Journal of Experimental Psychology: General
2009, Vol. 138, No. 4, 487–502

0096-3445/09/$12.00 DOI: 10.1037/a0016797

Compression in Visual Working Memory: Using Statistical Regularities to Form More Efficient Memory Representations

Timothy F. Brady and Talia Konkle
Massachusetts Institute of Technology

George A. Alvarez
Harvard University

The information that individuals can hold in working memory is quite limited, but researchers have typically studied this capacity using simple objects or letter strings with no associations between them. However, in the real world there are strong associations and regularities in the input. In an information theoretic sense, regularities introduce redundancies that make the input more compressible. The current study shows that observers can take advantage of these redundancies, enabling them to remember more items in working memory. In 2 experiments, covariance was introduced between colors in a display so that over trials some color pairs were more likely to appear than other color pairs. Observers remembered more items from these displays than from displays where the colors were paired randomly. The improved memory performance cannot be explained by simply guessing the high-probability color pair, suggesting that observers formed more efficient representations to remember more items. Further, as observers learned the regularities, their working memory performance improved in a way that is quantitatively predicted by a Bayesian learning model and optimal encoding scheme. These results suggest that the underlying capacity of the individuals' working memory is unchanged, but the information they have to remember can be encoded in a more compressed fashion.

Keywords: visual short-term memory, chunking, information theory, memory capacity, statistical learning

Every moment, a large amount of information from the world is transmitted to the brain through the eyes, ears, and other sensory modalities. A great deal of research has examined how the perceptual and cognitive system handles this overwhelming influx of information (e.g., Neisser, 1967). Indeed, this information overload provides the motivating intuition for why we need selective attention: to actively filter out irrelevant input to allow specific processing of the intended stimuli (Broadbent, 1958). However, since the world is filled with regularities and structure, the information transmitted to the brain is also filled with regularities (Barlow, 1989). In quantitative terms, there is significant redundancy in the input (Huffman, 1952; Shannon, 1948). An intuitive example of the redundancy in the visual input is to consider all the possible images that could be made from an 8×8 grid where any pixel can be any color. Most of the images will look like noise, and only a very tiny percentage of these images will actually look like a picture of the real world (Chandler & Field, 2007). This indicates that real-world images are not randomly structured, and in fact share many structural similarities (e.g., Burton & Moorehead, 1987; Field, 1987; Frazor & Geisler, 2006). Interestingly, computationally efficient representations of image-level redundancy produce basis sets that look remarkably like primary visual cortex, providing evidence that our visual perceptual system takes advantage of this redundancy by tuning neural response characteristics to the natural statistics of the world (Olshausen & Field, 1996).

Being sensitive to the statistics of the input has direct consequences for memory as well as for perception (Anderson & Schooler, 2000). Recent work on the rational analysis of memory, for example, suggests that the power laws of forgetting and practice approximate an optimal Bayesian solution to the problem of memory retrieval given the statistics of the environment (Anderson & Schooler, 1991; see also Shiffrin & Steyvers, 1997, 1998). Here we apply similar principles of rational analysis (Chater & Oaksford, 1999) to the capacity of the working memory system. We focus on the abstract computational problem being solved by the working memory system: the storage of as much information as possible in the limited space available.

Working Memory Capacity and Redundancy

According to information theory, in an optimal system more content can be stored if there are redundancies in the input (Cover & Thomas, 1991). In other words, if the input contains statistical structure and regularities, then each piece of information we encode limits the likely possibilities for the remaining information (e.g., given a *q*, the next letter is likely to be *u*). This makes it possible to encode more items in less space. If the human working

Timothy F. Brady and Talia Konkle, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology; George A. Alvarez, Department of Psychology, Harvard University.

This research was supported by a National Institutes of Health/National Eye Institute fellowship F32EY016982 to George A. Alvarez, a National Defense Science and Engineering Graduate fellowship awarded to Talia Konkle, and a National Science Foundation graduate research fellowship awarded to Timothy F. Brady.

