Chunking as a rational strategy for lossy data compression in visual working memory tasks.

Matthew R. Nassar[1], Julie Helmers[1], and Michael J. Frank[1]

[1]Department of Cognitive, Linguistic, and Psychological Sciences; Brown Institute for Brain Science, Brown University 02912-1821

Corresponding Author:
Matthew R. Nassar
Department of Cognitive, Linguistic and Psychological Sciences
Brown University
Providence, RI 02912-1821

Phone: 607-316-4932
E-mail: matthew_nassar@brown.edu

Abstract:

The amount of visual information that can be stored in working memory is inherently limited, and the nature of this limitation has been a subject of intense debate. The debate has relied on tasks and models that assume visual items are independently encoded in working memory. Here we propose an alternative to this assumption: similar features are jointly encoded through a "chunking" process to optimize performance on visual working memory tasks. We show that such chunking can: 1) facilitate performance improvements for abstract capacity-limited systems, 2) be optimized through reinforcement, 3) be implemented by neural network center-surround dynamics, and 4) increase effective storage capacity at the expense of recall precision. Human subjects performing a delayed report working memory task show evidence of the performance advantages, trial-to-trial behavioral adjustments, precision detriments, and inter-item dependencies predicted by optimization of task performance though chunking. Furthermore, by applying similar analyses to previously published datasets, we show that markers of chunking behavior are robust and increase with memory load. Taken together, our results support a more nuanced view of visual working memory capacity limitations: tradeoff between memory precision and memory quantity through chunking leads to capacity limitations that include both discrete (item limit) and continuous (precision limit) aspects.

Introduction:

People and animals are limited in their capacity to retain visual information in short-term memory; however, the exact nature of this limitation is hotly debated [1,2]. Competing theories have stipulated that capacity is constrained by either a discrete item limit (e.g., a fixed number of "slots") or by the distribution of a flexible resource across relevant visual information (resource) [3,4]. In their simplest form, these competing theories are both philosophically distinct and statistically identifiable, but experimental evidence has been mixed, with some studies favoring each theory and the best fitting computational models incorporating elements of each [3-13]. In particular, as the number of items to be retained increases, visual working memory reports tend to become less precise, as predicted by resource models, and more likely to reflect guessing, as predicted by slots models [13].

While the competing classes of visual working memory models have evolved substantially over the past decade, they have both relied on the underlying assumption that items are stored independently of one another in working memory. Recent work has challenged this assumption, instead highlighting the tendency of human subjects to exploit any available stimulus regularities to optimize performance [14-19]. It is still unknown to what extent such optimization affects the tasks used to probe capacity limits, and thereby our understanding of their nature. This is by design, in that these tasks minimize statistical structure, thereby eliminating the potential advantages conferred by optimization of stimulus encoding through lossless data compression or optimization of memory decoding through Bayesian inference (figure 1). Nonetheless, people may employ fast and frugal lossy data compression techniques optimized to reduce memory storage requirements at a small but acceptable cost to task performance.

One such compression strategy is the joint encoding, or chunking, of similar feature values into a single blended memory representation, or chunk. Such an encoding strategy would be in line with the broader psychological literature suggesting that humans efficiently chunk sequences of information to reduce effective memory demands [20]. Furthermore, similarity-based chunking could be thought of as an extreme extension of the center-surround dynamics prominent in the visual system [21-23]: where standard implementations of center-surround dynamics predict attraction of similar representations, chunking would require merging such representations completely. Indeed, single layer neural network models of working memory have observed this phenomenon and suggested that it might play a role in limiting precision in visual working memory [24]. More complex network architectures, in conjunction with experimental data, have suggested that the access to working memory is gated, subject to reinforcement learning, and used as a form of stimulus clustering, suggesting a mechanism through which this form of chunking might be optimized according to task success [25-27].

Here we directly test whether and how people compress visual information through chunking and examine how this affects notions of working memory capacity limitation. To do so, we first develop an abstract model of capacity-limited memory and use it to demonstrate that chunking improves simulated performance, can be optimized according to reward feedback, and offers larger performance advantages for more clustered stimulus arrays. Second, we establish that a similar chunking advantage is conferred at the implementational level by center-surround dynamics, and that these dynamics mediate a direct tradeoff between the number of items stored and the apparent precision with which those items are stored. Third, we demonstrate empirically that working memory performance in human subjects is enhanced for the most "chunkable" stimulus arrays and adjusted according to reward feedback, as predicted by the abstract model, and that this enhancement comes at the cost of reduced precision, as predicted by the center-surround model. Finally, we show that the behavioral markers of chunking are present across a large corpus of previously published working memory datasets and increase as a function of memory demand. Together, these results suggest that people optimize center-surround chunking to trade precision for recall probability when approaching capacity limits, effectively producing a hybrid between continuous and discrete capacity limitation.

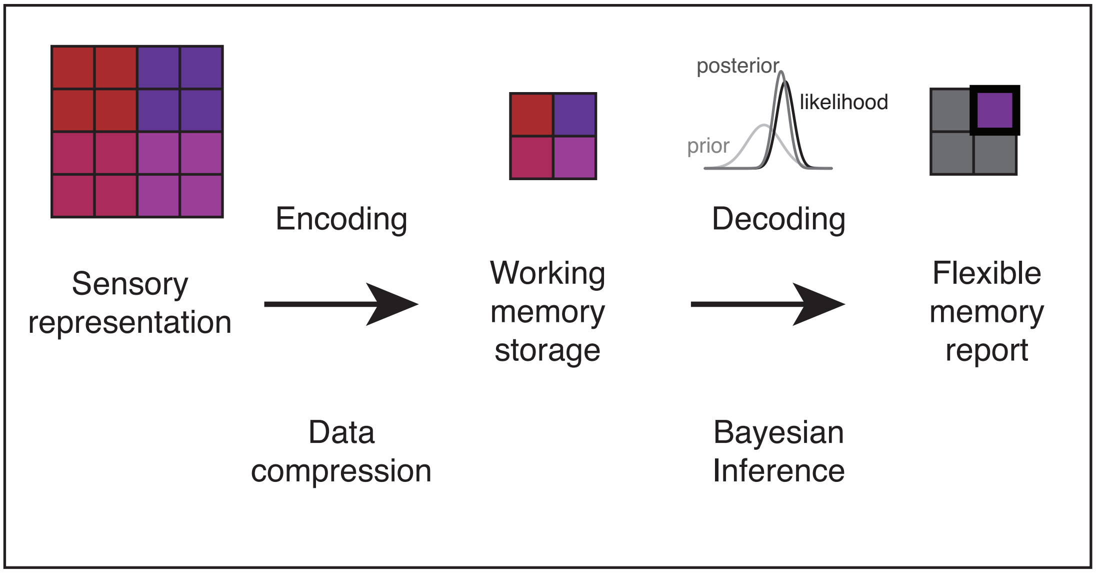


Figure 1: Working memory requires encoding sensory representations and decoding them during retrieval to inform downstream decisions. Both encoding and decoding involve transformations that can be optimized to maximize accuracy of memory reports. Optimization of encoding requires compressing data for efficient storage (data compression) [14,20], whereas optimization of decoding requires pooling information from relevant sources including prior expectations and conditional item dependencies (Bayesian inference) [15,16,18]. Standard delayed report visual working memory tasks minimize statistical structure that can be exploited through lossless data compression and Bayesian inference [3,11,28]. It is unknown whether performance advantages might be achieved in these tasks through a form of lossy data compression that deemphasizes encoding the visual information that is least relevant to successful performance.

Results:

We use three distinct approaches to examine whether, how, and why people might chunk visual information in standard working memory tasks. First, we provide a computational level analysis of the potential utility of chunking within an abstract memory system that stores visual features as binary words in a buffer with a fixed capacity and optimizes information content by adjusting encoding strategies as a function of task success. Second, we consider how chunking could be implemented through center-surround dynamics, and we examine its predicted impact on memory reports. Specifically, we extend a descriptive model of working memory to include item and feature interactions that would be expected to emerge from center-surround dynamics in memory attractor networks. Third, we test the unique behavioral predictions of each of these models with data from human subjects performing a novel delayed report task designed to test chunking model predictions. Finally, we establish that the behavioral patterns indicative of chunking in our dataset are also visible in a meta-analysis of behavioral datasets that have previously been used to evaluate competing models of working memory, but which heretofore have not considered a role for chunking.

Chunking similar object features can improve performance of capacity-limited systems.

Visual working memory capacity is typically measured using either delayed report or change detection tasks [4,8,28]. Here we focus on the former, as they have provided nuanced information about the shape of memory distributions and have formed the basis for competing models of capacity limitation [1,2,12,29].

Specifically, we consider a delayed report color reproduction task that requires storage of color and orientation information (figure 2). Each trial of the task consists of three core stages: stimulus presentation, delay, and probe. During stimulus presentation, five oriented colored bars are displayed. During the subsequent delay, the screen is blanked, requiring short-term storage of color and orientation information. During the probe stage, an oriented bar is displayed (in gray) and the participant is required to report the color that had been associated with that orientation in the preceding stimulus array.

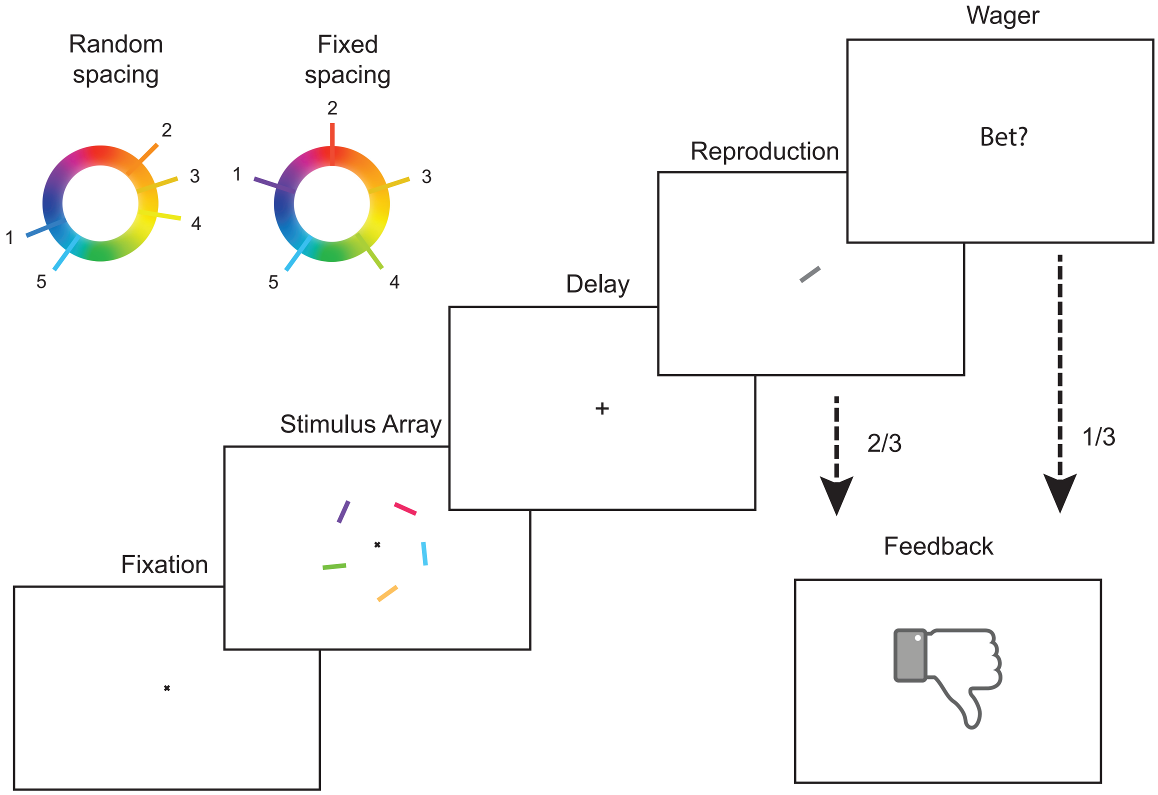


Figure 2: Delayed report color reproduction task. Each trial begins with central fixation for 500 msec, followed by stimulus presentation for 200 msec. Stimuli consist of five colored and oriented bars evenly distributed around a circle subtending 4 degrees of visual angle and centered on the point of fixation. Stimulus presentation is followed by a 900 msec delay, after which a single oriented bar is displayed centrally. The subject is required to report the color associated with the bar with the probed orientation in the previous stimulus array. After confirming the report, the subject receives feedback dependent on whether the absolute magnitude of the reproduction error was greater or less than a fixed threshold. Stimulus colors on any given trial are selected either: 1) randomly and independently as is standard in such tasks (random spacing; upper left) or 2) ensuring uniform spacing on the color wheel so as to minimize within-array color similarity (fixed spacing).

To first understand whether, in principle, information encoding could be optimized in this task, we developed a limited-capacity system for memory storage in which colors and orientations are represented with binary words (figure 3). The precision with which a color is stored depends on the number of binary digits (bits) used to represent that color: a single bit can be used to specify a half of the color wheel, a second bit can be added to specify a quadrant of the color wheel, and so on (figure 3, top). Capacity limitations within such a system can be easily implemented as a fixed limit on the number of bits stored during the delay period. These bits can be used to represent the individual colors in the target array, by, for example, dividing them evenly among the available targets (independent item encoding; figure 3). Alternatively, multiple similar colors could be jointly represented with a single binary word that is then linked to multiple orientations (chunking; figure 3). An intuitive advantage of the second encoding strategy is that reducing the number of binary color words increases the number of bits available to represent each word, potentially offsetting the biased encoding of the chunked items by better representing each encoded color.

