variable memory precision

Recently, some models have allowed for the possibility that memory is resource-limited and the amount of the resource might vary across trials (Fougnie, Suchow, & Alvarez, 2012; van den Berg et al., 2012). In particular, they have shown that participants remember items with different precisions on different trials (van den Berg et al., 2012) and even for different items on the same trial (Fougnie et al., 2012).

There are two possible explanations for variability in participants' precision. One is that it results from variability within participants (either within trials or across trials). However, resources that vary because of internal states of the observer or stochastic noise at encoding or maintenance are insufficient to explain the systematic differences we find across displays, because the differences we find are consistent across participants within a display but not consistent across

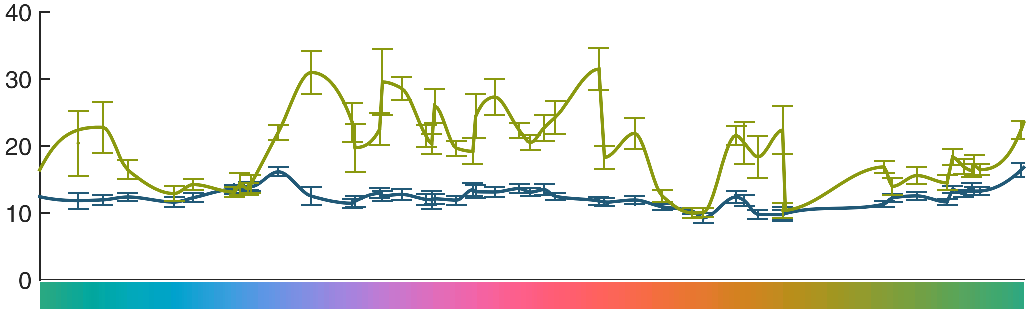

Figure 5. *SD* across the color wheel, from Set Size 1 experiment (blue) and the exact same colors in the Set Size 3 experiment, where they have varying contexts from the other two items on the display (green). Errors bars represent ±1 *SEM*, calculated by bootstrapping.

displays. Another possibility is that the number of resources available varies based on which exact configuration of colors is present on the display (either because of interactions between item's representations or because these models are incorrect in treating all memory as item-based, and some memory resources are actually used for encoding ensemble structures). This would be consistent with our finding of large variability between items and displays that is systematic across participants. To examine the relative contributions of these two sources, we analyzed our display-by-display data to look at how much variability was accounted for by across-display factors.

In particular, to demonstrate the variability in precision, van den Berg et al. (2012) showed that the standard model fits fail to account for the shape of the distribution of participants' responses; rather than being a von Mises distribution, participants responses are instead "peakier" than a von Mises distribution, consistent with a mixture of von Mises distributions with different precisions. To visualize this peakiness, they show the residuals of the data after taking into account the model fits. The residuals clearly indicate the models are not peaky enough to account for the data.

How much of this variability (e.g., the underrepresentation of near-the-target responses in the residual) can be accounted for by the reliable-across-participants differences we have observed? Figure 4 shows the residuals for the standard, aggregate model fit to all displays, and the residuals when fitting a separate model to each item on each trial. The residual is notably smaller when taking into account the variability across items that is consistent across participants.

In particular, the sum of squared residuals is 31% as large when fitting each display independently. This suggests that more than two thirds of the variability reported by (van den Berg et al., 2012) is reliable across participants (e.g., can be captured by modeling particular items from particular displays independently, rather than in aggregate). Approximately one third of this variability thus appears to be independent of the stimuli shown, perhaps due to variations in degradation over time (Fougnie et al., 2012).

Thus, a large portion of the variability in particular items' fidelity in memory appears to be reliable across participants because it is a function of the colors on the display and their relationship to each other.

Variance between displays results in different parameter estimates in item-based models

We find that model parameters fit to data aggregated across displays is reliably different than the average of parameters fit to individual displays. This is important because models of working memory make specific predictions as to how model parameters will vary with set size (e.g., Zhang & Luck, 2008), but the standard analysis method aggregates data across displays. We find that the standard deviation of participants' reports is significantly overestimated by this standard analysis technique at Set Size 3 (fit to aggregate: $SD = 22.5$; average of fits to individual displays: 20.0; overestimate: $t[143] = 4.71$, $p < 0.001$). Furthermore, in our control experiments run at Set Size 1 (fit to aggregate: $SD = 14.1$; average of fits to individual displays: 12.4; overestimate, $t[47] = 8.24$, $p < 0.0001$), *SD* is also overestimated. However, at Set Size 6, this same overestimate does not occur (fit to aggregate: $SD = 25.7$; average of fits to individual displays: 25.4; no overestimate: $t[287] = 0.41$, $p > 0.10$), and the overestimation at Set Size 3 is reliably greater than the overestimation at Set Size 6 ($t[430] = 2.69$, $p = 0.007$). This suggests that analyzing individual items and displays can change inferences even within item-based models like those used by Zhang and Luck (2008) to argue for a plateau in memory fidelity between Set Size 3 and Set Size 6. This suggests that the variance we observe in particular displays is meaningful for comparing working memory models, even when those models are fit to the average across all displays. Ignoring item and display-based effects can systematically distort the average precision and guess rates, and can do so differently at different set sizes.

This variance across items and displays is not substantially influenced by differences in precision for individual colors

There is some evidence to suggest participants are consistent in which colors they report most precisely even in single item displays (Bae, Olkkonen, Allred, Wilson, & Flombaum, 2014), as well as which items they tend to encode from particular displays (Morey, 2011). This could reflect reliable variance in how precisely participants perceive different colors or reliable variance in how likely they are to pick different colors from the response wheel.

However, there are several reasons to believe such effects do not account for the majority of the variance we demonstrate here. First, reliable differences in precision for particular colors should impact our estimates of *SD*, but not p_{mem}, yet we find significant variance across items and displays in both *SD* and in p_{mem}. Second, such effects are relatively small—much smaller than the large variance we observe across displays. To quantify this, we can examine our control experiment where we had 300 participants perform a memory task at Set Size 1. We measured performance at Set Size 1 using a subset of the colors used in our main Set Size 3 experiment (one from each display). As shown in Figure 5 (blue line), we find some inhomogeneity across the color wheel in how precisely participants represent colors at Set Size 1, but this variance is relatively small and confined to a small number of colors. In addition, this variance in *SD* is an order of magnitude smaller than the variance in *SD* we observe in our main experiment at Set Size 3 (green line), and the two are relatively uncorrelated, with the variance in *SD* by color in Set Size 1 accounting for only 5.0% of the variance we observe at Set Size 3.

Indeed, Bae et al. (2014) also found that such effects get smaller and less reliable with larger set sizes, perhaps reflecting the fact that constraints on memory capacity and additional contextual and ensemble factors begin to dominate any effects of individual colors at higher set sizes.

Thus, while there are reliable differences in the precision with which participants represent particular colors and/or reliable differences in their propensity to pick particular colors during the response window (e.g., Bae et al., 2014), these effects account for very little of the variance we observe across displays at Set Size 3. Instead, the relationship of the colors on the display to each other (e.g., perceptual grouping, ensemble effects, etc.) appears to be the primary determinant of this variance between items and displays at Set Size 3.