We would like to thank Aude Oliva for additional funding and Todd Thompson for valuable insight and discussion. A portion of this work was presented at the annual meeting of the Cognitive Sciences Society, July 2008.

Correspondence concerning this article should be addressed to Timothy F. Brady, Department of Brain and Cognitive Sciences, 46-4078, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. E-mail: tfbrady@mit.edu

memory system approximates an optimal memory system, it should be able to take advantage of statistical regularities in the input in order to encode more items into working memory.

However, while the capacity of short-term and working memory has been extensively studied (e.g., Alvarez & Cavanagh, 2004; Baddeley, 1986; Cowan, 2001, 2005; Zhang & Luck, 2008), little formal modeling has been done to examine the effects of redundancy on the system. Nearly all studies on visual working memory have focused on memory for arbitrary pairings or novel stimuli. While some studies have investigated the effects of associative learning on visual working memory capacity (Olson & Jiang, 2004; Olson, Jiang, & Moore, 2005), they have not provided clear evidence for the use of redundancy to increase capacity. For example, one study found evidence that learning did not increase the amount of information remembered, but that it improved memory performance by redirecting attention to the items that were subsequently tested (Olson et al., 2005).

Chunking

However, the effects of redundancy on working memory capacity have been well studied through the phenomenon of *chunking*, particularly in verbal working memory (Cowan, 2001; Miller, 1956; Simon, 1974). Cowan (2001) defined a chunk as a group of items where the intrachunk associations are greater than the interchunk associations. In other words, in the sequence *FBICIA* the letters *F*, *B*, and *I* are highly associated with each other and the letters *C*, *I*, and *A* are highly associated with each other, but the letters have fewer associations across the chunk boundaries. Thus, observers are able to recall the sequence using the chunks *FBI* and *CIA*, effectively taking up only two of the four chunks that people are able to store in memory (Cowan, 2001; Cowan, Chen, & Rouder, 2004). By comparison, when the letters are random, say *HSGABJ*, they are more difficult to remember, since it is more difficult to chunk them into coherent, associated units.

Chunking is not usually framed as a form of compression analogous to information theoretic views. In fact, in the seminal work of Miller (1956), chunking and information theoretic views of memory were explicitly contrasted, and the most naïve information theoretic view was found lacking in its ability to explain the capacity of working memory. However, at its root chunking approximates a form of compression: It replaces highly correlated items (which are therefore highly redundant) with a single chunk that represents all of the items. Thus, it is possible to frame the strategy of chunking as a psychological implementation of a broader computational idea: removal of redundancy to form compressed representations and allow more items to be stored in memory. At this level of description, chunking is compatible with information theoretic analyses. In fact, information theory and Bayesian probability theory may be able to explain exactly when human observers will form a chunk in long-term memory (e.g., Orbán, Fiser, Aslin, & Lengyel, 2008), in addition to how useful that chunk will be to subsequent working memory tasks. Thus, information theory may be not only compatible with chunking but in fact may provide useful constraints on theories of chunking.

In the present experiments we asked whether human observers learn and use regularities in working memory in a way that is compatible with an information theoretic compression analysis. In two experiments we presented observers with displays of colors that were either random or patterned. By presenting regularities in the displays over the course of the experiment, we examined if and how observers take advantage of these regularities to form more efficient representations. We then present a quantitative model of how learning occurs and how the stimuli are encoded using the learned regularities. We show that more items can be successfully stored in visual working memory if there are redundancies (patterns) in the input. We also show that this learning is compatible with the compressibility of the displays according to information theory.

Experiment 1: Regularities Within Objects

In classic visual working memory experiments, the stimuli used are generally colored oriented lines, shapes, and colored circles, and the aim is to quantify how many objects or features can be remembered. In one of the seminal articles in this field, Luck and Vogel (1997) proposed that people can remember four objects no matter how many features they contain. This view has since been tempered, with some arguing for independent storage of different feature dimensions (Magnussen, Greenlee, & Thomas, 1996; Olson & Jiang, 2002; Wheeler & Treisman, 2002; Xu, 2002) and others arguing for more graded representations, in which information load determines how many objects can be stored (Alvarez & Cavanagh, 2004; Bays & Husain, 2008). However, nearly all current work emphasizes that at best three or four features from a given stimulus dimension can be encoded successfully.