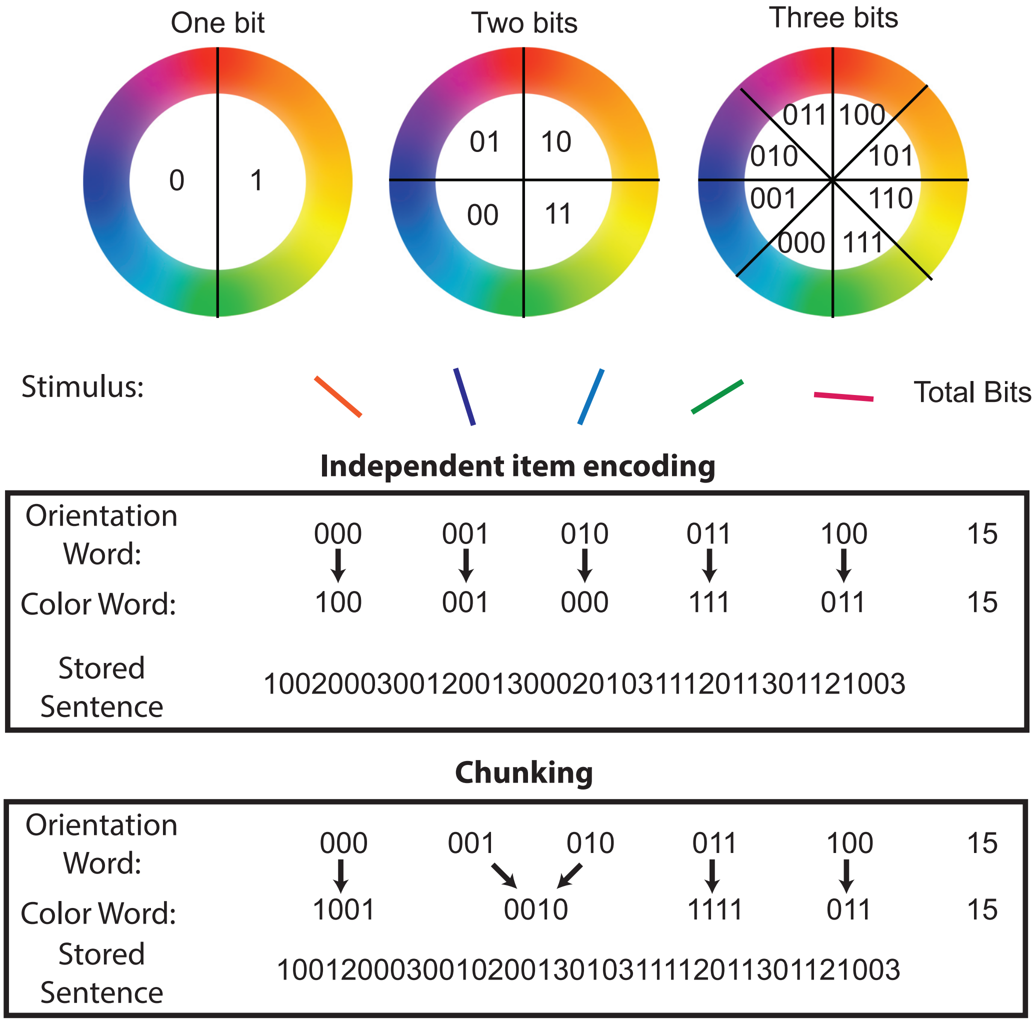


Figure 3: Binary encoding model of visual working memory. In order to formalize capacity limitations, it is useful to consider an abstract model of working memory that stores features in binary words. **Top**: Each color can be described by a binary word of fixed length, where the number of digits in the word determines the storage precision. **Middle & Bottom**: Stimulus arrays can be stored by linking ordered pairs of color and orientation words. Capacity limitations are modeled by a fixed limit on the length of the resulting "sentence" comprised of color and orientation words separated by word termination symbols (2/3 for color/orientation words, respectively). One strategy for storing ordered pairs involves alternating sequences of color and orientation words, such that each color is linked to a single orientation (Middle). Another strategy for storage would be to link two or more orientations to a single color. This reduces the number of colors that need to be stored, and thus increases the number of bits allotted to each color (Bottom).

To test this potential advantage quantitatively, we simulated the task performance of models with various levels of chunking under a variety of conditions. We parameterized chunking by setting a "partitioning criterion" that defines the minimum distance between two colors required for independent representation. If the distance between two colors is smaller than the partitioning criterion, the colors are represented as a single "chunk". Thus, a partitioning criterion of zero indicates that all items are represented independently, whereas a partitioning criterion of π indicates that all item colors will be chunked together (i.e. represented by a single binary word). Performance of models with various partitioning criterions was simulated for working memory color reproduction tasks ranging from easy (two targets) to difficult (eight targets).

Model performance depended on partitioning criterion as a function of task difficulty (figure 4). For easier tasks with few items to encode, the model's memory buffer was large enough to store each item independently with a reasonable number of bits, such that increasing the partitioning criterion beyond zero was detrimental to task performance (two targets; figure 4a). However, for harder tasks, in which storing each item with high precision was not possible due to limited buffer size, performance was best for moderate partitioning criterions that allow for joint representation of similar, but not divergent, colors (five targets; figure 4b). Across task difficulties, there was a monotonic relationship between the number of targets and the performance advantage that could be attained through some level of chunking (figure 4C). However, the magnitude of the chunking bonus also depended on the exact distribution of the target colors within the array. If the colors were randomly sampled independently, which is the standard method in such tasks, chunking offered large advantages. In contrast, the advantages of chunking were considerably smaller when colors were uniformly distributed in color space to maximize color separation (fixed spacing; figure 1 inset; compare the blue and yellow lines in figure 4A-C).

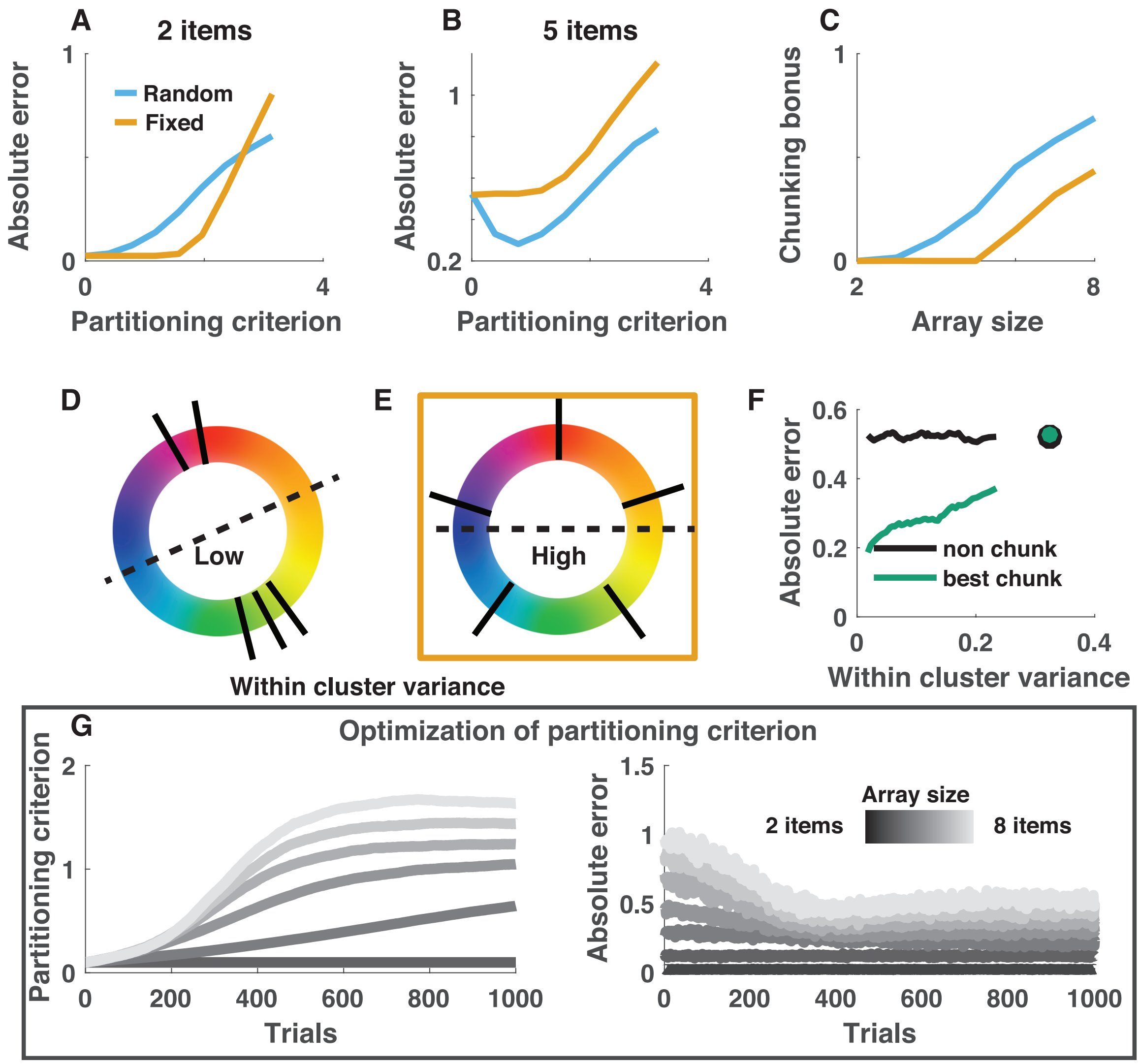


Figure 4: Chunking improves memory performance in binary encoding model. A&B: Mean absolute error (abscissa) for simulated performance of a binary encoding model with partitioning criterions (ordinate) ranging from 0 (all colors are stored independently) to π (all colors are stored in a single chunk). Blue/yellow lines reflect performance on tasks containing random/fixed spacing color arrays. When capacity is large relative to the array size, as in **A,** increased chunking produced larger errors and diminished performance. However, for larger memory arrays, as in **B**, chunking similar items produced smaller errors and improved performance, particularly when stimuli were randomly spaced. **C:** The chunking bonus, measured as the mean absolute error of the non-chunking model (partitioning criterion = 0) minus the mean absolute error of the best performing model, increased monotonically with memory load and was larger for random spacing stimulus arrays. **D&E:** Within cluster variance provides a measure of feature clustering within a stimulus array, with low values indicating more clustering (**D**) and high values indicating less clustering (**E**). **F:** Performance of the best chunking model, but not the non-chunking model, depends on the clustering of individual stimulus arrays, as assessed through within cluster variance. Mean absolute error is plotted for stimulus arrays grouped in sliding windows of within cluster variance for the best chunking (green) and non-chunking (black) models. Circles reflect the same values computed for fixed spacing trials. **G:** Chunking behaviors could be optimized by adjustment of the partitioning criterion as a function of trial-by-trial feedback about task success. Partitioning criterion (left) and mean absolute error (right) are plotted across trials (ordinate) for simulated models that adjust the partitioning criterion

on each trial according to the chunking and feedback on the previous trial using reinforcement learning (see Methods). Simulations are sorted by array size, with lighter colors corresponding to larger memory arrays.

To better characterize the aspects of the randomly spaced target arrays that enabled better performance in the chunking models, we computed the within cluster variance (WCV) as a simple metric of the "chunkability" of each stimulus array (see Methods). Target arrays with tightly clustered colors have low WCV, whereas target arrays with distantly spaced colors have high WCV (figure 4D&E). The performance of chunking models depended monotonically on WCV, with the smallest errors achieved on low WCV target arrays (figure 4F), supporting the notion that chunking advantages are achieved by more efficient representation of similar colors through joint encoding.

To examine whether the advantages of chunking could be learned online, we tested whether the partitioning criterion could be optimized on a trial-to-trial basis via reinforcement learning. When the partitioning criterion was adjusted on each trial according to the chunking (total number of chunks) and reward feedback (thresholded binary feedback) from the previous trial, it tended to increase rapidly until reaching a load-dependent asymptote (figure 4G). Trial-to-trial increases in the partitioning criterion corresponded to rapid improvements in overall task performance, as measured by average absolute error (figure 4G). Thus, chunking could be optimized across conditions to improve performance and facilitate greater chunking in higher memory load contexts.

Taken together, the results from the binary encoding model suggest that 1) chunking nearby feature values can improve performance in working memory tasks, 2) performance improvements from chunking increase with target number and are mitigated by uniformly spacing feature values, 3) performance of chunking models depends monotonically on WCV, and 4) chunking behavior can be learned through reinforcement of successful chunking behaviors. In summary, the binary encoding model clarifies why, and under what conditions, chunking could benefit working memory performance in standard tasks. In the next section, we extend these ideas to examine how chunking could be achieved by biological systems and to identify more detailed behavioral predictions of plausible chunking mechanisms.

Chunking advantages can be conferred by center-surround dynamics.

The brain is thought to implement visual working memory in neural networks that include individual neurons tuned to specific visual features and capable of maintaining representations in the absence of input [30-34]. Neural responses within such networks are thought to obey center-surround receptive field architectures that are present throughout the visual system [35,36], supported by lateral connectivity [37-39], and have recently been shown to bias working memory reports [40]. In computational models, the ability of a neural network to maintain feature representations in the absence of inputs depends critically on the recurrent connections between neurons [41-45]. In particular, persisting feature representations, like those of the colors in our task, are facilitated by local excitatory connections between similarly tuned neurons and by broadly tuned inhibition between neurons with dissimilar tuning (figure 5A). Such connectivity promotes interactions between multiple simultaneous representations, with the impact of each color representation on the others depending on their level of similarity. If the two stored colors are similar enough to promote mutual recurrent excitation, each represented color will experience biased excitation in the direction of its neighboring color, and eventually the two color "bumps" will merge to form a single representation (figure 5B) [24]. In contrast, if the stored colors are separated beyond the narrowly tuned recurrent excitation function, mutual recurrent inhibition will dominate, leading to a net repulsion of color representations from one another (figure 5C) [40,46]. Together, the interactive dynamics of the system can be described by a "difference of Gaussians" function that mediates the attraction and joint representation of similar colors and the repulsion of dissimilar ones (figure 5B&C; yellow shading).