The variance across items and displays is not due to swaps

We tested whether an alternative item-based model, the swap model of Bays et al. (2009), could account for the variance across items and displays. This model posits that error distributions include some proportion of misreports of the wrong item from the display. Such swaps would certainly lead to display-specific and item-specific error distributions. However, we find that this swap model does not account for a significant proportion of the current data. These swaps, when a model including them is fit across all displays, are estimated to occur less than 2% of the time. If such a model is fit to the displays individually, it is apparent that swaps are estimated to occur only on displays where the items are extremely nearby in color space, consistent with the possibility these putative swaps are in reality systematic shifts in the color reported (towards other colors in the display) rather than true swaps. A similar conclusion—few swaps, and most of them the result of nearby items in color space—holds in our control experiment at Set Size 6 (see Appendix).

While some working memory experiments likely result in many swap errors (e.g., when items are closely spaced and/or appear in unpredictable locations; Emrich & Ferber, 2012), the current experiment was designed to minimize the chance of such errors: Only three items are presented, in stable and predictable locations across trials, and these locations are widely spaced. Furthermore, sufficient encoding time is given (1000 ms). All of these factors likely result in few swap errors (e.g., Bays et al. 2009; Emrich & Ferber, 2012).

The variance across items and displays is not due to simple grouping or chunking

We find reliable variance in the precision and likelihood of remembering across both individual items and entire displays, and this variance does not appear to be due to swaps of items with other items. One possible explanation is thus an all-or-none grouping or chunking strategy. For example, a model where participants can sometimes represent multiple items as a single unit or chunk, as in cases where Gestalt grouping makes participants see two items as one (Xu, 2006; Xu & Chun, 2007). Such chunking or grouping models have not been formalized, but they nevertheless make specific predictions that are not supported by the current data.

Specifically, if participants are in fact storing multiple items as a single chunk, then items that are similar should tend to be reported as the exact same color. To examine this, we looked at all pairs of items on the same display that were less than 20° apart in

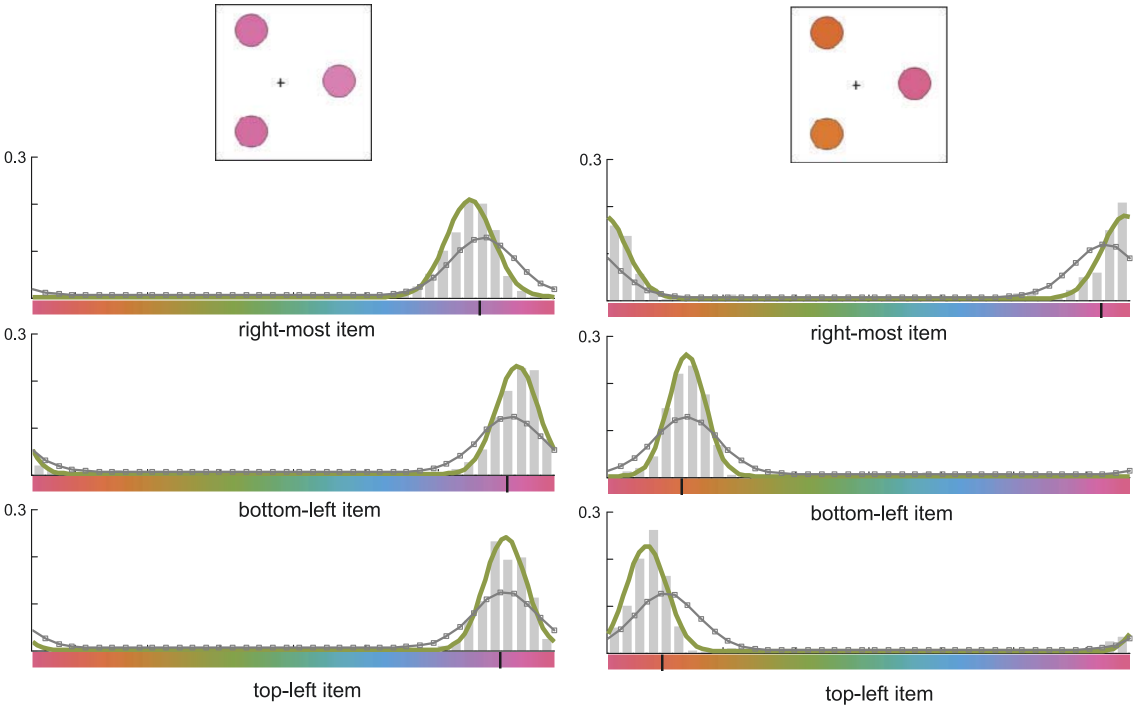

Figure 6. Two examples of displays with similar items, where we would expect chunking to be most prevalent. For each display, the responses of participants are shown as a histogram (gray bars) and the correct color is shown by a vertical black line on the color bar underneath. A mixture model fit to each item's responses individually is shown in green. The gray line with squares marks the average mixture model fit across all items in all displays. Differences between the green line fit to the histogram and the gray line reveal ways that responses to these items are different than the average item. Notice that even for items that are very similar (e.g., the middle and bottom row on both sides) participants responses are very clearly distinct; the distributions for the two items are not identical but are properly shifted so they are biased toward the correct response in each case.

color space and asked how often participants responses for these items tended to be nearly identical. We operationalized "nearly identical" as the responses being <5° apart (given this is approximately the noise usually seen when matching colors present on the screen; Brady, Konkle, Gill, Oliva, & Alvarez, 2013). The chunking account predicts that when participants remember the items successfully, they should respond nearly identically for the two items, as they have stored only a single color representing both locations. However, this will only be true for items that were remembered successfully; times where the chunk was forgotten (e.g., guess responses) will result in many responses further apart than this. Thus, we examined only relatively accurate responses—those within 20° of the correct answer for both items. We find that, on average only 36.1% of these accurate responses are within 5° of each other, whereas 38.5% of the actual item pairs considered are within 5° of each other. Thus, participants do not have a tendency to report items as nearly identical when they are similar but distinct.

Example displays are shown in Figure 6, along with their responses, to visualize this in particular displays that feature similar but distinct items. Note that the histogram of responses is clearly distinct even for items that are very similarly colored.

This conclusion is not substantially affected by the particular cut-offs we chose: Considering only responses within 10° of the correct answers (to further ensure this is not affected by guessing) gives qualitatively the same result, with 34.0% of responses being within 5° of each other. If we operationalize "nearly identical to" in a broader way, and examine the proportion of responses within 15° of each other, we find that 64.1% of participants' responses to these items were within 15° of each other, as compared to 53.8% of the actual item pairs. This remains inconsistent with an account where participants always report these similar items as the same exact color.