Here we modify the standard paradigm by introducing regularities in the displays for some observers. One group of participants was presented with colors drawn randomly, as in classical visual working memory tasks, such that all possible pairs of colors were equally likely to occur. A second group of participants was presented with colors that occurred most often paired with another color. For example, a particular observer might see red most often around yellow, white most often around blue, whereas a smaller percentage of the time these colors appeared with any other color. Because this manipulation introduces redundancy into the displays, in information theoretic terms these displays contain less information. An information theoretic view of memory therefore predicts that the observers presented with regularities should be able to encode more items into memory.

Method

Observers. Twenty naïve observers were recruited from the Massachusetts Institute of Technology (MIT) participant pool (age range 18–35) and received $10 for their participation. All observers gave informed consent.

Procedure. Observers were presented with displays consisting of four objects around the fixation point (see sample display in Figure 1). Each object was made up of two different colored circles, with one circle inside the other. Observers were informed that their task was to remember the locations of each of the eight colors. At the start of a trial, the colors appeared and remained visible for 1,000 ms. Then the colors disappeared, and placeholder circles were present for the next 1,000 ms (long enough to prevent observers from relying on iconic memory; Sperling, 1960), and then either the inside or outside circle on a random object was darkened.

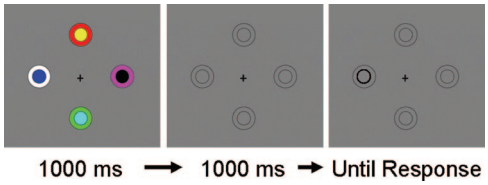

Figure 1. A sample trial from Experiment 1. Eight colors were presented as four two-color objects. The colors disappeared for 1 s and then either the inside or outside of an object was cued. Observers had to indicate what color was at the cued location.

The task was to indicate which of the eight colors had been presented at the indicated location, by pressing one of eight color-coded keys. Observers completed 600 trials, presented in 10 blocks of 60 trials each. Afterward, they completed a questionnaire, reporting the strategies they employed and whether they noticed the presence of patterns in the displays.

The stimuli were presented using MATLAB software with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). The eight colors used were red, green, blue, magenta, cyan, yellow, black, and white.

Manipulation. Observers were randomly assigned to two groups, *patterned* and *uniform*, which differed in how the colors for each trial were chosen. For observers in the uniform condition, the locations of the colors in each trial were chosen randomly, with only the constraint that each color had to appear exactly once in a display.

For observers in the patterned condition, the stimuli for each trial were not chosen randomly. First, for each subject a joint probability matrix was constructed to indicate how likely each color was to appear inside and outside of each other color. This matrix was made by choosing four high-probability pairs at random (probability = .2151) and then assigning the rest of the probability mass uniformly (probability = .0027). As in the uniform condition, all eight colors were present in each display. In order to achieve this, the diagonal of the joint probability matrix was set to zero in order to prevent the same color from appearing twice in the same display.

The pairs were constrained so that each color was assigned to exactly one high-probability pair. For example, if {blue–outside, red–inside} was a high-probability pair in this joint probability matrix, the observer would often see blue and red appear together, in that configuration. However, blue and red each would also sometimes appear with other colors, or in a different configuration. So, for example, {blue–outside, yellow–inside} and {red–outside, blue–inside} could also appear with low probability. High-probability pairs accounted for approximately 80% of the pairs shown during the experiment, and low-probability pairs constituted the other 20%.

In the final block of the experiment in the patterned condition, the distribution from which the displays were drawn was changed to a uniform distribution. This eliminated the regularities in the display, and allowed us to assess whether observers had used the regularities to improve their performance. Further, this manipulation gives a quantitative measure of learning: the difference in performance between Block 9 and Block 10.