To incorporate these dynamics into a descriptive model of working memory performance, we implemented attractive and repulsive forces among stored memories in accordance with narrowly tuned excitation and broadly tuned inhibition functions. On each trial, each color from the target array was: 1) perturbed by a mean zero random variable to simulate neural noise, 2) chunked with each other color in the array with probability proportional to the excitation tuning function, 3) repulsed by each other color in the array with magnitude proportional to the inhibition tuning function, and 4) probabilistically stored across the delay period according to a Poisson process. The proportionality constants allowed us to examine the performance of models ranging from those that were not affected by recurrent dynamics (zero proportionality constant) to those that were highly affected (high proportionality constant).

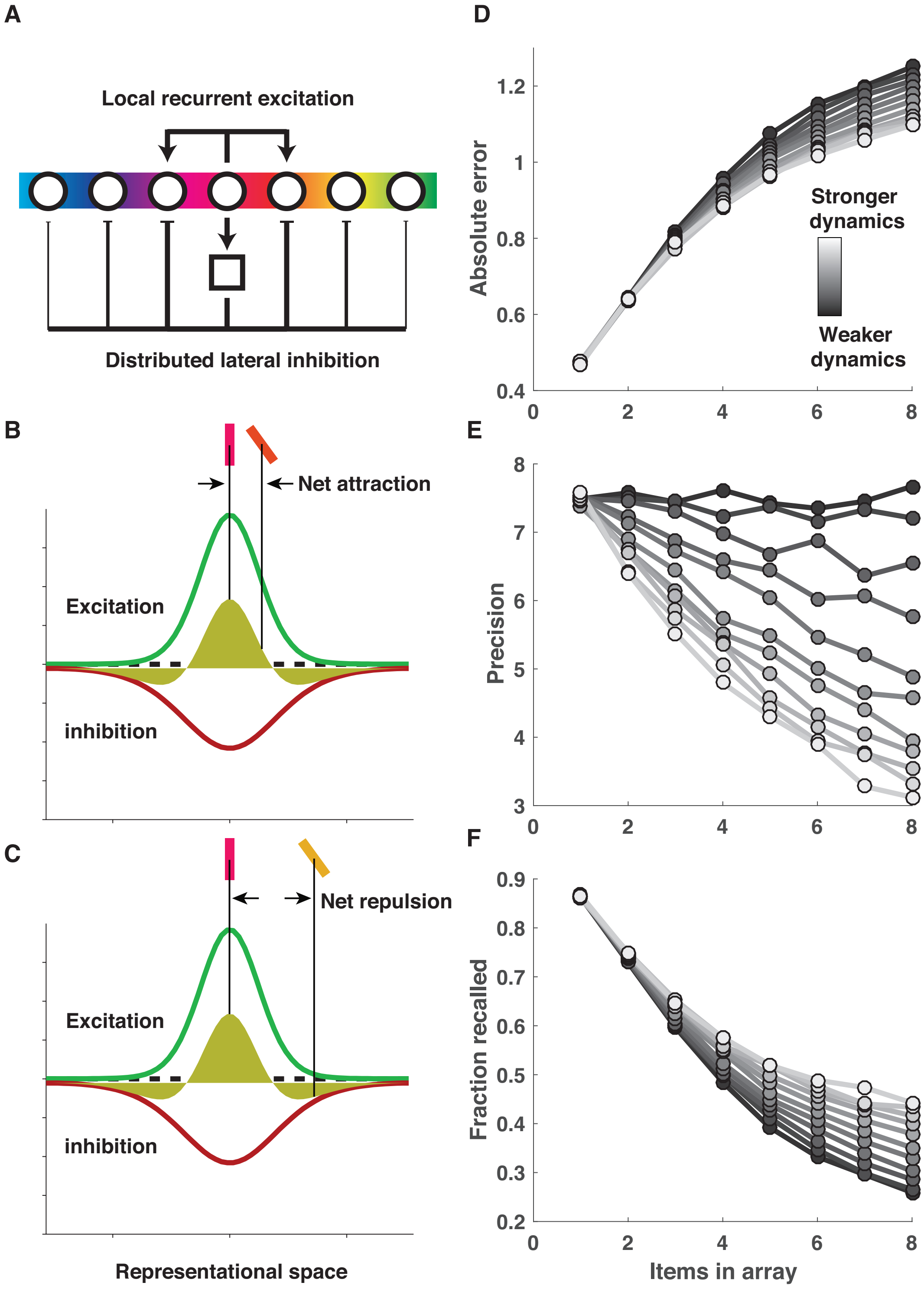


Figure 5: Attractive and repulsive forces characteristic of center-surround connectivity can mediate chunking to improve recall at the cost of precision. A) Local recurrent excitation and distributed lateral inhibition are thought to be key components of biological networks responsible for visual working memory storage. These patterns of neural connectivity give rise to two counteracting forces: recurrent excitation facilitates attraction of neighboring representations through "bump collisions" [24], whereas broadly tuned lateral inhibition facilitates repulsion of distinct

bumps of neural activity [40,46]. **B&C)**Together, these forces produce a difference of Gaussians tuning function (yellow shading; **B**&**C**) that facilitates attraction of closely neighboring representations but repulsion of more distant ones. Here we model these effects at the cognitive level by assuming that two imprecise internal representations of color are chunked, and jointly represented by their mean value, with a fixed probability defined by a narrowly tuned von Mises distribution (green curve; **B**&**C**) in order to mimic the effects of narrowly tuned excitation. After probabilistic chunking, each color representation exerts a repulsive influence over all other representations with a magnitude defined by a broadly tuned von Mises distribution (red curve; **B**&**C**) in order to mimic the effects of broadly tuned inhibition. The model stores a Poisson number of the representations, chunked or otherwise, for subsequent recall. **D**) The influence of center-surround dynamics over model performance can be manipulated by applying a gain to the amplitude of the excitation and inhibition functions such that larger values correspond to greater item interdependencies and lead to smaller errors on average (lighter colors correspond to higher gain). **E**&**F**) The performance improvement mediated by increasing center-surround dynamics relies on a tradeoff between recall probability and precision, through which increased attractive and repulsive forces reduce precision (lighter bars; **E**), but enhance recall probability (lighter bars; **F**).

Models implementing greater recurrent dynamics achieved better performance through a recall/precision tradeoff. Performance was simulated on delayed report tasks in which target number (array size) ranged from one to eight. Performance of models employing recurrent dynamics was slightly worse for easier tasks but dramatically improved for more difficult ones, similar to the effects observed in the binary model above (figure 5D; lighter lines represent stronger recurrent dynamics). Here, though, performance differences were characterized by opposing effects of recurrent dynamics on precision and recall. Models employing recurrent dynamics showed improved recall, particularly in the hardest tasks, as "bump" collisions allowed for the storage of multiple target features in a single representation (figure 5F). On the other hand, models employing recurrent dynamics showed reduced precision, as both attractive and repulsive forces reduced the fidelity of stored color representations (figure 5E). In standard models of resource limitations, decrements in precision that occur with increased array sizes have been attributed to the division of a limited resource. However, in the recurrent dynamics models, the decrement in precision is caused by the increase in inter-item interactions that occurs when additional items are added to the memory array. Thus, the inclusion of recurrent dynamics affects the nature of capacity limitations: minimizing the impact of center-surround forces leads to a specific decay in recall as a function of array size, as predicted by "slots" models, whereas maximizing the impact of center-surround forces leads to decays in precision across set size, heretofore linked to resource depletion accounts [1-4].

In summary, inter-item dependencies that emerge from center-surround dynamics are sufficient to mediate the performance bonuses of chunking, but do so at the cost of precision. Thus, if working memory is optimized through chunking in this way, it should lead to more faithful recall for clustered stimulus arrays but more precise recall of less clustered ones. In principle, such optimization could be guided in cognitive or real-world tasks by implicit or explicit feedback to favor successful chunking strategies and avoid unsuccessful ones.

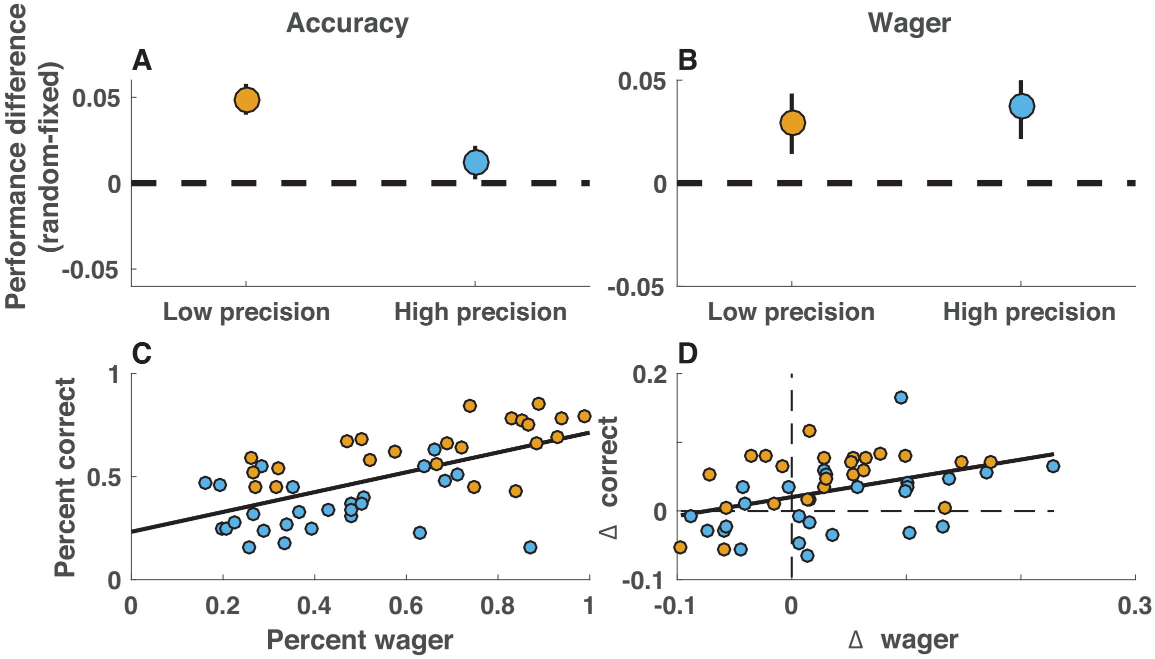


Figure 6: Uniformly spaced stimulus configurations degrade task performance and confidence in human subjects. Subject performance was assessed in terms of accuracy (percent of trials eliciting positive feedback) and confidence (percent of trials eliciting high post-decision wagers) separately according to precision condition (23 subjects were required to achieve an error of less than π/3 to elicit positive feedback [low precision], whereas 24 subjects were required to achieve an error of less than π/8 to elicit positive feedback [high precision]). **A)** Subjects in the low precision condition were more accurate for random spacing, as opposed to fixed spacing, stimulus configurations (orange; t=5.6, p < 10e-4), whereas subjects in the high precision condition attained similar overall performance in both configurations (blue; t=1.5, p = 0.15). Points/lines indicate group mean/SEM. **B)** Subjects in both conditions indicated higher confidence for random-spacing, as opposed to fixed-spacing, stimulus configurations (t = [2.3, 2.0] and p = [0.03, 0.06] for high and low precision conditions, respectively). **C)** Subjects that were most accurate, as assessed online according to a fixed error threshold, also tended to make higher post-decision wagers. Orange and blue points indicate subjects in low and high precision conditions, respectively. **D)** Furthermore, the improvement in accuracy from fixed- to random-spaced arrays was greater for subjects that showed the largest increase in confidence across the same conditions.

People are more accurate and confident when remembering chunkable target arrays.

To directly test key predictions made by the binary encoding and center-surround models, we collected behavioral data from human participants in the task described in figure 2. The critical manipulation in the task is that the colors in some trials were uniformly spaced on the color wheel (fixed spacing) whereas the colors in interleaved trials were randomly and independently sampled from the color wheel (random spacing). This manipulation allowed us to test 1) the prediction of the binary coding model that chunkable random spacing arrays will lead to better performance (figures 4b and 2) the prediction of the center-surround encoding model that such recall improvements will be accompanied by decrements in precision (figure 5E&F). The task also required a post-decision wager on one third of trials and provided explicit feedback on each trial, determined by comparing error magnitude to a fixed threshold. These task features allowed us to determine whether participants were aware of any performance bonuses attributable to chunking, and to test the extent to which chunking behaviors are adjusted trial-by-trial in accordance with their effects on task performance.