Participants represent more than just individual items

Existing work has shown the presence of ensemble/texture representations (Alvarez, 2011; Brady & Alvarez, 2015; Haberman & Whitney, 2012; Rosenholtz et al., 2012) and shown the influence of these ensemble representations on individual item representations

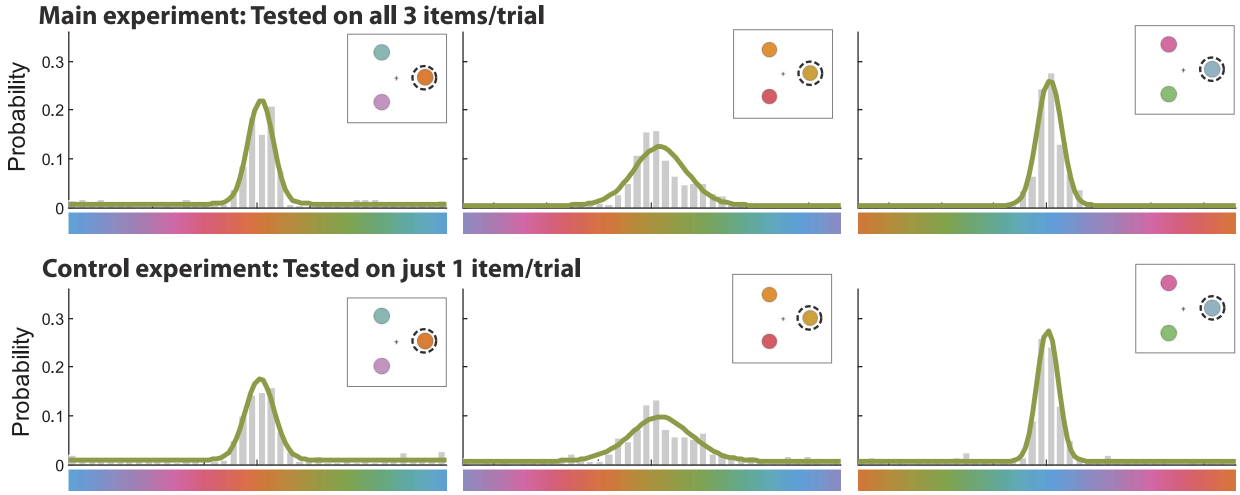

Figure 7. Example error distributions for three items, when tested in the main experiment (top), and in the control experiment where only one item is tested per trial (bottom). Responses are nearly identical whether participants are tested only a single item or on all three items. Note that the tested item is always shown on the right, but in the actual experiment, different participants saw the display with the three items in different locations.

(Brady & Alvarez, 2011; Brady & Tenenbaum, 2013). In the current data, we find evidence of such ensemble representations even when participants don't seem to possess any individual item memory. In particular, participants seem to be aware of ensemble properties like the variability of the colors on the display (e.g., whether all the colors similar or all different) even when they fail to remember any of the individual item colors. If we examine only trials where the colors participants' reported were entirely wrong for every item (not within the correct one fourth of the color wheel; i.e., error >45° for each item), we still find a significant correlation between the variability of the colors participants' report and the variability of the items on the display ($r = 0.40$; $p = 0.004$). Thus, when all three colors are relatively similar, participants report similar colors for all three, even if they are entirely wrong about the particular colors; when all three colors are different, participants report colors that are further apart in color space, and this is true even when participants do not remember any of the individual colors. This provides direct evidence for ensemble representations separately from individual item representations. In addition, this provides strong evidence against chunking and grouping as the only mechanism causing variability across items and displays, since these accounts do not predict any direct representation of ensemble properties or summary statistics.

The variance across displays is not caused by asking participants to report all of the items

One possibility is that the nonindependence between items could be an artifact of the task we had participants perform. In particular, participants were asked to report all the items in each array, rather than just one item (as is usually done). This may have caused participants to adopt a global or holistic strategy to encode all the items. Furthermore, it is possible that retrieving one item at test may have biased subsequent reports, for example, through proactive interference or priming.[1]

Thus, we ran an additional control experiment. This experiment was identical to the main experiment, except that participants were asked about only a single item from each display. We can then examine whether participant's error distributions are similar for the main experiment as for this control experiment, to ask whether these concerns might have caused the nonindependence we observer between items.

We find that participants' error distributions are nearly identical when tested on only one item (see Figure 7 for examples). In addition, fitting the standard model parameters to each of the 48 items tested here and comparing them with the same 48 items in the main experiment shows correlations of $r = 0.98$ ($p < 0.0001$) for *bias*, $r = 0.90$ ($p < 0.0001$) for *SD*, and $r = 0.76$ ($p < 0.0001$) for p_{mem}. These values are comparable to the reliabilities estimated within each experiment ($r = 0.98$, 0.87, and 0.86, respectively). Thus, there does not appear to be any effect of testing multiple items on participants' error distributions.

A structured representation model of working memory explains some aspects of performance

While item-based or chunk-based models cannot account for the variability across items and displays, a

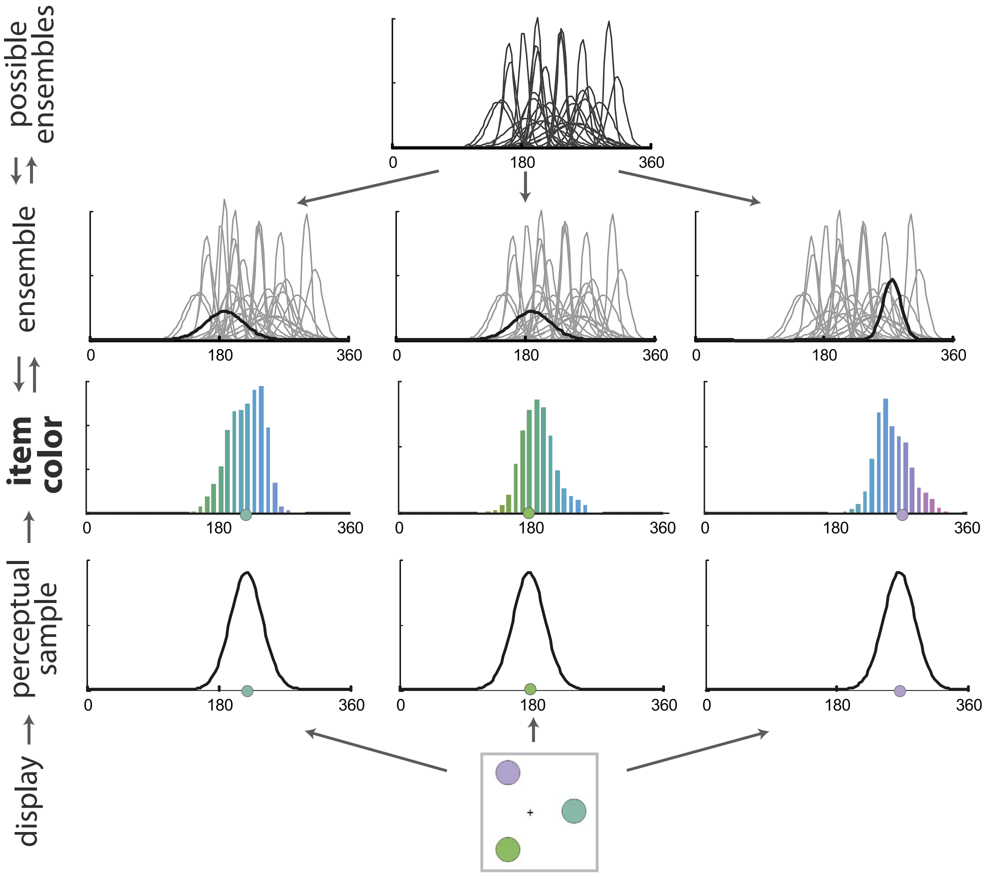

Figure 8. Schematic of the structured representation model. Participants see a display (bottom row) and encode noisy samples of the colors present (perceptual samples; second row from bottom). Participants then attempt to infer, based on these samples, the true colors the items were (item color; middle row). To do so, they make use of not only the samples they encoded, but are influenced, by potential ensemble groupings of the colors (top row). Thus, the colors that participants report are influenced not only by the sampled colors, but also by structure of potential clusters that could be present in the display. For example, both the first and second item might be seen as coming from a particular ensemble group with a particular center color and variance (highlighted in black; second row from top), in which case both this ensemble information and the sample that was encoded will inform the color that participants' report. Alternatively, all three colors might be seen as coming from a single high variance ensemble, and then all three colors would be pulled towards each other. Importantly, the model posits that participants do not simply choose a single ensemble that could explain the display, but instead integrate over all possible ensembles in making their response.

model based on structured individual item and ensemble representations *can* successfully explain some of the variance in participants' performance (Brady & Alvarez, 2011; Orhan & Jacobs, 2013). In particular, the data are compatible with a model where participants treat the display as being made up of clusters of items, representing both individual item information and the mean color and variance of these clusters in a single structured representation (as in Brady & Alvarez, 2011).