Results

We estimated the number of colors observers could successfully hold in memory using the following formula for capacity given an eight-alternative forced choice (see the Appendix for a derivation of this formula):

$$K = [(\mathrm{PC} \times 8 \times 8) - 8]/7.$$

By correcting for chance we can examine exactly how many colors from each display observers would have had to remember in order to achieve a given percent correct (PC). It should be noted that K is a way of quantifying the number of colors remembered that does not necessarily reflect what observers actually represent about the displays. For instance, observers may remember all eight colors with uncertainty rather than some subset of the colors with perfect certainty (see, e.g., Bays & Husain, 2008; Wilken & Ma, 2004; however, see Rouder et al., 2008, and Zhang & Luck, 2008, for evidence of discrete fixed-resolution representations).

Performance across groups. Observers in the uniform condition remembered 2.7 colors on average throughout the experiment (see Figure 2). This is consistent with previous results on the capacity of visual working memory for colors (e.g., Vogel & Awh, 2008, in which the K values varied from less than 1 to more than 6 across 170 individuals; $M = 2.9$, $SD = 1$).

Critically, we found that observers in the patterned condition could successfully remember $K = 5.4$ colors after learning the regularities in the displays (Block 9). This memory capacity is significantly higher than the $K = 3.0$ colors they were able to remember when the displays were changed to be uniformly distributed in Block 10 (see Figure 2), $t(9) = 4.90$, $p = .0009$, two-tailed; note that this is a within-subject test, and so the between-subjects error bars in Figure 2 underestimate the reliability of this effect. In addition, capacity for colors increased significantly across the first nine blocks of the experiment: $F(8, 72) = 12.28$, $p < .0001$, one-way repeated measures analysis of variance. There was also a significant interaction between color capacity in

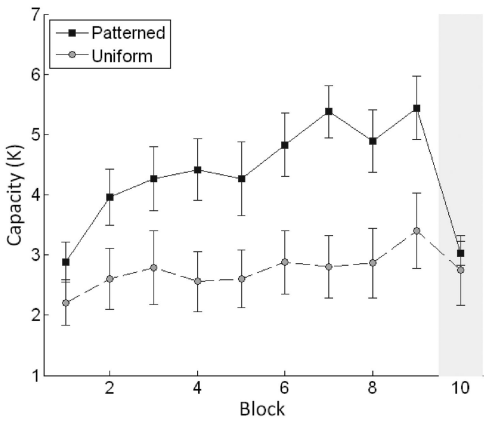

Figure 2. Results of Experiment 1: Memory performance is plotted as a function of experimental block. In the patterned condition some colors appeared on the same object more frequently than did others. Observers' performance increased over time as they learned these regularities. In Block 10 (shaded for emphasis) the regularities were removed, so that all of the colors were drawn from a uniform distribution for both groups of observers. Error bars correspond to ±1 *SEM*.

the uniform condition and color capacity in the patterned condition across blocks, with observers in the patterned condition increasing their capacity more over time, $F(8, 144) = 2.85$, $p = .006$.

Seven of 10 observers in the patterned condition reported noticing regular patterns in the display. The magnitude of the decrease in memory performance from Block 9 to 10 was the same for observers who explicitly noticed the regularities ($M = 26\%$) and those who did not ($M = 27\%$), and 9 of 10 observers showed such decreases (mean decrease across all observers = 26%). In addition, 1 of 10 observers in the uniform condition reported noticing regular patterns in the configuration of colors, although no such patterns existed.

Postperceptual inference. One concern is that observers might simply have remembered one color from each pair and then inferred what the other colors were after the display was gone. This would suggest that observers were actually remembering only three or four colors and were using a postperceptual guessing strategy to achieve a higher performance in the memory test. This leads to two predictions. First, when a color from a low-probability pair is tested (20% of the time), observers should guess wrong and thus should show worse performance on these pairs over time. Second, on these trials they should guess wrong in a specific way—that is, they should guess the high-probability color of the item in the adjacent location. For example, if an observer remembers only that the outside color of an object was blue, and the inside color is tested, they should wrongly infer and report the high-probability color that is often paired with blue.