In accordance with the predictions about behavioral optimization through chunking, participants were more accurate and confident when presented with randomly spaced stimuli. Subject accuracy, as assessed using the same error threshold procedure used to determine feedback, was greater on random spacing trials than on fixed spacing trials (figure 6A; $t = 4.4$, $p < 10e-4$). This accuracy improvement was more prevalent for subjects that experienced a liberal accuracy criterion (low precision, absolute error $< \pi/3$) than for those that experienced a conservative accuracy criterion (high precision, absolute error $< \pi/8$; two sample t-test for group difference: $t = -2.5$, $p = 0.02$). Participants also wagered more frequently on random spacing than fixed spacing trials, suggesting that they were cognizant of the conferred performance advantage (figure 6B; $t = 3.1$, $p = 0.003$). Subjects tended to gauge their wagering behavior reasonably well, with subjects who achieved higher levels of accuracy also betting more frequently (figure 6C; Pearson's rho =0.68, $p < 10e-6$). Furthermore, individual differences in the adjustment of wagering as a function of color spacing configuration correlated with the change in accuracy that subjects experienced across the configurations (figure 6D; Pearson's rho =044, $p = 0.002$). Taken together, these data suggest that subjects experienced and were aware of performance advantages for randomly spaced stimuli, but that the extent of these advantages differed across individuals.

To better understand these performance advantages, we tested the extent to which trial-to-trial accuracy and confidence scores depended on stimulus clustering within the randomly spaced stimulus arrays. Specifically, we computed within cluster variance (WCV) for each color array to evaluate whether this clustering statistic could be used to predict subjects' accuracy of, and confidence in, color reports. As predicted by binary encoding chunking models (figure 4F), subjects were more accurate for low WCV trials; performance on high WCV trials was similar to that in the fixed spacing configuration (figure 7A). Furthermore, subject wagering also decreased monotonically with WCV, such that betting behavior on the highest WCV (least chunkable) color arrays was similar to that on fixed spacing trials (figure 7B).

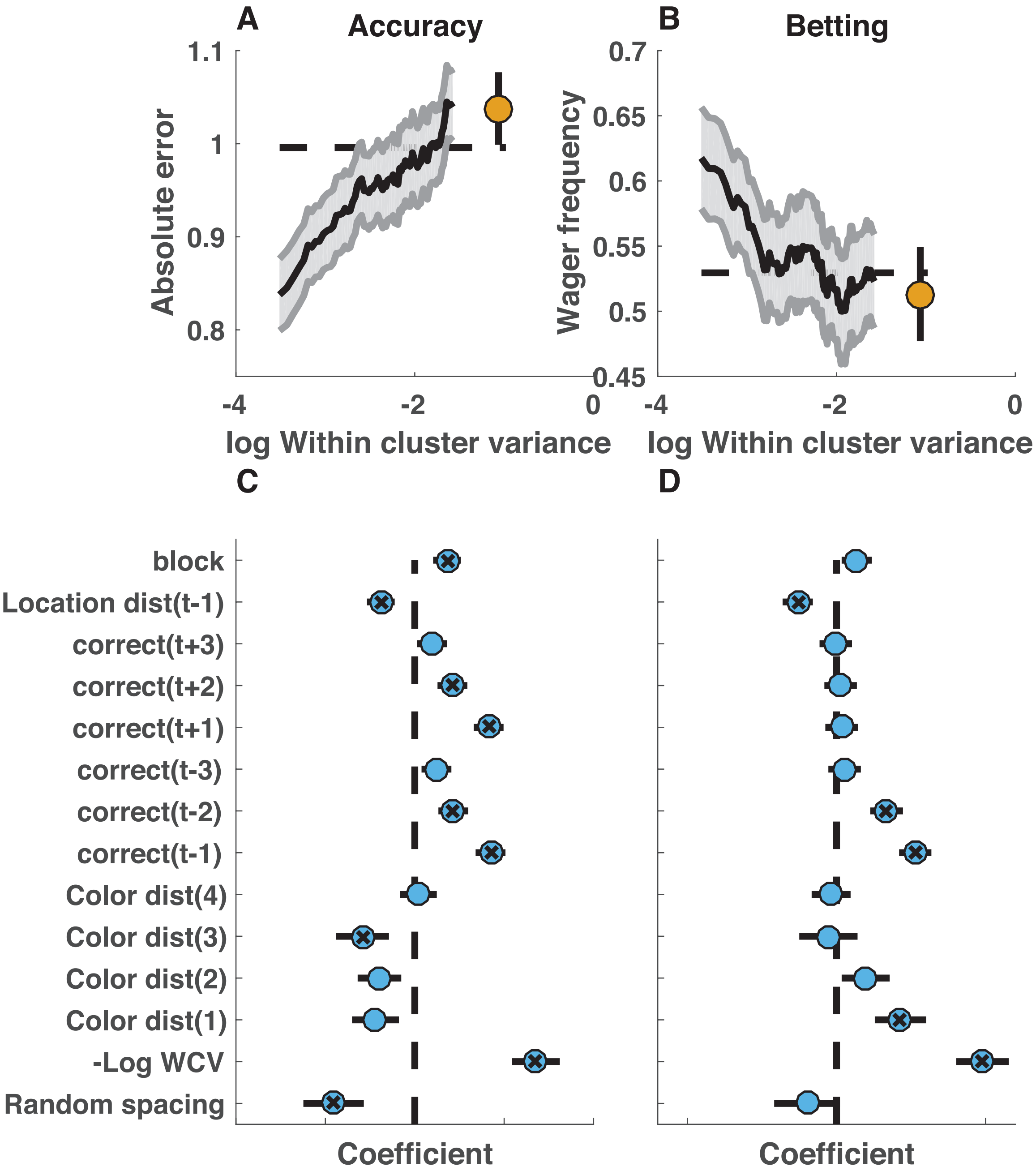


Figure 7: Memory performance and confidence vary from trial to trial according to stimulus clustering. A&B) Memory performance and confidence increase linearly with the chunkability of stimulus arrays. Mean absolute error magnitude (**A**) and high wager frequency (**B**) was computed per subject in sliding bins of within cluster variance (larger values = decreased chunkability) for random (lines) and fixed spacing conditions (points). Lines and shading reflect mean and SEM across subjects. **C&D)** Effects of chunkability on performance and confidence persist after accounting for potential confounding factors and feedback-dependent performance adjustments. Coefficients from a mixed-effects logistic regression model of binary accuracy (**C**) and wager (**D**) are plotted on the abscissa. Circles/lines reflect mean/SEM, and X marks indicate coefficients significantly different from zero ($p < 0.05$).

To determine whether the effects of WCV on confidence and accuracy could be mediated by alternative explanations, we applied a generalized linear model (GLM) to the binary accuracy and confidence data. The GLM included –log(WCV) as well as variables that would better explain the data under competing explanations. In particular, we included the distances between the target color and each non-probed color as nuisance variables to determine whether the apparent WCV effects could be better explained by a tendency to report non-probed colors, which are often closer to the target color for more chunkable stimulus arrays. When this model was applied separately to accuracy and post-decision wagers, coefficients for –log(WCV) were greater than zero, suggesting that the advantage is conferred by the configuration of all stimuli, and not simply by the proximity of the target color to neighboring colors (figure 7C&D; accuracy $\beta = 0.028$, $t = 5.3$, $p < 10e\text{-}6$; confidence $\beta = 0.051$, $t = 5.8$, $p < 10e\text{-}8$).

Choice accuracy is modulated by reward feedback.

The GLM also included terms to account for other factors that could potentially affect task performance, including feedback from previous trials. Positive feedback on the previous trial was predictive of higher accuracy and confidence for the current trial, in a manner consistent with trial-by-trial behavioral adjustment (figure 7C&D, “correct (t-1)” coefficient; accuracy $\beta = 0.017$, $t = 5.0$, $p < 10e\text{-}6$; confidence $\beta = 0.026$, $t = 4.8$, $p < 10e\text{-}5$). This predictive relationship was unlikely to be driven by autocorrelation in performance, as such an explanation should also predict that confidence measurements relate to accuracy on future trials, and this relationship was not evident in the data (figure 7D, “correct (t+1)” coefficient; confidence $\beta = 0.0017$, $t = 0.3$, $p = 0.75$). Despite seemingly robust feedback-driven effects, overall performance improvements across the session were somewhat modest, as evidenced by a relatively small positive coefficient for a term in the GLM relating block number to accuracy (figure 7C, “block” coefficient; accuracy $\beta = 0.007$, $t = 2.3$, $p = 0.02$). Thus, the GLM results suggest that subjects gained a performance advantage for chunkable target arrays, modulated behavior in response to previous feedback, and improved slightly over the course of task sessions.

People are less precise when remembering chunkable target arrays.

If the chunking advantages described above are mediated through center-surround dynamics in recurrent networks, then they should come at the cost of precision (figure 5E&F). To test this prediction, we pooled the error distributions across all subjects separately for random- and fixed-spacing trials and examined the difference in error distributions for the two conditions (figure 8, left column). The error distributions from both conditions were consistent in shape with those previously reported in similar tasks (figure 8A&D) [13]. However, the error distributions differed subtly between the two conditions: in the random-spacing condition, subjects made more small errors, but did not have more perfect recalls (figure 8G). This pattern of difference was also evident in data simulated from the center-surround chunking model (figure 8, middle column) but not in data simulated from an independent encoding model fit to subject behavior (figure 8, right column). Thus, both the subjects and the center-surround chunking model reported more colors that were slightly offset from the target color in the random-spacing condition than in the fixed-spacing condition, consistent with a reduction in precision resulting from chunking.

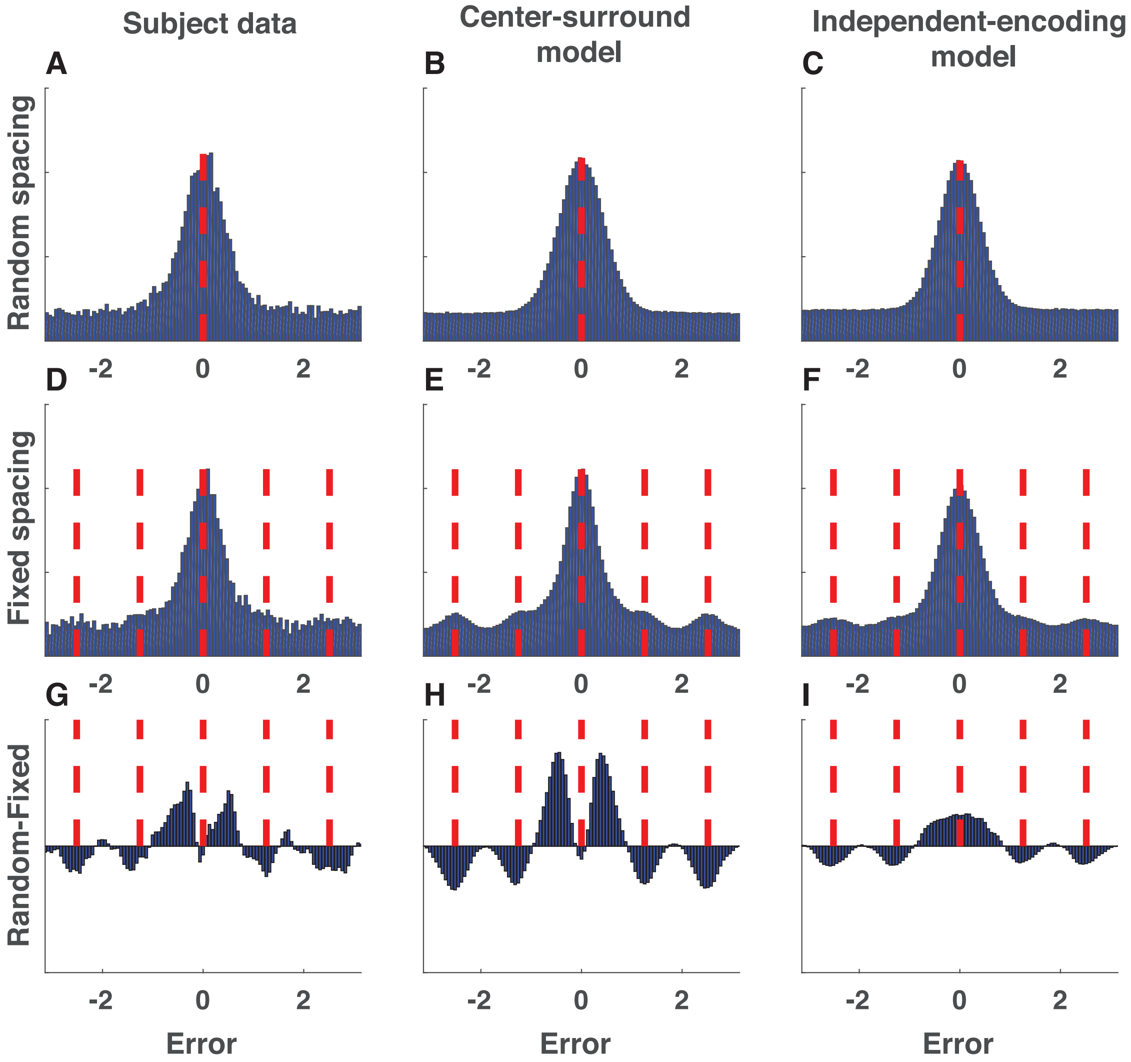


Figure 8: Error distributions reveal evidence for center-surround chunking. A-C) Signed color reproduction errors made in the random spacing condition by (**A**) subjects, (**B**) center-surround chunking models, and (**C**) independent encoding models. Data is collapsed across all simulated or actual sessions. **D-F)** Same as **A-C** but for the fixed spacing condition. Red dashed lines indicate probed and non-probed target locations. Note that alignment of non-probed target locations emphasizes the prominence of non-probed target reports (binding errors), which would appear uniformly distributed in the random spacing condition. **G-I)** Difference in above error distributions for random minus fixed. To aid in visualization, bin count differences were smoothed with a Gaussian kernel (standard deviation = 1 bin). Subjects and the center-surround chunking model show increased moderately small, but non-zero, errors in the random spacing condition. Note that differences of reports near the non-probed targets are present in both models, as they simply reflect an artifact of alignment of binding errors in the fixed spacing condition.