Within this framework, item-level and cluster-level representations are not independent, but are instead integrated. Consequently, the structured representation model predicts a shift in the representation of each item toward the mean of the other items represented in the same cluster, as well as different precisions (*SD*s) for items that cluster well with other items versus those that do not (Brady & Alvarez, 2011; Orhan & Jacobs, 2013). A similar hierarchical representation is known to be present in memory for visual size and in spatial memory (Brady & Alvarez, 2011; Orhan & Jacobs, 2013); and such representations can be considered optimal under certain conditions, since they make use of information from all of the items to inform judgments for any particular item (Brady & Alvarez, 2011; Orhan & Jacobs, 2013). Furthermore, such models have explicit representations of ensemble properties like the mean and variance of a cluster and of the entire set of items on the display, which is consistent with our finding that participants sometimes retain this information in the absence of specific item memories (e.g., know the variance of the colors even when they do not remember the actual colors).

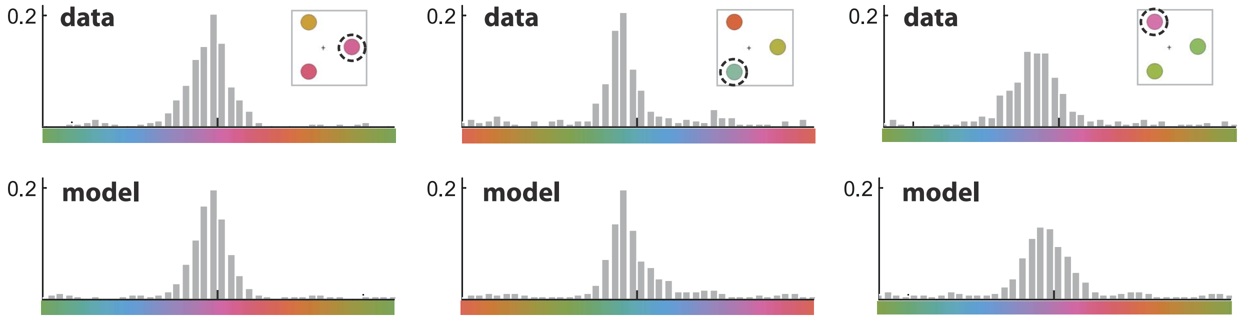

Figure 9. Predictions for three items from different displays. Notice that on each kind of display—displays where the item groups well (left), a display where no items group (middle), and a display where an item is an outlier relative to the others (right)—the model correctly predicts shifts in the mean and the direction of the skew, in addition to the width of the distribution and the rate of nontarget related responses.

How do participants decide which items to cluster together? In the case of spatial memory, it has been suggested that, rather than picking a single clustering of the display, the visual system infers a probability distribution over many such clusterings (Orhan & Jacobs, 2013). This allows the representation of items at multiple levels of abstraction simultaneously, and results in biases toward the means of all possible clusters to which an item might be assigned. We propose a similar model can explain important variance in visual working memory for colors. Thus, in estimating the color of the set of items on the display, participants would integrate out their uncertainty about the clustering of the set of items, weighting each possible ensemble representation by how well it would explain the colors that are present, but ultimately considering many such ensembles in their response (see Figure 8).

We formalized an implementation of this model and examined its fit to the data. The details of the model are outlined in the Appendix. We allowed three free parameters to vary: how noisy participants initial encoding of the colors was (taking into account both perceptual and memory noise), equivalent to *SD* in the item-based models; the concentration parameter, which captures how likely different items are to come from the same cluster as each other (low values: more clustering; high value: less clustering); and the rate at which a particular cluster of items is corrupted or lost from memory (equivalent to p_{mem} in the item-based models). The best-fit parameters were equivalent to an *SD* of 22.5° and a p_{mem} of 0.70, with a concentration parameter suggesting mild clustering ($C = 15$). Note that while the proposed memory representation is hierarchical, and there are many parameters in the model whose value depend on the particular colors present on a given display, the only parameters that are not 100% determined by the colors on the particular display are the *SD*, p_{mem}, and C parameters. Thus, when fit to the data of a given participant or set of participants, the model is not hierarchical in the traditional statistical sense (e.g., Morey, 2011). Instead, it is simply a model that makes distinct predictions on different displays depending on the colors present on those displays, with all predictions yoked to the set of three parameters that are free to vary to fit to the data.

The intuition behind the model is expressed in Figure 8. Rather than thinking of the entire model, consider a single example first. For instance, imagine that you remember each item both as an individual and as part of a particular cluster, with a particular mean and variance (as represented by the ensemble representation with a darker line in the second row of Figure 8). What is the optimal inference about the color of the item, if you have some belief about what color the item was from your perceptual encoding of that item (with noise, based on the *SD* parameter; fourth row), but you also believe the item to be drawn from a particular higher level cluster (second row)? It turns out that, according to Bayes rule, the optimal response is not to report a color based purely on your perceptual memory, but to bias your response toward the ensemble mean (e.g., Brady & Alvarez, 2011; Huttenlocher, Hedges, & Vevea, 2000). For example, if you encoded the left item in Figure 8 as blue/green with uncertainty (*SD*), but you also encoded that this item is in a cluster with the green item, then it is more likely that the item was actually greener than your perceptual estimate (as opposed to bluer than your perceptual estimate). Thus the optimal response distribution is centered not around the center of the individual item distribution, but is centered around a weighted average of this individual item memory *and* the associated ensemble, with the ensemble given more weight when the individual item memory is noisy (e.g., *SD* was high) or when the ensemble was itself narrow (e.g., the cluster you encoded was very tight, as if both items were nearly the same color).

This example provides the intuition behind the structured representation model. However, in the actual model, items are not assigned a single ensemble representation. Instead, many possible ensemble distributions are considered and are weighted based on how well they explain the items on the display. For example, if all three items are pink, an ensemble representation with a center on green would be very unlikely; an ensemble with low variance centered on pink would be much more likely; but a slightly wider ensemble representation with a center on more red would also be possible, etc. The *C* parameter controls how likely multiple items are to be drawn from the same ensemble representation as each other. If this value were close to 0, there would be a strong preference for ensemble representations where all three items are drawn from the same ensemble; if it were very high, there would be a strong preference for ensembles where each item is assigned a unique distribution (and thus there is very little pull toward the other items).

The middle row of Figure 8 shows the kind of item color distributions that result from this model. The ways in which these distributions differ from the perceptual encoding distributions in the fourth row is based on pulls toward the different possible ensemble representations, weighted by how likely those ensemble representations are. For example, the middle column shows the green item. The distribution for this item has its mode on roughly the correct answer, but all of its skew is toward the blue/purple side. This is because the only ensembles that are high likelihood on this display account for either only this item (and thus result in little skew), or result in clustering this item with either the blue and purple item or just the blue or just the purple item, which in all cases results in skew in that direction.