To test these two predictions, we separated out trials where the tested item was from a high-probability pair and those where the tested item was from a low-probability pair. In other words, if blue often appeared inside red, we considered only the ≈20% of trials where blue appeared with another color or in another configuration. On these trials, an explicit inference process would cause observers to report the wrong color. However, we still found that performance improved over blocks (see Figure 3). Capacity (K), the number of colors remembered, was significantly greater in Block 9, when the low-probability pairs were in the context of high-probability pairs, than in Block 10, when all the pairs were of low probability, $t(9) = 4.08$, $p = .003$.

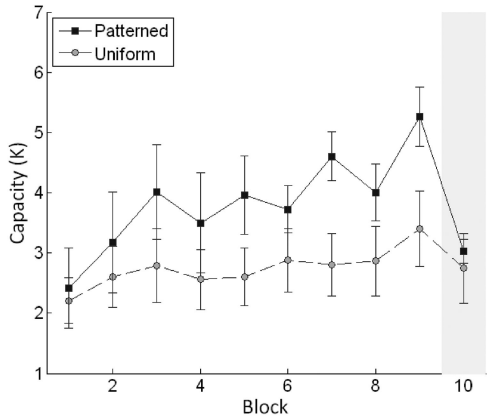

Figure 3. Results of Experiment 1 when considering only cases where the colors appeared in low-probability pairings. The dark squares represent data from observers in the patterned condition for the 20% of trials where a low-probability pair was tested; the gray circles represent the data from observers in the uniform condition. The gray circle in Block 10 corresponds to 100% of uniform trials, because all pairs were of low probability in this block (which is shaded for emphasis). Error bars correspond to ±1 *SEM*.

We next analyzed trials in the first nine blocks where a color from a low-probability pair was tested and observers answered incorrectly (on average there were 35 such trials per observer, for a total of 350 such trials across all 10 observers in the first experiment). If observers do not know what color was present and are explicitly inferring what was on the display using the high-probability pairings, then their responses should more often reflect the high-probability color of the adjacent item. However, on these trials, observers reported the high-probability color of the adjacent item only 9% of the time (where chance is 1/7, or 14%). Further, observers wrongly reported the high-probability color of the tested color only 2% of the time. In fact, the only systematic trend on these low-probability error trials is that observers tended to swap the inner and outer colors much more often than chance: 41% of the time when observers were incorrect, they mistakenly reported the adjacent color. Interestingly, the rate of swaps with the adjacent color was lower in the high-probability pairs: On trials where a high-probability pair was tested, only 27% of error trials were explained by observers incorrectly reporting the adjacent color. This could be taken to suggest that the high-probability pairs tend to be encoded as a single perceptual unit or chunk.

This analysis strongly argues against a postperceptual account of increased memory capacity, where unencoded items are inferred during the testing stage. Not only do observers mostly get trials with the low-probability pairs correct—suggesting they are not performing postperceptual inference—but even on the trials where they do make mistakes, they do not tend to report the associated high-probability colors, as would be predicted by an inference account.

Instead we suggest that observers learned to encode the high-probability pairs using a more efficient representation. For example, suppose a display contains two high-probability pairs and two low-probability pairs. Over time, the high-probability items are encoded more efficiently, leaving more memory resources for the low-probability items. Such an account explains why even colors presented in low-probability pairs showed improved memory performance relative to the uniform group, but only when they were on the same displays as high-probability pairs. In addition, an analysis across trials demonstrated that, on trials with more high-probability pairs in the display, more items were successfully encoded ($K = 3.2$, 3.2, 3.6, 4.0, and 4.7 for 0, 1, 2, 3, and 4 high-probability pairs in the display, averaged across the entire experiment). This increase in capacity as a function of the number of high-probability pairs was significant, $F(4, 36) = 4.25$, $p = .0065$. Furthermore, the only difference between displays containing three or four high-probability pairs was whether the remaining pair's colors were presented in the proper inner–outer configuration. Nevertheless, there was a trend for performance in these two conditions to differ, suggesting that learning may have been specific to the spatial configuration, $t(9) = 1.78$, $p = .11$. Together with the fact that observers did not often flip the inner and outer colors of the high-probability pairs, this suggests that observers may have been encoding the inner and outer colors as a single bound unit or chunk.