Errors are modulated by nearest neighbors consistent with chunking via recurrence

To better understand the nature of the increased moderate sized errors in the random spacing condition, we sorted trials according to the non-probed target color that was most similar to the probed target color (nearest neighbor color; see Methods for details). This procedure revealed structure in individual color reports related to the nearest neighbor non-probed color (see supplementary figure 1). To determine whether such structure persisted systematically across subjects, we fit a descriptive mixture model to error distributions pooled across subjects in sliding windows of nearest neighbor distance. The model contained free parameters to examine 1) the precision of error distributions, 2) the bias of error distributions toward the nearest neighbor non-probed target color, and 3) the relative proportion of trials that were recalled, forgotten, or mis-bound (in keeping with nomenclature from previous literature [11,47]).

The model fits revealed that subject precision and bias depended on nearby stimulus colors in a manner consistent with chunking through recurrent dynamics. In particular, subject memory reports were biased towards the nearest neighbor color if it was sufficiently similar to the probed target color, but biased away from it if it was sufficiently dissimilar (figure 9A). Qualitatively, this pattern of bias maps onto the idea of a narrowly tuned excitation function promoting attraction of nearby targets and a broadly tuned inhibition function promoting repulsion of more distant ones (see figures 5B&C). Precision also depended on nearest neighbor color distance. Subject precision was maximal when the nearest neighbor color was most dissimilar to the probe color and minimal when it was moderately similar (figure 9D). In addition, fits revealed an increase in the proportion of correct recalls, and a corresponding decrease in the number of uniform guesses, when a nearby neighbor color existed in the stimulus array (Figure 9G&J). This pattern of results was strikingly consistent with those produced by a chunking model based on recurrent dynamics (figure 9, middle column) but not with those produced by the best fitting independent encoding model (figure 9, right column).

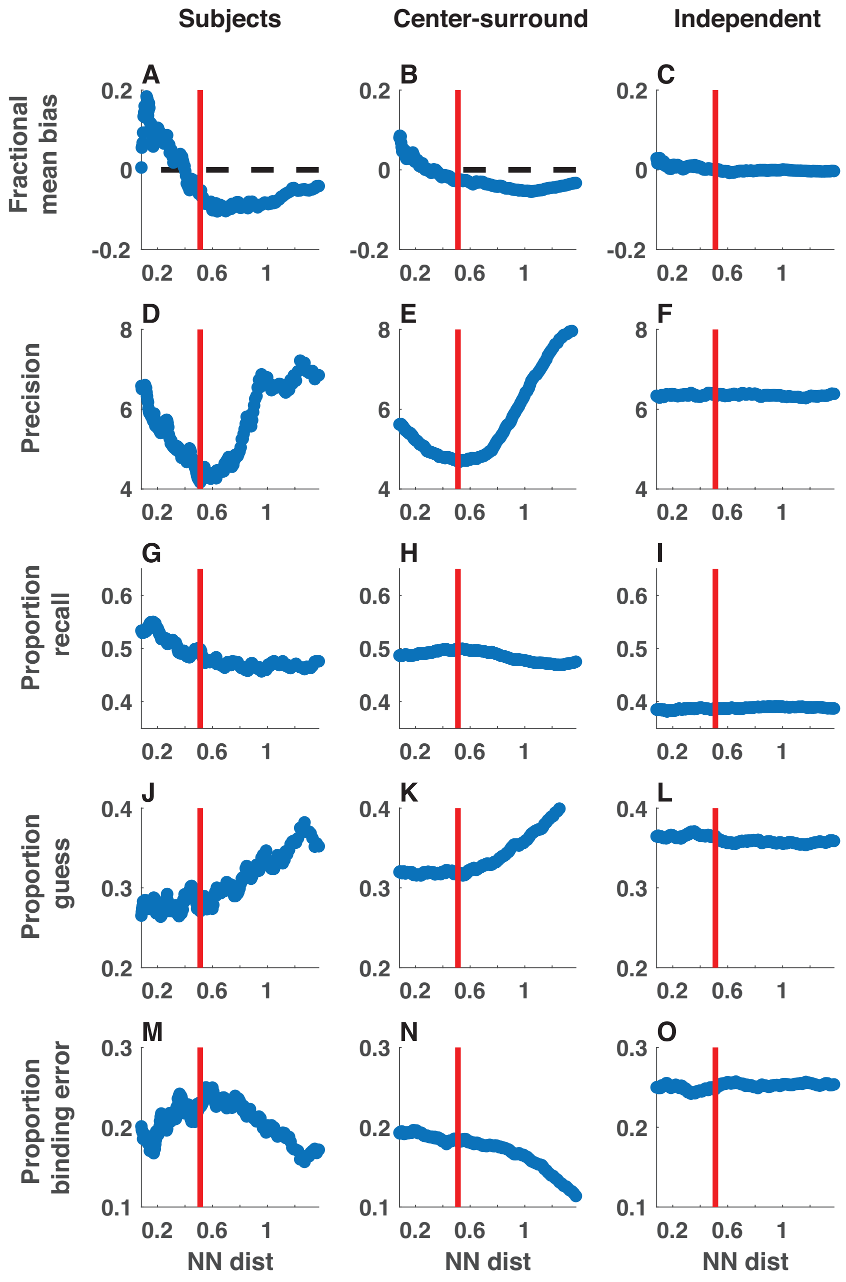


Figure 9: Neighboring stimulus features affect bias, precision, and recall probability as predicted by the center-surround chunking model. Subject (left) and simulated (center = center-surround, right = independent encoding) data were collapsed across all sessions and binned in sliding windows according to the absolute distance between the probed target color and the most similar non-probed target color (NN distance; abscissa). Data in each bin was fit with a mixture model that included free parameters to estimate 1) the bias of memory reports towards the closest color in the target array expressed as a fraction of distance to that target (**A-C**), 2) the precision of memory reports (**D-F**), and 3) the proportion of reports generated from: the von Mises "memory distribution" (**G-I**), the uniform "guess distribution" (**J-L**), or the mixture of von Mises "binding error distribution" (**M-O**). The qualitative trends present in subject data are also present in data simulated

from the center-surround chunking model but not in those simulated from the independent encoding model.

Jointly accounting for recall, precision, and bias reveals feedback-dependent adjustments of chunking.

To account for the predicted effects of chunking on precision, recall, and bias, we extended the basic mixture model of subject memory reports described above to include coefficients simultaneously quantifying WCV effects on recall (figure 7) and nearest neighbor distance (NND) effects on precision and bias (figure 9). Coefficients for these modulatory terms indicated that the probability of remembering a target decreased as a function of WCV, while precision and bias increased according to NND-based predictions (figure 10, gray points; WCV effect on recall: t = -6.8, p < 10e-7; precision modulation by NN distance: t = 4.24, p = 10e-4; mean shift modulation by NN distance: t = 2.5, p = 0.02). This same pattern of results held for data simulated from the center-surround chunking model, but not for data from an independent encoding model (figure 10, gold and blue points, respectively).

We also tested whether people adjusted chunking strategies online as a function of reward feedback in a manner similar to that used to optimize performance in the binary encoding model (figure 4G). In particular, in the binary encoding model, the partitioning criterion was adapted based on the previous trial's chunking and feedback and selectively contributed to performance improvements for the most chunkable stimulus arrays (figure 4F). To explore the possibility that people adjust chunking in a similar way, we included additional terms in the mixture model to allow recall probability to vary as a function of previous trial feedback (pCorr), proxies for previous and current trial chunkability (pWCV, WCV), and their predicted interactions (see Methods). The best fitting coefficients revealed an overall recall bonus on trials following correct feedback (pCorrect: t = 5.4, p < 10e-5), but also that the magnitude of this bonus was greater for more chunkable trials (pCorrect * WCV: t = -2.1, p < 0.05) and for trials that had a chunkability matched to that of the previous trial (pCorrect * WCV * pWCV: t = 2.0, p < 0.05; figure 10D). Consistent with optimization of chunking via reinforcement learning, these interactions capture a tendency toward larger feedback-driven changes in task performance when both the current and previous trial color arrays were highly chunkable (figure 10E&F).

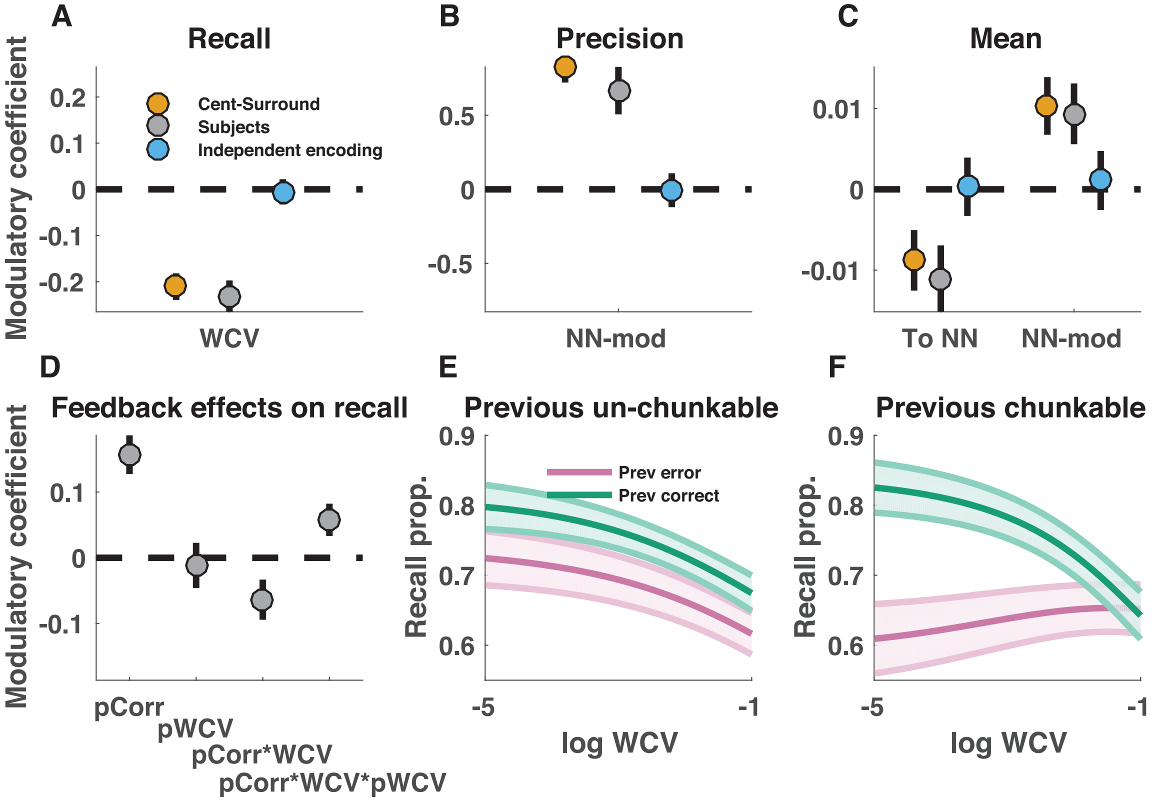


Figure 10: Extended mixture model reveals chunking effects on recall, precision, and bias, together with feedback-based adjustments in chunking behavior. Subject and simulated data was fit with a descriptive model in which errors were generated from a mixture of three processes: successful recall, uniform guesses, and binding errors (recalling the color of a non-probed target). The model was extended in three ways to account for the effects of chunking on recall, precision, and bias. **A)** Coefficients describing the logistic effect of within cluster variance (WCV) on recall probability are plotted separately for subjects (gray) and simulated data (center-surround and independent encoding in yellow and blue, respectively). Mean and SEM are reflected by circles and lines, respectively. **B)** Coefficients describing the effect of neighboring color distance on overall precision. Positive values indicate adherence to the relationship in figure 9D. **C)** Coefficients describing the overall bias in recall distributions towards the nearest neighbor non-probed target color (left) and the modulation of this bias according to the nearest neighbor color distance. Positive values indicate adherence to the (center-surround) relationship in figure 9A. **D)** Recall probability was modulated by feedback and chunkability, suggesting trial-to-trial adjustments of chunking. The extended mixture model included additional terms to account for the effects of feedback on subsequent trial performance. Mean/SEM coefficients for terms accounting for 1) the overall effect of positive feedback (pCorr) on subsequent performance, 2) the overall effect of previous trial log within chunk variance (plWCV) on subsequent performance, 3) the multiplicative interaction pCorr*plWCV, and 4) the three-way interaction pCorr*lWCV*plWCV are represented from left to right with points/bars. **E&F)** Recall probability of best fitting descriptive models plotted as a function of the log(WCV) for the current trial and divided according to previous feedback (color) and the log(WCV) for the previous trial [E (unchunkable): plWCV =-1, F (chunkable): plWCV=-5]. Lines/shading reflect mean/SEM across subjects.

Chunking effects generalize across tasks and scale with memory load.

Finally, to test whether our empirical findings were robust to changes in task conditions and to examine how they vary with memory load, we fit the mixture model described above to a meta-analysis dataset that included 101 subjects performing eight different experiments [13]. Consistent with previous reports, recall proportion decayed as a function of the number of targets in the stimulus array (array size) across all subjects and experiments (figure 11A; yellow), and precision was greatest for the smallest array sizes (figure 11B) [13].