This model provided a strong qualitative fit to the data (see Figure 9). In addition, model comparison metrics showed that this model provided a vastly better explanation of the data than the standard model that assumed all displays were interchangeable despite having the same number of free parameters (BIC difference: 4,401; *SEM*: ±308; $t[299] = 14.3$, $p < 0.001$). The predictions of this model are shown in Figure 8 on particular individual displays. To ensure that this model was not subject to overfitting, we collected an entirely new set of data with 300 new subjects, using the same displays as before but with the items rotated 180° in color space (e.g., red items became green). The same model with the exact same parameter values derived from the previous data provided a very strong fit. This zero parameter model easily outperformed the model that treated all displays interchangeably (BIC difference: 3,630) on this independent dataset.

The fit of this model provides some evidence that participants' representations are multifaceted, composed of individual item representations that are modulated by ensemble statistics (Brady & Alvarez, 2011). In particular, if we wish to fit models that provide strong explanatory value—for example, reduce the space of a complicated set of data to just a few simple principles and a few parameters—then this model provides a significant improvement over fitting a standard mixture model that assumes the same parameters hold on each individual item and each individual display.

In addition, this model explicitly represents both a cluster's mean and cluster's variance. Thus, this model has the ability to explain why participants sometimes forget all of the individual colors in a display but nevertheless retain the information about the variance of the items (e.g., ensemble information). In particular, if the cluster means become corrupted but not the cluster variances, then the model predicts the report of incorrect colors, but the correct retention of how spread out on the color wheel these items are.

How good of an explanation is this structured representation model?

The structured representation model we propose provides a better fit than the standard model that assumes that all items on all displays share a single precision and guess rate, even with the same number of parameters fit to the data. How does this model compare to a model that assumes all displays are independent and all items have entirely separate *bias*, *SD*, and p_{mem}?

The completely independent model, with 432 parameters, actually outperforms the structured representation model (BIC difference: 2,693; *SEM*: ±307, $t[299] = 8.8$, $p < 0.001$). This suggests that, while the structured representation model is an improvement over the standard model that assumes the displays and items are all interchangeable, it is still far from a completely satisfactory explanation of the data. However, given that the model with 432 parameters provides little to no explanatory value—it does not reduce the space of a complicated set of data much at all, instead simply providing a complicated restatement of the data—the structured representation model may still be considered an improvement, just as the standard mixture model provides a useful simplification of the data. In particular, the structured representation model provides just a few simple principles and parameters, and this model provides a significant improvement over fitting a standard mixture model that assumes the same parameters hold on each individual item and each individual display.

What makes the structured representation model fail to out-fit the model that assumes representations are entirely independent for each item? Some factors are likely to be systematic: For example, there is some evidence that items that are very close by in color space actually repel each other, rather than attracting each other, possibly because they are explicitly coded as distinct (e.g., Johnson et al., 2009). There is some tendency for this in our data (e.g., see Figure 6) and it is directly the opposite of the prediction of the hierarchical attraction in the structured representation model.

Other factors are likely to be more idiosyncratic. In particular, there may be factors that cause variance across items and displays in ways that are very difficult to model (e.g., particular color combinations may have semantic meaning, causing attraction to pre-existing prior states, like red–green–blue).

General discussion

The current results demonstrate a failure of item-based and chunk-based models to account for data from individual working memory displays. In particular, these data show that performance on working memory tasks is strongly affected by interactions between items within a display, and provide evidence that higher level ensemble properties are explicitly encoded even in simple color memory displays (e.g., the variance of colors on a display). These results provide evidence against influential item-based models that predict (or assume) no variability across displays, including recent slot models and resource models of visual working memory capacity. In fact, the present results suggest that individual object representations may compose only a part of working memory representations, and suggest that hierarchically structured representations may play an important role in memory, even for the canonical color memory task. We propose that these findings require a new framework for studying visual working memory, focusing on structured memory representations composed of multiple, interacting levels—including both individual-item information and interitem ensemble information.

Implications for estimating the capacity of working memory

The present finding suggests that estimates of working memory capacity must take interitem effects into account in order to accurately estimate item limits. Our analyses suggest that current capacity estimates, which are generally derived from item-based models and which assume 100% of participants' responses are based on individuated item representations (e.g., Bays et al., 2009; Zhang & Luck, 2008), are likely overestimating capacity for individual items. In particular, even displays of three colors tend to have some structure in them, which people can use to remember the colors, and that could be seen as a way for participants to inflate their capacity estimates beyond the core capacity of their visual working memory systems, in the same way that all-or-none chunking is seen as inflating capacity estimates (e.g., Cowan, 2001). For example, Cowan (2001) advocated using nonsense stimuli and eliminating the ability to chunk (thereby combining information about multiple items) to get at the core capacity of the system. However, if even displays of three colors show some structure, then, if this logic is correct, we would need to look at the least structured displays—those that reveal the lowest capacity—to show the true, core capacity of the visual memory system.

Thus, examining p_{mem} in the displays with the lowest capacity estimates should provide an estimate of how many individual items can be remembered independent of grouping and ensemble factors that may help participants remember information in some displays by combining information sources across items. For Set Size 3 displays, the lowest reliable capacity was only 1.8 items, as compared to the mean across displays of nearly 2.6 items—a 44% overestimate. For Set Size 6 displays, the lowest reliable capacity was 1.3 items, compared to the mean across displays of 2.5 items. These low capacity displays (visible in Figure 3), tend to have the least structure and fewest relations between items.

Thus, our data indicate that existing models of working memory capacity overestimate the capacity of memory for individual items by confounding all encoding strategies participants use (ensembles, grouping, individual items) with an estimate of how many individual items they can remember. When display structure is at a minimum, estimates of individual item capacity are nearly 50% less than estimated by slot and resource models. Note, however, that this estimate is not a meaningful measure of the total amount of information participants can store, but only their capacity for individuated objects; participants are able to remember significantly more on average because their representations appear to be much richer than just individuated objects (for example, they appear to independently store the variability of the items in color space, as a kind of ensemble representation).

Implications for working memory architecture

Recent theories of working memory have hinged on accurately estimating how many items can be remembered, and the precision with which they are stored (for a review, see Brady et al., 2011). In particular, much work has focused on how the precision of item representations changes with set size. In addition to showing that capacity is overestimated by the standard analysis method, our data also provide evidence that fitting the aggregate data across displays, as is typically done, systematically misrepresents the parameters of participants' memory representations in a way that undermines current claims about cognitive architecture.

For example, many recent papers have debated whether the precision of memory plateaus at Set Size 3 (i.e., whether an identical *SD* is present at Set Size 3 and Set Size 6), because such a plateau is taken to support a slot model of memory (Bays et al., 2009; Van den Berg & Ma, 2014; Zhang & Luck, 2008, 2011). However, we find that fitting separate models for each item significantly changes the conclusions about how fidelity changes with set size. In particular, such an analysis shows that the standard deviation of participants' reports is significantly overestimated by the standard analysis technique at Set Size 1 and Set Size 3, but not Set Size 6. Thus, at least part of the reason that recent papers have found similar or equivalent standard deviations at Set Size 3 and Set Size 6—a fundamental claim of the slot model (van den Berg & Ma, 2014; Zhang & Luck, 2008)—is that variability across displays leads to a greater overestimation of the standard deviation at lower set sizes than Set Size 6. Thus, our data cast serious doubt on claims that participants' representations do not continue to get noisier after Set Size 3 (Zhang & Luck, 2008), and strongly undermine this particular piece of evidence for a simple slot-like architecture for working memory, even if we continued to (incorrectly) assume that all of participants' performance results from individual item representations.