Discussion

The present results indicate that, if we consider working memory capacity in terms of the number of colors remembered, observers were able to use the regularities in the displays to increase their capacity past what has been assumed to be a fixed limit of approximately three or four colors. When colors are redundant (i.e., are correlated), then observers can successfully encode more than simply three or four colors. This suggests that the information content of the stimuli is incredibly important to determining how many can be successfully stored (see Alvarez & Cavanagh, 2004, for converging evidence of fewer high-information-load items being stored).

These data can also be interpreted with respect to current psychological constructs for analyzing the capacity of visual working memory (*slots*) and working memory more broadly (*chunks*). In visual working memory, it has been argued that objects with multiple features (e.g., color and orientation) can be stored in a single slot as effectively as objects with only a single feature (Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001). In these models, the unit of memory is thus considered an *object*, a collection of features that are spatiotemporally contiguous (Luck & Vogel, 1997; see Scholl, 2001, for evidence pertaining to the definition of objects in mid-level vision). However, it has been found that memory for objects with two values along a single feature dimension does not show the expected within-object advantage, suggesting that what can be stored in a slot is a single value along each feature dimension, rather than an entire object (e.g., a single object with two colors on it, as in the present experiment, is not represented in a single slot; see for further discussion Olson & Jiang, 2002, Wheeler & Treisman, 2002, and Xu, 2002). This is consistent with the present data from the uniform group, where capacity was approximately three colors rather than three multicolor objects (six colors).

The data from the patterned group represent a challenge to this view. The ability of the patterned group to remember up to six colors represents a capacity of more than a single color per object, suggesting that capacity cannot be fixed to 3–4 objects with a single value along each feature dimension. Instead, the present data can be framed in terms of a slot model only if slots can hold not just one color, but multiple colors from the same object as the objects are learned over time. In this sense, slots of visual working memory become more like chunks in the broader working memory literature (Cowan, 2001). We return to this issue in Experiment 2, when we explore whether these regularities can be used when they are present across objects.

We next performed an information theoretic analysis of the current data to examine if observers have a fixed working memory capacity when measured in bits. We can estimate the amount of redundancy in the displays to test the hypothesis that observers actually have the same amount of resources to allocate in both uniform and patterned conditions. On this account, the difference in memory performance comes from the fact that the patterned displays allow observers to allocate their memory space more effectively. This allows us to make quantitative predictions about working memory capacity given a specific amount of redundancy in the display.

Modeling

Modeling provides a formal framework for theories of compression and allows us to test the hypothesis that there is a limit of visual working memory capacity, not in terms of the number of colors that can be remembered, but in terms of the amount of information required to encode those colors. The modeling has four stages. First, we model how observers might learn the color regularities based on the number of times they saw each pair of colors. The probability of each color pair is estimated with a Bayesian model that accounts for the frequency with which each color pair appeared, plus a prior probability that the colors will be paired uniformly. Second, we assess how these learned statistics translate into representations in bits, using Huffman coding (Huffman, 1952). Huffman coding is a way of using the probabilities of a set of symbols to create a binary code for representing those symbols in a compressed format. This allowed us to estimate the number of bits required to encode each item on the display. Third, we show that the information theoretic model successfully predicts observers' data, suggesting they perform near optimal compression. Finally, we show that a discrete chunking model can also fit the data. Importantly, the best-fitting chunking model is one that closely approximates the information theoretic optimal. MATLAB code implementing the model can be downloaded from the authors' website (http://visionlab.harvard.edu/members/tim/).