Modulators of recall, precision, and bias also varied as a function of array size in a manner consistent with chunking (figure 11C-F). To quantify these effects across tasks and individuals, we summarized parameters fit to each subject with a regression model that included two terms: one intercept term that captured the average parameter value across array sizes and one slope term that captured the change in the parameter as a function of array size. Coefficients describing the effects of WCV on recall were negative on average, indicating better recall of more chunkable arrays (figure 11G; median intercept = -0.60, permutation $p < 10e-4$). Negative coefficients were also more extreme for larger array sizes, suggesting that subjects appropriately increased chunking as a function of memory load (figure 11H; median WCV slope = -0.17, permutation $p = 0.007$). Across array sizes, there was no main effect of precision modulation to NND (median intercept = 0.001, permutation $p = 0.92$), but precision modulation coefficients tended to be positive for the largest target arrays (figure 11D), which drove an increase in precision modulation coefficients across array size (figure 11H; median slope = 0.20, permutation $p = 0.02$). Overall, there was a slight bias of memory reports towards nearest neighbors (figure 11E&G; median intercept = 0.20, permutation $p = 0.06$), and the magnitude of this bias increased with array size (figure 11H; median slope = 0.15, permutation $p = 0.001$). Furthermore, this bias was modulated according to NND, as predicted by the center-surround chunking model and observed in our empirical data (figure 11F&G; median intercept = 0.42, permutation $p < 10e-4$). There was no significant effect of array size on the magnitude of this modulatory effect (figure 11H; median slope = -0.15, permutation $p < 0.13$). Taken together, these results support the robustness of our behavioral markers for chunking and supply evidence that chunking is more prominent for larger target arrays, in which it could provide the largest advantages (see predictions from figure 4C).

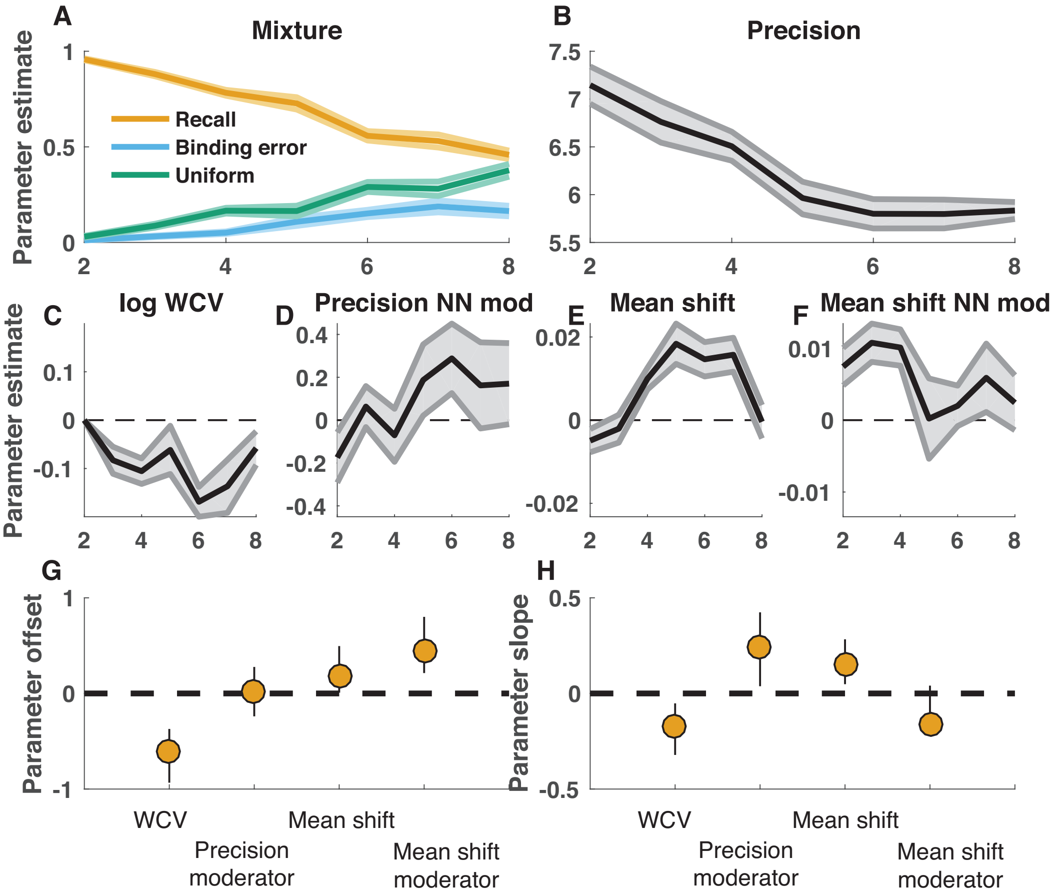


Figure 11: The effects of chunking on recall, precision, and bias are robust across previous working memory studies and depend on overall memory load. The extended mixture model was fit to a meta-analysis dataset that includes delayed memory reports from 101 subjects collected in eight labs using a variety of specific task features (see Methods). **A**) Fit mixture proportions for recall, guess, and binding error distributions are plotted as a function of the number of items in the memory array. **B**) Precision is plotted as a function of memory array size. **C-F**) Coefficients describing the effects of chunking on recall, precision, and bias are plotted as a function of array size. From left to right, plots indicate (**D**) the logistic effect of WCV on recall, (**D**) the modulation of precision by the distance to the closest neighboring feature value, (**E**) the tendency of reports to be biased towards nearest neighbor feature values, and (**F**) the modulation of bias according to the proximity of the nearest neighbor feature value. **G**) Main effects of each chunking coefficient are plotted across memory array size. **H**) Slopes of per-subject chunking coefficients are plotted across memory array size. Positive values indicate larger coefficients for larger memory arrays. Circles and lines in bottom panels reflect median and bootstrapped 95% confidence intervals.

Discussion:

Our work builds on two parallel lines of research. One has focused on how encoding and decoding of working memories is optimized under various statistical contingencies [14-18,48], whereas the other has focused on understanding the nature of capacity limitations in visual working memory [3,4,7,8,11-13]. Here, we explore how people optimize encoding in the same tasks that have formed the basis of our understanding of capacity limitations. Our findings shed light on both the nature of memory capacity limitations and on the encoding strategies employed to minimize their impact.

With regard to encoding strategies, the binary encoding model showed that selective chunking allowed performance advantages that grew as a function of array size (figure 4). This chunking advantage could also be achieved by adding biologically inspired center-surround dynamics to a process model of working memory task performance (figure 5). These dynamics arbitrate a tradeoff between recall and precision, and they predict array-specific inter-item dependencies (figure 9). Human subjects showed the performance benefits predicted by both chunking models and the costs in precision and inter-item dependencies predicted by center-surround chunking in particular (figures 7-9).

These findings are in line with previous work that highlights the effects of center-surround processing on perception and memory, as well as the use of chunking as a mnemonic strategy in a wide range of working memory tasks [22,49,50]. Chunking was first used to describe mnemonic strategies for storage of sequential information, for example, encoding the digits 2-0-0-5 as a single date (2005) rather than as its constituent digits [20,51,52]. In the visual domain, visual features are in some sense chunked into objects [53]. Recent work has suggested that people can chunk arbitrary visual information when that information is inherently clustered and visible for an extended duration [18]. Here, we extend on this work to show that a simple form of chunking, joint encoding of similar feature values, is rapidly implemented by human visual working memory systems to improve performance in tasks that have heretofore been thought to lack exploitable statistical structure.

An important question stemming from this extension is to what extent chunking can be adjusted to optimize working memory accuracy under different conditions. Our modeling shows that a simple learning rule is capable of rapidly adjusting the amount of chunking to optimize performance given the current memory demands, leading to greater chunking for higher memory loads. Consistent with this notion, our meta-analysis shows that recall benefits from chunking are greater for larger memory arrays, suggesting that people selectively engage chunking when it is most advantageous (compare figure 4C with figure 11C). Furthermore, feedback-dependent modulation of chunking behaviors in our experimental data is indicative of online optimization of the chunking process (figure 10D-F), such as the process that allowed learning of the partitioning criterion in the binary encoding model (figure 4G). Yet these trial-to-trial adjustments occur despite only minimal performance improvements across task blocks (figure 7). There are several possible explanations for this discrepancy, including 1) that a priori processing strategies are well-calibrated to our task, 2) that optimization in our task occurs on a different timescale than our measurements, and 3) that the presence of unchunkable arrays hindered learning overall. Distinguishing between these possibilities will require a better understanding of what exactly is being adjusted in response to feedback. For example, reward feedback could promote the prioritization of storing chunked arrays over non-chunkable ones, or it could modulate center-surround inhibition dynamics (e.g., via fine tuning of feature selective attention and/or altering local excitation-inhibition balance [24,54]). In any case, our work, along with other recent research showing an adaptive tradeoff of precision and recall [55], strongly motivate future work to better understand the scope, time course, and mechanism for this type of this optimization process.

While we do not have direct measurements pertaining to the biological underpinnings of the chunking behaviors identified here, the center-surround chunking model provides insight into a potential mechanism. In particular, our descriptive model is based on a mechanistic account in which each item is stored in a modified attractor network that includes sharply tuned recurrent excitation and broadly tuned recurrent inhibition (figure 5) [24,37-39,41,45]. This account is supported by the frequent observation of sustained activity during the delay period of memory tasks in both parietal and prefrontal cortices [56-58] (but see also [59]). Here we have considered the network to store features on a single dimension (color); however, it is clear that at some level, conjunctive coding across features (e.g. color and orientation) is necessary to bind information to the dimension used to probe memories [60]. In our task, it is unknown whether any sustained representations reflect information about the report feature (color in our task), probe feature (orientation in our task), or some conjunction of the two. Recent work has hinted that in some cases, sustained representations in prefrontal cortex may only encode the probe dimension, which could point back to relevant sensory representations at time of recall [59,61-63]. Previous computational instantiations of this process have relied on the basal ganglia to learn appropriate prefrontal representations that can be jointly cued by multiple disparate perceptual features, based on reward feedback [27,62,64]. Analogously, feedback effects observed in our data could be driven by the basal ganglia learning to selectively engage prefrontal units that are prone to representation of multiple probe feature values. This interpretation could expand on a large body of work that implicates the basal ganglia in gating working memory processes by stipulating a novel and testable role for the basal ganglia in optimizing joint feature encoding [25,27,65-67].

Implications for capacity limitations. Working memory limitations have been theorized to result from either a discrete limitation on available "slots" or a continuous limitation by a divisible "resource". The distinction between these theories is most evident when additional targets are added to a memory array. A discrete limitation predicts that after all slots are filled, additional targets will be forgotten and will be reported as random guesses [2]. In contrast, a resource limitation predicts that additional targets will cause each target to be encoded with lower precision [1]. While individual studies have provided support for each theory [3-12,68], a recent meta-analysis provides simultaneous support for the core predictions of both: increasing memory load leads to both increased guessing and decreased precision [13].

Our results suggest that a joint capacity limitation over recall and precision may result from a rational chunking procedure implemented through center-surround dynamics to effectively trade precision for recall (figure 5). This procedure allows subjects to achieve performance improvements for chunkable stimulus arrays at the cost of precision (figures 6-8). It also provides two reasons for precision to decrease with array size: 1) benefits of chunking increase with memory load, incentivizing higher recall probability and lower precision representations (figure 4C&G) and 2) larger memory arrays have a greater probability of containing a non-probed target that biases the encoding of the probed target feature (figure 9D&E). In addition to accounting for known influences on precision, our model also predicted that measured precision should vary across trials, an established feature of human behavioral data [12,29], and correctly predicted that this variability in precision should depend on the features of non-probed targets (figure 9D&E).

However, our findings call the interpretation of precision measurements into question. The center-surround model predicts that internal representations apply attractive and repulsive forces to one another, systematically biasing memory reports. When averaged across trials with differing stimulus configurations, such interactions are interpreted as variability in memory reports, as the net forces on a probed target vary randomly from one stimulus configuration to the next. Yet, since much of this variability is simply an artifact of averaging across disparate conditions, our work raises an important question: how much of the variability in memory reports across trials and individuals is truly reflective of imprecision, rather than bias? While the notion that imprecision can emerge from systematic inter-item dependencies is somewhat at odds with the basic resource limitation model, it is consistent with the recent proposal of a specific form of resource limitation in which the constrained resource is the representational space itself [69,70].

Within such a framework, it is interesting to reconsider the meaning of individual differences in memory storage recall and precision. Previous work has shown that individual differences in the number of items successfully retained in visual working memory tasks, but not differences in precision, are related to fluid intelligence and psychiatric conditions such as schizophrenia, among other factors [71,72]. These results have been interpreted in terms of differences in a discrete capacity for memory storage, or in filtering irrelevant information [73], but our results raise questions regarding whether some of these individual differences may be driven instead by differences in chunking behavior or the optimization thereof (figure 6D). In particular, performance-based measures of capacity may be sensitive to individual differences in lateral connectivity profiles that favor a spectrum from independent to merged feature storage policies, and to the ability to override such policies through learned top-down modulation of lateral connectivity [74-76].

In summary, our results show that humans readily exploit chunking strategies to improve performance on visual working memory tasks. The implementation of chunking is consistent with a form of center-surround dynamics that combines similar representations and facilitates mutual repulsion of disparate ones. This implementation leads to a fundamental tradeoff between the number of items stored and the precision with which they are stored, providing a natural bridge between slots and resource accounts of working memory capacity limitations. People optimize this tradeoff from trial-to-trial according to stimulus statistics and evaluative feedback. These results provide the first joint account of how and why discrete and continuous factors contribute to working memory capacity limits.