Beyond individual items, and towards structured representations

Most research on visual working memory attempts to minimize the role of individual-display variability so that it can be ignored when modeling performance. However, this approach can only be successful if it is possible to control for factors that might vary across displays, and if the controlled factors are actually irrelevant to models of working memory capacity. We argue that both of these requirements are untenable. For instance, we have shown that the standard practice of generating random displays of simple items results in substantial individual-display variability, and that parameters derived from models of aggregate data can be misleading. Furthermore, it is either difficult or impossible to manipulate key variables, such as the number of items to remember, while holding interitem factors completely constant. That is, it would be nearly impossible to generate displays such that a six-item display has the same amount of perceptual organization and the same utility of ensemble representations as a one- or two-item display. Thus, if these interitem factors are not taken into account, it is impossible to isolate changes in performance that are due to changes in set size alone.

Even if contextual factors could be controlled to minimize their contribution to performance, doing so presupposes that their role is irrelevant in the encoding and storage of information in working memory. The results shown here, and in previous studies (Brady & Alvarez, 2011, 2015; Brady & Tenenbaum, 2013), support the proposal that visual working memory representations are in fact richly structured, composed of both texture/ensemble features and individual item features, in addition to the utility of all-or-none perceptual grouping (Xu & Chun, 2007). Therefore, to fully understand our ability to hold information in mind, it is more important to understand how memory maintains structured information than to estimate the number of individual objects that can be remembered.

Based on our findings that interitems factors play an important role in working memory storage, we propose a new framework for studying visual working memory in particular, and visual cognition more broadly, which employs modeling individual-display variation to constrain models of cognitive function (as in item response theories and similar attempts to estimate item effects: e.g., Baayen, Davidson, & Bates, 2008; Lord, 1980). This approach is analogous to the use of individual-person variation to constrain cognitive theory (Peterzell & Teller, 1996; Vogel & Awh, 2008; Wilmer & Nakayama, 2007; Yovel & Kanwisher, 2005), and many of the same techniques and inferences can be drawn using an individual-display approach.

In the current work, we have examined primarily variance across displays, and, in doing so, we averaged across participants. However, we ultimately wish to understand not only variance in particular displays, but also how this variance is effected by the variance across participants (as in Vogel & Awh, 2008). Doing so will require hierarchical models (e.g., Morey, 2011) and mixed-effects models, which are popular when examining item effects in other areas of research (e.g., Baayen et al., 2008; see also item-response theory: Lord, 1980, and attempts to examine both items and participants: Clark, 1973).

The main advantage of this approach—whether averaging across subjects or fitting mixed-effects models—is that collecting large amounts of data on particular displays, and using simple modeling techniques to account for individual-display variability, allows more stringent tests of model predictions. The current study demonstrates the importance and utility of this method: We use it to show that there is substantial individual-display variation that has a major impact on the representations participants' form of simple working memory displays, and that working

memory capacity cannot be accurately modeled without taking into account the structure and use of ensembles in working memory representations. This provides strong evidence against all existing item-based models of working memory and opens many new questions about the nature of working memory representations.

Keywords: visual short-term memory, working memory capacity, ensemble statistics

Acknowledgments

Funding from NSF CAREER BCS-0953730 to G. A. A.

Commercial relationships: none.
Corresponding author: Timothy F. Brady.
Email: timbrady@ucsd.edu.
Address: Department of Psychology, University of California, San Diego, San Diego, CA, USA.

Footnote

[1] We thank an anonymous reviewer for this suggestion and for the suggestion of the control experiment.

References

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29, doi.org/10.1016/j.jecp.2009.11.003.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131.

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

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics: Efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, *106*, 7345–7350.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*(7), 622–628.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412, doi.org/10.1016/j.jml.2007.12.005.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Bae, G., Olkkonen, M., Allred, S., Wilson, C., & Flombaum, J. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision*, *14*(4):7, 1–23, doi:10.1167/14.4.7. [PubMed] [Article]

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10):7, 1–11, doi:10.1167/9.10.7. [PubMed] [Article]

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368, doi.org/10.1093/pan/mpr057.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392.

Brady, T. F., & Alvarez, G. A. (2015). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *41*(3), 921–929.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5):4, 1–34, doi:10.1167/11.5.4. [PubMed] [Article]

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, *24*(6), 981–990.

Brady, T. F., & Oliva, A. (2012). Spatial frequency integration during active perception: Perceptual hysteresis when an object recedes. *Frontiers in Psychology*, *3*, 462, doi.org/10.3389/fpsyg.2012.00462.

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

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5, doi.org/10.1177/1745691610393980.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466, doi.org/10.1016/S0022-5371(80)90312-6.

Emrich, S., & Ferber, S. (2012). Competition increases binding errors in visual working memory. *Journal of Vision*, *12*(4):12, 1–16, doi:10.1167/12.4.12. [PubMed] [Article]

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Neuroscience*, *8*(12), 1643–6, doi.org/10.1038/nn1608.

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12):27, 1–11, doi:10.1167/10.12.27. [PubMed] [Article]

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229, doi.org/10.1038/ncomms2237.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201, doi.org/10.1038/nn.2889

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, *17*(5), 673–679.

Godden, D., & Baddeley, A. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224, doi.org/10.1002/bdm.1753.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753.

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

Howard, M., & Kahana, M. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision*, *10*(2):24, 1–27, doi:10.1167/10.2.24. [PubMed] [Article]

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology, General*, *129*(2), 220–241.

Im, H. Y., & Halberda, J. (2012). Accurately modeling Visual Working Memory performance at the individual trial level. *Journal of Vision*, *12*(9):858, doi:10.1167/12.9.858. [Article]

Jiang, Y. V., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*(3), 683–702.

Johnson, J., Spencer, J., Luck, S. J., & Schoner, G. (2009). A dynamic neural field model of visual working memory and change detection. *Psychological Science*, *20*(5), 568–577.

Kahana, M. J., Zhou, F., Geller, A. S., & Sekuler, R. (2007). Lure-similarity affects visual episodic recognition: Detailed tests of a noisy exemplar model. *Memory & Cognition*, *35*(6), 1222–1232.

Kubovy, M., & Pomerantz, J. (Eds.). (1981). *Perceptual organization*. Hillsdale, NJ: Erlbaum.

Lin, P., & Luck, S. (2009). The influence of similarity on visual working memory representations. *Visual Cognition*, *17*(3), 356–372.

Lin, P., & Luck, S. (2012). Proactive interference does not meaningfully distort visual working memory capacity estimates in the canonical change detection task. *Frontiers in Psychology*, *3*, 42.

Lord, F. M. (1980). *Applications of item response to theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281, doi.org/10.1038/36846.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, *55*(1), 8–24, doi.org/10.1016/j.jmp.2010.08.008.

Nosofsky, R., & Kantner, J. (2006). Exemplar similarity, study list homogeneity, and short-term perceptual recognition. *Memory & Cognition*, *34*(1), 112–124.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 61–65.