Learning the Color Pairs

We used a Dirichlet-multinomial model (Gelman, Carlin, Stern, & Rubin, 2003) to infer the probability distribution that the stimuli were being drawn from, given the color pairs that had been observed. We let d equal the observations of color pairs. Thus, if the trial represented in Figure 1 is the first trial of the experiment, after this trial d = {1 yellow–green, 1 black–white, 1 blue–red, 1 magenta–cyan}. We assume that d is sampled from a multinomial distribution with parameter θ. In other words, we assume that at any point in the experiment, the set of stimuli we have seen so far is a result of repeated rolls of a weighted 64-sided die (one side for each cell in the joint probability matrix; i.e., one for each color pair), where the chance of landing on the ith side of the 64-sided die is given by θ_i. Note that this is a simplification, since the experiment included the additional constraint that no color could appear multiple times in the same display. However, this constraint does not have a major effect on the expected distribution of stimuli once a large number of samples has been obtained and was thus ignored in our formalization.

We set our a priori expectations about θ using a Dirichlet distribution with parameter α. The larger α is, the more strongly the model starts off assuming that the true distribution of the stimuli is a uniform distribution. The α parameter can be approximately interpreted as the number of trials the observers imagine having seen from a uniform distribution before the start of the experiment. Using statistical notation, the model can be written as

$$\theta \sim \text{Dirichlet}(\alpha)$$

$$d \sim \text{Multinomial}(\theta)$$

To fit the model to the data we set a fixed α and assumed that the counts of the pairs that were shown, d, are observed for some

time period of the experiment. Our goal is to compute the posterior distribution $p(\theta|d, \alpha)$. The mean of this posterior distribution is an observer's best guess at the true probability distribution that the stimuli are being drawn from, and the variance in the posterior indicates how certain the observer is about the estimate. The posterior of this model reduces to a Dirichlet posterior where the weight for each color pair is equal to the frequency with which that color pair appears in d, plus the prior on that pair, α_i.

Encoding the Color Pairs

Any finite set of options can be uniquely encoded into a string of bits. For example, if we wished to encode strings consisting of the four letters *A*, *B*, *C*, and *D* into strings of bits, we could do so by assigning a unique two-bit code to each letter and then concatenating the codes. Imagine we had assigned the following codes to the letters: $A = 00$, $B = 01$, $C = 10$, $D = 11$. The string *ACAABAA* could then be written as 00100000010000 (14 bits) and uniquely decoded to retrieve the original string.

Importantly, however, this naïve method of generating a code performs quite badly in the case where some letters are much more likely to appear than others. A better method gives items that occur most frequently the shortest codes, while less frequent items are assigned longer codes. So, for example, if $p(A) = .5$, and $p(B) = .2$, $p(C) = .2$, and $p(D) = .1$, then we can achieve a great deal of compression by representing strings from this language using a different code: $A = 0$, $B = 10$, $C = 110$, $D = 111$. Using this code, the string from above, *ACAABAA*, would be represented as 0110001000 (10 bits), a significant savings even for such a short string (29%). Note that it can still be uniquely decoded, because no item's code is the same as the beginning of a different item's code.

Huffman coding (Huffman, 1952) is a way of using the probabilities of a set of symbols to create a binary code for representing those symbols in a compressed format (as in the example of *A*, *B*, *C*, and *D* above). Here, we used Huffman coding to estimate how much savings observers should show as a result of the fact that the color pairs in our experiment were drawn from a nonuniform distribution. In the Appendix, we demonstrate that the same results also hold for another way of assessing compression using self-information.

We used the probabilities of each color pair, as assessed by the Bayesian model described above, to generate a unique bit string encoding the stimuli on each trial, averaged for each block of the experiment. We supposed that if observers were using some form of compression to take advantage of the redundancies in the display, the length of the code that our compression algorithm generates should be inversely proportional to how many objects observers were able to successfully encode. In other words, if there were many low-frequency color pairs presented (as in Block 10), these items should have longer codes, and observers should be able to successfully remember fewer of them. Alternatively, if there are many high-frequency color pairs presented, the better they should be able to compress the input, and the more colors they will remember.

Information Theory

With these learning and coding models, we can compute a prediction about the memory performance for each subject for each block. In order to assess the fit between the model and the behavioral data, we used the following procedure. For each display in a block, we calculated the number of bits required to encode that display based on the probabilities from the learning model. Next, we correlated the average number of bits per display from the model with the memory performance of the observers. We expected that the fewer bits per display needed, the better observers' memory performance, and thus we expected a negative correlation.