Methods:

Delayed report task. Human subjects completed five blocks (100 trials each) of a delayed report color reproduction task (figure 2). Each trial of the task consisted of four primary stages: stimulus presentation, delay, probe, and feedback. During stimulus presentation, subjects were shown five oriented bars (length = 2 degrees visual angle) arranged in a circle (radius = 4 degrees visual angle) centered on a fixation point. Bar positions were equally spaced around the circle and jittered uniformly from trial to trial. Bar orientations were uniformly spaced, jittered from trial to trial, and independent of position or color. Bar colors were chosen from a fixed set of colors corresponding to a circle in CIELAB color space (L= 80, radius in a*, b* = 60) and referred to by angular position for convenience. In the “random spacing” condition, all five colors were sampled independently of one another from the color space, allowing for the possibility of two similar colors in the same stimulus array. In the “fixed spacing” condition, colors were uniformly spaced along the CIELAB color wheel and randomly assigned to bar locations. Stimuli were presented for 200 msec, after which the screen was blanked.

The subsequent delay period lasted 900 msec, during which subjects were forced to remember the colors and orientations of the preceding stimulus array. During the subsequent probe stage, subjects were shown a gray oriented bar in the center of the screen for one second, before being asked to report the color that had been associated with that orientation in the preceding stimulus array. Color reports were made by adjusting the color of the oriented bar using a mouse. The initial position of the mouse on the color wheel was randomly initialized on each trial. On a subset (1/3) of trials subjects were asked to make a post-decision wager about the accuracy of their report by choosing to bet either 0 or 2 points. Binary feedback was provided on each trial based on whether subject reporting accuracy fell within a certain error tolerance window (π /3 radians – low precision condition or π /8 radians – high precision condition). Subjects were paid bonuses according to total accumulated points. All human subject procedures were approved by the Brown University Institutional Review Board and conducted in agreement with the Declaration of Helsinki.

Binary encoding model. To explore the potential advantage of chunking in delayed report tasks, we developed a flexible and computationally tractable model for capacity-limited storage. This model stores color and orientation information symbolically in a set of binary "words" concatenated to form a "sentence". During the stimulus presentation phase, target colors and orientations are "encoded" as an alternating sequence of binary words reflecting the position on a circular feature space (figure 3). The number of binary digits (bits) in a word controls the precision with which the feature is stored. For example, a single digit can encode which half of the feature space contains the color of a bar, whereas three bits can narrow the stimulus color down to one eighth of the color space (figure 3, top). Each binary word is followed by a "stop" symbol denoting the type of information in the preceding word (e.g. color or orientation). A capacity limitation is implemented in the model as a limit on the number of bits that can be stored in memory. Specifically, we applied a fixed limit of 15 bits for storage of color information. Similar results were achieved by applying a limit to the total bits, i.e. including orientation information, but here we allow for perfect orientation storage in order to isolate the effects of capacity limitations on the recall dimension (color). Bits were evenly distributed among represented colors, as this strategy for allocation of bits is optimal for standard cost functions (e.g. minimizing squared or absolute errors).

During the probe phase, the model is presented with a single orientation and recalls the color word that immediately precedes that orientation in the stored binary sentence. A report is then sampled from a uniform distribution across the range of colors consistent with that stored binary color word. For example, if the color word contains one, two, or three bits, it is sampled from uniform distribution over one half, quarter, or eighth of the color space.

Chunking was parametrically implemented in the binary encoding model by adding a "partitioning criterion" that specifies the minimum distance between two colors in color space that is necessary for independent storage. Colors separated by distances smaller than the partitioning criterion are "chunked" into a single color representation. The distance computation is completed during the "encoding" phase, before colors are converted to binary words. Distances are corrupted with a small amount of noise consistent with variability in the visual representation or the chunking processes (normally distributed with standard deviation equal to 0.4 times the partitioning criterion). After chunking, bits are allocated evenly across all represented colors, as described above.

Model performance was simulated for the delayed estimation task across eight different array sizes [1-8] with two different color generation conditions (fixed- and random-spacing) for nine different partitioning criterions ranging from zero to π. For each condition and model, mean absolute error was computed across 5000 simulated trials. The "best-chunking" model was defined as the model with the lowest mean absolute error, whereas the "non-chunking" model was the model with partitioning criterion equal to zero (such that every color was stored independently). For each condition, chunking bonus was computed as the difference in absolute error between the non-chunking and best-chunking models.

For the trial-to-trial optimization of the partitioning criterion (figure 4g), we adjusted the partitioning criterion on each trial according to the following rule:

$$PC = PC - \ \alpha\, \delta\, \Delta C$$

where PC is the partitioning criterion, α is a learning rate, δ is a reward prediction error (previous trial feedback minus long term average feedback), and ΔC is the number of "chunks" into which the previous stimulus array was divided minus the long term average of that quantity. Thus, if by chance the model did more chunking on a given trial, the ΔC would take a negative value, and positive feedback would drive a positive δ and a corresponding increase in the partitioning criterion, leading to an increase in chunking on subsequent trials. Negative feedback for the same trial would lead to a negative δ and corresponding decrease in the partitioning criterion, leading to a decrease in chunking on subsequent trials.

Center-surround chunking model. To determine the effects that center-surround dynamics would have on visual working memory task performance, we extended the standard descriptive model of delayed memory reports to incorporate features of center-surround dynamics. In particular, on each trial, internal representations of each color were generated from a von Mises distribution with fixed concentration (7 for simulations). Pairwise distances (in color space) were computed for each pair of internal representations. Chunking probability was computed as a scaled von Mises function of this distance ($\mu = 0$, $\kappa = 12$ for simulation), corresponding to the narrow excitatory "center" over which local representations are likely to attract one another (figure 5A-C). Representations were merged in accordance with these chunking probabilities by replacing the color associated with each merged representation with the mean of the merged colors. After probabilistic chunking, distances were recomputed between representations, and each representation applied a repulsive force on neighboring representations as defined by a scaled von Mises function of the re-computed distance ($\mu = 0$, $\kappa = 2$ for simulation), corresponding to the broadly tuned "surround" over which representations repulse one another (figure 5A-C). Applying these forces leads each representation to be reset according to the following equation:

$$color_i \longleftarrow color_i + \sum_{X \neq i} W_{surround} \ \left(\frac{e^{\kappa \cos(color_i - color_x)}}{2\pi \mathrm{I}_0(k)}\right) \left(\frac{|color_i - color_x|}{color_i - color_x}\right)$$

Where $W_{surround}$ is a weight that controls the overall magnitude of surround effects, the second term in the sum is the probability density function for a von Mises distribution, and the final term serves to ensure that targets exert repulsive forces on neighboring targets. For the simulations in figure 5, the weight parameters for both center and surround were set to equal values ranging from 0 to 0.7. For comparisons to subject data, $W_{surround}$ was set to 0.6 and W_{center} was set to 1.2.

Probabilistic recall was implemented in the model according to a Poisson memory process [13,16]. On each trial, the model accurately recalled some number of representations drawn from a Poisson distribution ($\lambda = 2$ for simulations). Similar results were achieved using an inhibition based forgetting process inspired by Wei and colleagues [24]; however, here we use a more standard Poisson process for simplicity. In the case that a representation that was not successfully recalled was probed, the model reported either a uniformly distributed guess (p = 0.65) or the color of an alternative representation (binding error, p = 0.35).

Computing array "chunkability". In order to assess the potential benefits of chunking on each trial we computed a clustering statistic, within cluster variance (WCV), for each stimulus array. WCV was computed by dividing the array colors into two clusters that minimized within cluster variance (see figure 4). WCV was defined as the average circular variance over colors within these clusters.

Logistic regression models. Binary accuracy and betting data were concatenated across all subjects and interrogated with a mixed-effects logistic regression model that included terms to account for fixed effects of 1) –log(WCV), a proxy for stimulus array chunkability, 2) the color distance between the probed target and each other color in the array, ordered from smallest to largest, 3) feedback on previous and subsequent trials, 4) spatial distance between the location of the probed target and the location of the previously probed target, and 5) task block. In addition, the model included dummy variables to account for random intercepts specific to individual subjects.

Nearest neighbor analysis. For each trial, the nearest neighbor color was identified as the color of the non-probed target that was most similar to that of the probed target. Target colors and subject reports were transformed for each trial such that the probed target color corresponded to zero and the nearest neighbor color ranged from zero to π. Trials were then sorted according to absolute nearest neighbor distance (see supplementary figure 1) and binned in sliding windows of 50 trials according to nearest neighbor distance. Binned data was combined across all subjects and fit with a mixture model that assumed data were generated from a mixture of 1) a von Mises distributed memory report (free parameters: mean, precision, and mixture weight), 2) uniformly distributed guesses (free parameters: mixture weight), and 3) binding errors that were von Mises distributed and centered on non-probed targets (no free parameters required, as mixture weight forms simplex with the other mixture components). Maximum posterior probability parameter estimates for the mixture model fits to subject and model simulated data are reported in figure 9 (prior distributions for all modulator terms were centered on zero with σ = 0.5 for recall modulators, σ = 2 for precision modulators, and σ = 0.05 for bias modulators).

Mixture model. We extended the standard mixture model of memory reports [3,11] to allow for modulation of recall probability, precision, and bias according to WCV, nearest neighbor distance, and feedback. The standard mixture model assumes reports are generated from a mixture of "correct recall", "guessing", and "binding error" processes. These three mixture components were specified using two free parameters: one dictating the probability with which an item would be successfully stored $(correct\ recall + binding\ error)$ and one specifying the probability with which a stored item would be correctly reported $\left(\frac{correct\ recall}{correct\ recall+binding\ error}\right)$. We allowed the parameter dictating successful storage to be modified as a logistic function of 1) log(WCV), 2) previous feedback, 3) previous log(WCV), 4) previous feedback*log(WCV), 5) previous feedback*log(WCV)*previous log(WCV). All potential modulators of successful storage were mean-centered (before and after interaction) and constrained by priors favoring values near zero [~normal(0, 0.5)]. Since our successful storage parameter is the probability the subject will not elicit a uniform guess, it affects both correct recall and binding error mixture components. However, since reports were far more likely to correspond to correct recalls (median mixture proportion = 0.50 across subjects) than binding errors (median binding error proportion = 0.17 = across subjects), modulator coefficients had larger effects on recall than binding errors, and we refer to them in the results as modulating recall for simplicity. We also considered an alternative model in which modulators affected the recall term directly and found similar results, although this alternative model provided a worse overall fit of the data.

The model also contained a modulator term to account for changes in precision occurring as a function of the distance to the nearest neighboring target color (see figure 9D). Trial-wise precision estimates were generated by interpolating the basic fits displayed in figure 9D using a leave-one-subject-out procedure to ensure that precision estimates for each subject were not informed by their own data. These precision estimates were then included as a linear modulator of the precision parameter in the mixture model, which was constrained to take values near zero using normal priors [~normal(0, 2)].

Finally, the model also included two terms to allow the mean of the reported response distribution to be shifted 1) a fixed amount toward the nearest neighbor color and 2) according to the distance-dependent function in figure 9A, computed using the same leave-one-subject-out procedure described above.

Meta-analysis. In order to test the robustness of our findings and determine how the behavioral hallmarks of chunking scale with the size of the stimulus array, we applied a modified version of our mixture model to a meta-analysis dataset. The meta-analysis dataset included eight studies and a total of 101 subjects [3,11,12,28,72,77].

Seven of the datasets, available online at http://www.cns.nyu.edu/malab/resources.html, were originally compiled by van den Berg et al. and have previously been described in detail [13]. Three of the studies compiled by Van den Berg et al. were excluded from our analyses due to retraction of the original studies, although the inclusion of these studies did not qualitatively change our results. The eighth dataset (28 subjects) comprised the control subjects in a psychiatric comparative study of visual working memory [72]. Each study differed in experimental details but involved a delayed report working memory task with at least two different array sizes.

We analyzed the data for each subject and array size separately using a simplified version of the mixture model described above. The simplified version did not include any parameters related to the feedback or stimulus configuration on the previous trial, as most of the included studies did not provide feedback. For each subject and parameter in the model, parameter estimates were analyzed with a subject-level regression to quantify 1) the mean parameter value across array sizes (intercept) and 2) the extent to which parameter estimates increased with array size (slope). A bootstrapping procedure was used to non-parametrically estimate 95% confidence intervals over the median slope and intercept for each parameter.

Acknowledgements:
We thank Ronald van den Berg, Wei Ji Ma, and James Gold for sharing data from previously published delayed report studies. We thank Anish Aitharaju, Anthony Jang, Ji Sun Kim, Michelle Kulowski, and Ezra Nelson for aiding in data collection. We thank Nicholas Franklin and Nathan Vierling-Claasen for helpful discussion. This work was funded by NIMH grant F32 MH102009 (MRN), NIMH grant R01 MH080066-01 (MJF), and NSF grant #1460604 (MJF).

Competing interests:
The authors declare no competing interests.

References:

1. Ma, W. J., Husain, M. & Bays, P. M. Changing concepts of working memory. *Nature Neuroscience* **17,** 347–356 (2014).
2. Luck, S. J. & Vogel, E. K. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences* **17,** 391–400 (2013).
3. Zhang, W. & Luck, S. J. Discrete fixed-resolution representations in visual working memory. *Nature* **453,** 233–235 (2008).
4. Bays, P. M. & Husain, M. Dynamic shifts of limited working memory resources in human vision. *Science* **321,** 851–854 (2008).
5. Rouder, J. N. *et al.* An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences* **105,** 5975–5979 (2008).