Oliva, A. (2005). Gist of the scene. In *Neurobiology of Attention* (pp. 251–257). New York: Elsevier.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*(1), 72–107.

Olson, I., & Marshuetz, C. (2005). Remembering "what" brings along "where" in visual working memory. *Perception & Psychophysics*, *67*(2), 185–194.

Orhan, A., & Jacobs, R. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review*, *120*(2), 297–328.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian Stochastic Modelling in Python. *Journal of Statistical Software*, *35*(4), 1–81.

Peterzell, D., & Teller, D. (1996). Individual differences in contrast sensitivity functions: The lowest spatial frequency channels. *Vision Research*, *36*(19), 3077–3085.

Portilla, J., & Simoncelli, E. P. (2000). A Parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70, doi.org/10.1023/A:1026553619983.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4):14, 1–17, doi:10.1167/12.4.14. [PubMed] [Article]

Rouder, J. N., Morey, R., 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*, *105*(16), 5975–5979.

Simoncelli, E., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Suchow, J., Brady, T., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(10):9, 1–8, doi:10.1167/13.10.9. [PubMed] [Article]

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

Tulving, E., & Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *60*(5), 352–373.

Van den Berg, R., & Ma, W. J. (2014). "Plateau"-related summary statistics are uninformative for comparing working memory models. *Attention, Perception, & Psychophysics*, *76*(7), 2117–2135.

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*, *109*(22), 8780–8785, doi.org/10.1073/pnas.1117465109.

Victor, J. D., & Conte, M. M. (2004). Visual working memory for image statistics. *Vision Research*, *44*(6), 541–546.

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):8, 244–256, doi:10.1167/5.3.8. [PubMed] [Article]

Viswanathan, S., Perl, D., Visscher, K. M., Kahana, M. J., & Sekuler, R. (2010). Homogeneity computation: How interitem similarity in visual short-term memory alters recognition. *Psychonomic Bulletin & Review*, *17*(1), 59–65.

Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, *17*(2), 171–176, doi.org/10.1111/j.1467-8721.2008.00569.x.

Wilken, P., & Ma, W. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12):11, 1120–1135, doi:10.1167/4.12.11. [PubMed] [Article]

Wilmer, J. B., & Nakayama, K. (2007). Two distinct visual motion mechanisms for smooth pursuit: evidence from individual differences. *Neuron*, *54*(6), 987–1000, doi.org/10.1016/j.neuron.2007.06.007.

Woodman, G. F., Vecera, S. P., & Luck, S. J. (2003). Perceptual organization influences visual working memory. *Psychonomic Bulletin Review*, *10*(1), 80–87.

Xu, Y. (2006). Understanding the object benefit in visual short-term memory: The roles of feature proximity and connectedness. *Perception & Psychophysics*, *68*(5), 815–828.

Xu, Y., & Chun, M. M. (2007). Visual grouping in human parietal cortex. *Proceedings of the National Academy of Sciences*, *104*(47), 18766–18771.

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*(24), 2256–2262, doi.org/10.1016/j.cub.2005.10.072.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory. *Psychological Science*, *22*(11), 1434–1441, doi.org/10.1177/0956797611417006.

Appendix

Extended swap model fits

One simple model that could account for display-by-display variance is one in which participants sometimes report incorrect items from the same display. It is visually apparent in our data (e.g., Figure 2) that this does not account for the majority of the difference in individual items and displays in the current dataset. However, it is also possible to analyze such swaps quantitatively by using a model that takes into account that participants might sometimes report an incorrect item (Bays et al., 2009; see Methods for a description of this model).

The quantitative fit confirms swaps are, at best, a very small component of the difference between items and displays. Fitting to the aggregate data across all displays gave us an estimate of 1.6% for $p_{distractor}$, the likelihood of incorrectly reporting a distractor item. This is a very small rate, and thus not likely of significant relevance to our conclusions (when taking into account swaps, p_{mem} was estimated at 78%). In addition, fitting a swap model to individual displays reveals that only three items in the entire set of 144 items have an estimated swap rate of more than 15%, and they all come from the same display—a display in which all three items are nearly the same color (all are a shade of pink). On this display swaps cannot be distinguished from correct responses or slightly shifted correct responses, and thus all three items have a large degree of uncertainty in their parameter estimates between p_{mem} and $p_{distractor}$. Taken in the context of the results from other displays, these responses almost certainly reflect correct memories rather than swaps. Thus, there is little evidence for reporting of incorrect items in our data, or that it could explain the differences we observe between different displays.

According to Bays et al. (2009), participants commit many more swaps at larger set sizes (e.g., they find a nearly 30% swap rate at Set Size 6). In our control experiment at Set Size 6, we did find a higher estimate of swaps when fitting the Bays et al. (2009) model. However, we still find an average swap rate of only 12%, well below Bays et al.'s (2009) estimate of nearly 30% at Set Size 6.

In addition, in our data, these swap estimates are driven by a small number of items and displays. Using model comparison, we find that the swap model is preferred to a model without swaps for only 18.4% of our 288 items (six items each in 48 displays). As at Set Size 3, these displays tend to be one with many similar colored items (e.g., five orange/pink colors and one blue color). On these displays, participants often misreported the pinks as more orange and vice versa; and this is interpreted as evidence of swaps in the model comparison between a standard mixture model and a swapping model. However, participants almost never reported an orange/pink color as blue or vice versa (as predicted by a pure swap account). Thus, this data is not consistent with swaps between all items or swaps based on location. Instead, it may be the case that there are very few true swaps: The error distributions may simply be complex, involving clusters and hierarchical encoding, which the swap model can mimic in some circumstances. In particular, at higher set sizes, people may be considerably more likely to take a clustering approach, and remember a broad color category, particularly on where the items fall into only a few clusters.

Other experiments may find more true swap errors than we do here. In particular, the locations the items appear in this experiment are stable and predictable locations across trials and these locations are widely spaced. Furthermore, sufficient encoding time is given (1000 ms). All of these factors likely result in few swap errors (e.g., Bays et al. 2009; Emrich & Ferber, 2012).

Structured representations/hierarchical Bayesian model

The hierarchical Bayesian model we use to conceptualize structured representations was similar to the Bayesian Finite Mixture Model of Orhan and Jacobs (2013), with two sets of modifications: First, the distribution for sampling the data and the cluster centers were modified to be von Mises distributions rather than normal distributions, to account for the circular nature of color data; second, we added a new component to account for guessing, which allowed for participants' representations to sometimes become corrupted prior to being reported.

In particular, we fit a mixture model made up of $K = 5$ components. All parameters of the model were determined by the colors present on the displays themselves, as opposed to the data (e.g., participants' responses), except for three parameters, which we fit based on participants' error distributions: κ_{sample}, which controls the sampling error of individual items, and which we set equal to seven (equivalent to an *SD* of 22.5°); *C,* which is the concentration parameter which indicates how likely items are to cluster, which we set equal to 15; and *g,* which is the guess rate (the inverse of p_{mem}), which we set equal to 0.30. The role of these parameters is explained below.

The model is specified as follows, where *i* ranges from 1 . . . *N*, for the $N = 3$ items; and *j* ranges from 1 . . . *K*, for the $K = 5$ clusters. First, the propensity for items to be in certain clusters:

$$\pi \sim \text{Dirichlet}_K(\alpha_p)$$

$$z_i \sim \text{Multinomial}(\pi)$$

where π represents the mixing proportions of the clusters (e.g., how many items are likely to be drawn from each cluster). This is distributed as a Dirichlet, with α_p as the concentration parameter, saying how likely we believe items are to come from the same clusters as each other; we put a Gamma (*C*, 1) prior on α_p and treat *C* as a free parameter. z_i designates the current cluster assignment of item *i*.

We gave the cluster locations, $\mu_{j,}$ an empirical prior based on how often participants tend to report particular colors in another dataset (participants have slight preferences for some colors over others: Im & Halberda, 2012; Bae et al., 2014). Inferences were not significantly affected by this choice compared to using a Uniform(0, 360) prior on $\mu_{j.}$

$$\kappa_j \sim \text{Gamma}(\alpha_\kappa, \beta_\kappa)$$

To model the expected width of the ensemble distributions, we used a gamma distribution for concentration parameter κ_j with scale parameter α_κ and shape parameter β_κ. We set $\alpha_\kappa = 10$ and put a Gamma (1, 1) prior on β_κ. These priors resulted in reasonable ensemble distributions. However, inference was not strongly affected by these priors (e.g., setting $\alpha_\kappa = 5$ or $\alpha_\kappa = 15$ results in nearly identical inferences). Next we specified the distribution for the item colors, θ_i:

$$\theta_i \sim \text{vonMises}(\mu_{z_i}, \kappa_{z_i})$$

These item centers are drawn from von Mises distributions, based on the cluster they are assigned to (z_i). Finally, the actual samples participants store in memory are noisy samples from these item colors:

$$x_i \sim \text{vonMises}(\theta_i, \kappa_{sample}).$$

There is some noise inherent in sampling the colors and storing them in an internal representation; κ_{sample} captures the effects of both the sensory and memory noise involved in generating these internal observations, and was treated as a free parameter along with the concentration parameter *C*.

To model guessing, we assume that participants do not always report the inferred value of θ_i for each color. Instead, their response might be based on a corrupted or lost representation instead, in which case they will instead report a random color sampled from their prior over colors (e.g., a guess). Thus, participants' actual reported values depend on an additional variable indicating whether or not the item is corrupted:

$$\omega_i \sim \text{Bernoulli}(n_i)$$

Based on pilot data, we assumed that such corruption will happen more often to items that do not fit into an ensemble with other items than those that do fit in with the other colors. Thus, n_i takes on the value *g* if the item is alone in its own cluster; g^2 if it is in a cluster with one other item; and g^3 if it is in a cluster with all three items. This guessing parameter *g* is treated as a free parameter. If ω_i is true, then the item is corrupted and participants reported a sample from their prior over colors; if ω_i is false, they reported the value of θ_i they have inferred.

We fit this model using default Markov chain Monte Carlo (MCMC) function of PyMC (Patil, Huard, & Fonnesbeck, 2010). The Python code for this model is included in the next section of the Appendix.

We based our choice of priors and model structure on those used by Orhan and Jacobs (2013); however, because designing this clustering model inherently involves degrees of freedom above and beyond the free parameters (including the exact form of the guessing distribution, which we fit based on pilot data), we collected a totally independent dataset that was never used to test this model except for the single time we evaluated its performance. The model generalized extremely well, still providing a significantly better fit than models based on assuming all items are exchangeable. The results of this analysis are presented in the main text.

Python code for Hierarchical Bayesian model

```
import numpy as np
import pymc as pm
import matplotlib.pyplot as plt

def main():
  # Parameters
  sampleK = 7.0
  C = 15.0
  g = 0.30

  # Clusters
  K = 5

  # Observed data (samples from each item):
  observedSamples = [334.0, 353.0, 101.0] # values for first display

  # Helper values:
  numItems = len(observedSamples)
  itemList = range(numItems)
  # For each value 0 ... 360, what is the probability participants report it?
  priorReports = np.array([…])

  # Distribution of how often each cluster is sampled from:
  alphaP = pm.Gamma('alphaP', alpha=C, beta=1)
  alpha = np.ones(K)*alphaP / K
  pi = pm.Dirichlet('pi', theta=alpha)
  cpi = pm.CompletedDirichlet('cp', D=pi)

  # Which cluster does each item come from?
  z = pm.Categorical('z', p=cpi, size=numItems)

  # Assumptions about the clusters:
  # - center is drawn from prior over what colors people tend to report:
  @pm.stochastic(__class__=pm.CircularStochastic, lo=0, hi=360)
  def mu_ks(value=np.linspace(0.0, 360.0, num=5, endpoint=False)):
    return np.sum([np.log(priorReports[int(value[i])]) for i in range(K)])
  # - width is based on a relatively uninformative prior:
  beta = pm.Gamma('beta', alpha=1, beta=1)
  kap_ks = [pm.Gamma('kap_%d' % i, alpha=10, beta=beta) for i in range(K)]

  # Given its cluster, what is this item's true color?
  @pm.stochastic(__class__=pm.CircularStochastic, lo=0, hi=360)
  def itemMus(mu=mu_ks, kap=kap_ks, z=z, value=observedSamples):
      return np.sum([logVonMises(value[i], mu[z[i]], kap[z[i]]) for i in itemList])

  # Sample the actual color that gets stored in the internal representation:
  # (Note this is the only things that is 'observed')
  @pm.stochastic(observed=True, __class__=pm.CircularStochastic, lo=0, hi=360)
  def x(itemMu=itemMus, value=observedSamples):
      return np.sum([logVonMises(value[i], itemMu[i], sampleK) for i in itemList])

  # How likely is any given item to be corrupted? Depends on z (in particular,
  # the number of items that are grouped with the current item):
  @pm.deterministic
  def corruptChance(z=z):
    return [g**(sum(z==z[i])) for i in itemList]

  # Did a particular item get corrupted or not?
  corrupt = pm.Bernoulli("corrupt", p=corruptChance)

  # Reported value of participants depends on both 'corrupt' and the inferred
  # itemMu:
  @pm.deterministic
  def reportedValue(itemMus=itemMus, corrupt=corrupt):
    n = np.zeros(numItems)
    for i in range(numItems):
      if corrupt[i]:
        # If corrupt, sample from our prior on colors:
        n[i] = np.argmax(np.random.multinomial(1, priorReports)==1)
      else:
        # Otherwise, report our best guess about the item color:
        n[i] = itemMus[i]
    return n

  # Call MCMC to setup inference in this model:
  M = pm.MCMC([alphaP, beta, mu_ks, kap_ks, pi, cpi, z, itemMus,
              corrupt, corruptChance, reportedValue, x])

  # Sample from the model:
  M.sample(iter=80000, burn=20000, thin=5)

  # Show histogram of reported responses for each item:
  fig = plt.figure()
  for item in range(3):
    ax = fig.add_subplot(3,1,item+1, xlim=(0,360))
    n, bins, patches = ax.hist(M.trace('reportedValue')[:][:,item], 50,
       normed=1, facecolor=(0.5,0.5,0.5), alpha=0.75)
    l = ax.plot(observedSamples[item], 0, 'x', linewidth=5,
       markeredgewidth=3, markerfacecolor='green', markersize=10)
  plt.show()


# Convert to radians in this helper function so we can work natively in degrees:
def logVonMises(x, mu, k):
  return pm.von_mises_like(x / 180 * np.pi, mu / 180 * np.pi, k)

main()
```