6. Cowan, N. & Rouder, J. N. Comment on "Dynamic shifts of limited working memory resources in human vision". *Science* **323,** 877–author reply 877 (2009).
7. Zhang, W. & Luck, S. J. Sudden death and gradual decay in visual working memory. *Psychol Sci* **20,** 423–428 (2009).
8. Zhang, W. & Luck, S. J. The number and quality of representations in working memory. *Psychol Sci* **22,** 1434–1441 (2011).
9. Donkin, C., Tran, S. C. & Nosofsky, R. Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory. *Atten Percept Psychophys* (2013). doi:10.3758/s13414-013-0561-7
10. Donkin, C., Nosofsky, R. M., Gold, J. M. & Shiffrin, R. M. Discrete-slots models of visual working-memory response times. *Psychological Review* **120,** 873–902 (2013).
11. Bays, P. M., Catalao, R. F. G. & Husain, M. The precision of visual working memory is set by allocation of a shared resource. *J Vis* **9,** 7.1–11 (2009).
12. van den Berg, R., Shin, H., Chou, W.-C., George, R. & Ma, W. J. Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences* **109,** 8780–8785 (2012).
13. van den Berg, R., Awh, E. & Ma, W. J. Factorial comparison of working memory models. *Psychological Review* **121,** 124–149 (2014).
14. Brady, T. F., Konkle, T. & Alvarez, G. A. Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General* **138,** 487–502 (2009).
15. Brady, T. F. & Alvarez, G. A. Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychol Sci* **22,** 384–392 (2011).
16. Sims, C. R., Jacobs, R. A. & Knill, D. C. An ideal observer analysis of visual working memory. *Psychological Review* **119,** 807–830 (2012).
17. Orhan, A. E. & Jacobs, R. A. A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review* **120,** 297–328 (2013).
18. Lew, T. F. & Vul, E. Ensemble clustering in visual working memory biases location memories and reduces the Weber noise of relative positions. *J Vis* **15,** 10 (2015).
19. Brady, T. F. & Alvarez, G. A. Contextual effects in visual working memory reveal hierarchically structured memory representations. *J Vis* **15,** 6 (2015).
20. Cowan, N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav Brain Sci* **24,** 87–114– discussion 114–85 (2001).
21. Cavanaugh, J. R. Selectivity and Spatial Distribution of Signals From the Receptive Field Surround in Macaque V1 Neurons. *Journal of Neurophysiology* **88,** 2547–2556 (2002).
22. Xing, J. & Heeger, D. J. Measurement and modeling of center-surround suppression and enhancement. *VISION RESEARCH* **41,** 571–583 (2001).
23. Weliky, M., Kandler, K., Fitzpatrick, D. & Katz, L. C. Patterns of excitation and

inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns. *Neuron* (1995).

24. Wei, Z., Wang, X.-J. & Wang, D.-H. From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *Journal of Neuroscience* **32,** 11228–11240 (2012).
25. O'Reilly, R. C. & Frank, M. J. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput* **18,** 283–328 (2006).
26. Todd, R. M., Niv, Y. & Cohen, J. D. Learning to use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. *Advances in neural information processing systems* 1–8 (2008).
27. Collins, A. G. E. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological Review* **120,** 190–229 (2013).
28. Wilken, P. & Ma, W. J. A detection theory account of change detection. *J Vis* **4,** 11–11 (2004).
29. Fougnie, D., Suchow, J. W. & Alvarez, G. A. Variability in the quality of visual working memory. *Nat Commun* **3,** 1229 (2012).
30. Fuster, J. M. & Jervey, J. P. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science* **212,** 952–955 (1981).
31. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14,** 477–485 (1995).
32. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24,** 167–202 (2001).
33. Curtis, C. E. & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences* **7,** 415–423 (2003).
34. Warden, M. R. & Miller, E. K. The Representation of Multiple Objects in Prefrontal Neuronal Delay Activity. *Cerebral Cortex* **17,** i41–i50 (2007).
35. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol. (Lond.)* **148,** 574–591 (1959).
36. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology* **28,** 229–289 (1965).
37. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological cybernetics* (1982).
38. Somers, D. C., Nelson, S. B. & Sur, M. An emergent model of orientation selectivity in cat visual cortical simple cells. *J. Neurosci.* **15,** 5448–5465 (1995).
39. Ben-Yishai, R., Bar-Or, R. L. & Sompolinsky, H. Theory of orientation tuning in visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **92,** 3844–3848 (1995).
40. Kiyonaga, A. & Egner, T. Center-Surround Inhibition in Working Memory. *Curr. Biol.* **26,** 64–68 (2016).
41. Wang, X. J. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *Journal of Neuroscience* **19,** 9587–9603 (1999).
42. Durstewitz, D. & Seamans, J. K. The computational role of dopamine D1

receptors in working memory. *Neural Netw* **15,** 561–572 (2002).

43. Barak, O., Sussillo, D., Romo, R., Tsodyks, M. & Abbott, L. F. Progress in Neurobiology. *Progress in Neurobiology* **103,** 214–222 (2013).
44. Kilpatrick, Z. P., Ermentrout, B. & Doiron, B. Optimizing Working Memory with Heterogeneity of Recurrent Cortical Excitation. *Journal of Neuroscience* **33,** 18999–19011 (2013).
45. Murray, J. D. *et al.* Linking Microcircuit Dysfunction to Cognitive Impairment: Effects of Disinhibition Associated with Schizophrenia in a Cortical Working Memory Model. *Cerebral Cortex* **24,** 859–872 (2014).
46. Felsen, G., Touryan, J. & Dan, Y. Contextual modulation of orientation tuning contributes to efficient processing of natural stimuli. *Network* **16,** 139–149 (2005).
47. Fallon, S. J., Zokaei, N. & Husain, M. Causes and consequences of limitations in visual working memory. *Annals of the New York Academy of Sciences* n/a–n/a (2016). doi:10.1111/nyas.12992
48. Brady, T. F. & Tenenbaum, J. B. A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review* **120,** 85–109 (2013).
49. Johnson, J. S., Spencer, J. P., Luck, S. J. & Schöner, G. A Dynamic Neural Field Model of Visual Working Memory and Change Detection. *Psychol Sci* **20,** 568–577 (2009).
50. Cowan, N. The magical number 4 in short-term memory: A reconsiderationof mental storage capacity. 1–99 (2001).
51. Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* (1956).
52. Chen, Z. & Cowan, N. Chunk Limits and Length Limits in Immediate Recall: A Reconciliation. *J Exp Psychol Learn Mem Cogn* **31,** 1235–1249 (2005).
53. Luria, R. & Vogel, E. K. Come Together, Right Now: Dynamic Overwriting of an Object's History through Common Fate. *J Cogn Neurosci* 1–11 (2014). doi:10.1038/nature02447
54. Störmer, V. S. & Alvarez, G. A. Feature-Based Attention Elicits Surround Suppression in Feature Space. *Current Biology* **24,** 1985–1988 (2014).
55. Fougnie, D., Cormiea, S. M., Kanabar, A. & Alvarez, G. A. Strategic Trade-Offs Between Quantity and Quality in Working Memory. *Journal of Experimental Psychology: Human Perception and Performance* (2016). doi:10.1037/xhp0000211
56. Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* (1971).
57. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* **61,** 331–349 (1989).
58. Gottlieb, J. Simultaneous Representation of Saccade Targets and Visual Onsets in Monkey Lateral Intraparietal Area. *Cerebral Cortex* **15,** 1198–1206 (2004).
59. Lara, A. H. & Wallis, J. D. Executive control processes underlying multi-item working memory. *Nature Publishing Group* **17,** 876–883 (2014).
60. Matthey, L., Bays, P. M. & Dayan, P. A Probabilistic Palimpsest Model of Visual

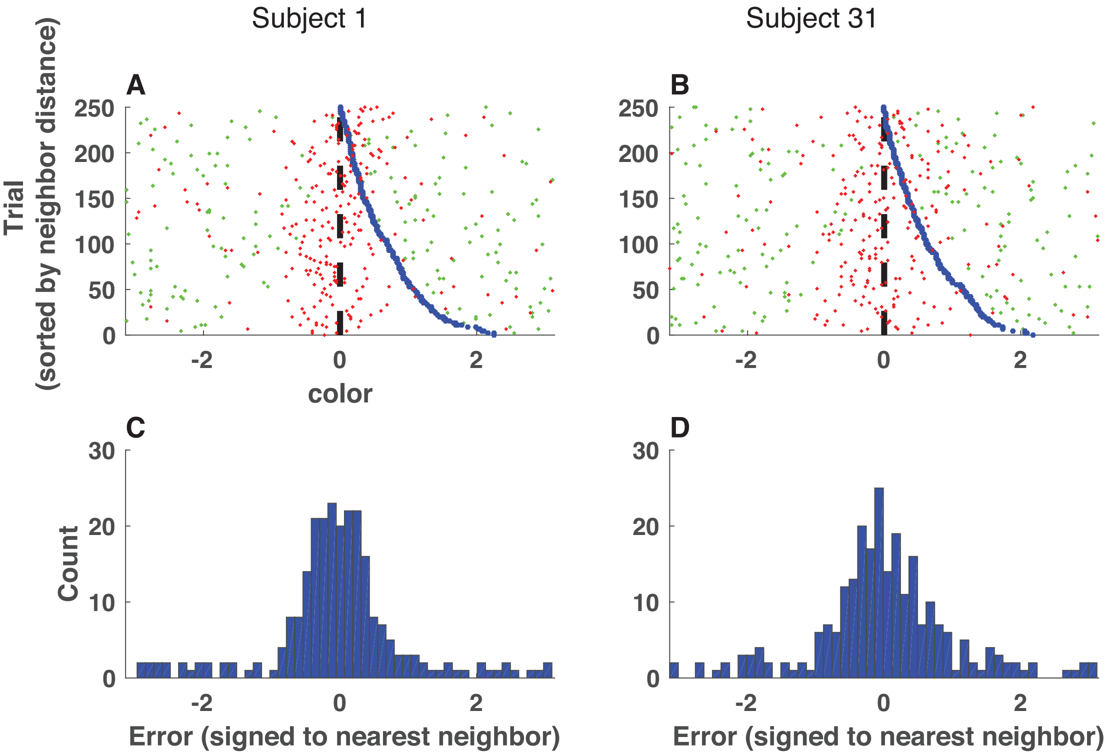
Short-term Memory. *PLoS Comput Biol* **11,** e1004003 (2015).

61. Lara, A. H. & Wallis, J. D. The Role of Prefrontal Cortex in Working Memory: A Mini Review. *Front Syst Neurosci* **9,** 173 (2015).
62. Kriete, T., Noelle, D. C., Cohen, J. D. & O'Reilly, R. C. Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences* **110,** 16390–16395 (2013).
63. Ester, E. F., Sprague, T. C. & Serences, J. T. Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron* **87,** 893–905 (2015).
64. Frank, M. J. & Badre, D. Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis. *Cerebral Cortex* **22,** 509–526 (2012).
65. Voytek, B. & Knight, R. T. Prefrontal cortex and basal ganglia contributions to visual working memory. *Proceedings of the National Academy of Sciences* **107,** 18167–18172 (2010).
66. Hazy, T. E., Frank, M. J. & O'Reilly, R. C. Banishing the homunculus: Making working memory work. *Neuroscience* **139,** 105–118 (2006).
67. Chatham, C. H., Frank, M. J. & Badre, D. Corticostriatal Output Gatingduring Selection from Working Memory. *Neuron* **81,** 930–942 (2014).
68. Pratte, M. S., Park, Y. E., Rademaker, R. L. & Tong, F. Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance* **43,** 6–17 (2017).
69. Franconeri, S. L., Alvarez, G. A. & Cavanagh, P. Flexible cognitive resources:competitive content maps for attentionand memory. *Trends in Cognitive Sciences* **17,** 134–141 (2013).
70. Cohen, M. A., Rhee, J. Y. & Alvarez, G. A. Limits on perceptual encoding can be predicted from known receptive field properties of human visual cortex. *Journal of Experimental Psychology: Human Perception and Performance* **42,** 67–77 (2016).
71. Fukuda, K., Vogel, E., Mayr, U. & Awh, E. Quantity, not quality: the relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review* **17,** 673–679 (2010).
72. Gold, J. M. *et al.* Reduced capacity but spared precision and maintenance of working memory representations in schizophrenia. *Arch. Gen. Psychiatry* **67,** 570–577 (2010).
73. Vogel, E. K., McCollough, A. W. & Machizawa, M. G. Neural measures reveal individual differences in controlling access to working memory. *Nature* **438,** 500–503 (2005).
74. Lowe, M. X. *et al.* Sensory processing patterns predict the integration of information held in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance* **42,** 294–301 (2016).
75. Freeman, E., Sagi, D. & Driver, J. Lateral interactions between targets and flankers in low-level vision depend on attention to the flankers. *Nature Neuroscience* **4,** 1032–1036 (2001).
76. Freeman, E., Driver, J., Sagi, D. & Zhaoping, L. Top-Down Modulation of Lateral

Interactions in Early Vision. *Current Biology* **13,** 985–989 (2003).

77. Rademaker, R. L., Tredway, C. H. & Tong, F. Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *J Vis* **12,** 21 (2012).

Supplementary Information:



Supplementary Figure 1: Sorting trials according to the nearest neighbor non-probed target color reveals structure in memory reports. A&B: Signed error of memory reports (red points) for all trials completed by two sample subjects (left = subject 1, right = subject 46). Trial errors are sorted by the distance from the probed target to the most similar color in the target array (nearest neighbor distance, NND) and transformed according to the direction of the nearest neighbor target (blue points). Green points reflect the positions of other colors in the target array, relative to the probed color and transformed as described above. Note the asymmetry in error distributions appears to change as a function of the nearest neighbor distance. **C&D:** Error histograms for the same two example subjects, transformed as described above. Note that in some cases apparent structure in the sorted errors (A) is no longer visible after collapsing across nearest neighbor distances (C).