This prediction held quite well, with the maximum fit between this Huffman code model and the human data at $\alpha = 34$, where r, the correlation coefficient between the human and model data, is –.96 (see Figure 4; $p < .0001$). This large negative correlation means that when the model predicts there should be long bit strings necessary to encode the stimuli, human visual working memory stores a low number of items. This is exactly as you would expect if visual working memory took advantage of a compression scheme to eliminate redundant information. In addition, this modeling suggests that if observers encoded the displays completely optimally, they would be able to remember approximately 6.1 colors. By Block 9, observers were remembering 5.4 colors on average, significantly better than with no compression at all, but not quite at the theoretically maximal compression.

The fit between the human data and the model is reasonably good across a broad range of values for the prior probability of a uniform distribution (see Figure 5). The fit is not as high where the prior is very low, since with no prior there is no learning curve—the model immediately decides that whatever stimuli it has seen are completely representative of the distribution (as a non-Bayesian model would do). The fit is also poor where the prior is very high, because the model never learns anything about the distribution of the stimuli, instead generating codes the entire time as though the distribution was uniform. However, across much of the middle range, the model provides a reasonable approximation to human performance.

Importantly, this model allows us to examine if there is a fixed information limit on memory capacity. The Huffman codes provide a measure of the average number of bits per object, and the memory performance gives a measure in number of colors remembered. Thus, if we multiply the average bits per item specified by

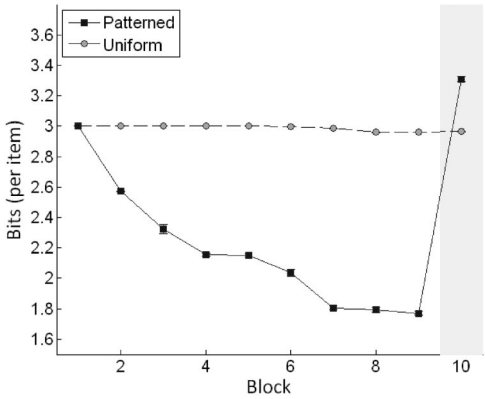

Figure 4. The average length of the Huffman code for a single color, by block. Block 10 is shaded to emphasize that in this block the regularities were removed, so all colors were drawn from a uniform distribution for both groups of observers.

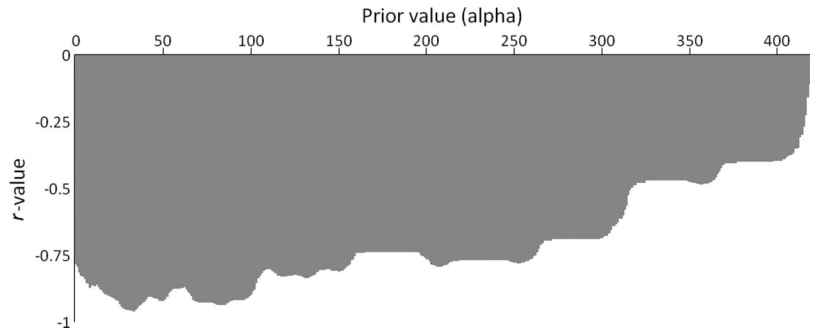

Figure 5. The correlation between the information theoretic model and the human behavioral data as a function of the value of the prior, α.

the Huffman code times the number of items remembered, we get an estimate of the number of bits of information a given set of observers recalled in a given block (see Figure 6).

Notice first that both groups of observers in the uniform condition and the patterned condition show roughly the same total capacity in bits, despite the overall difference in the number of items remembered between the groups. Second, the total bit estimate remains remarkably constant between Block 9 and Block 10 in the patterned group, even though the memory performance measured in number of colors showed a significant cost when the statistical regularities were removed. Thus, while the patterned group was able to remember more colors throughout the experiment, this increase was completely explained in the model by the fact that the items to be remembered were more redundant and presumably took less space in memory.

Chunking Model

The information theoretic modeling gives a way of formally specifying how compressible a set of input is, given the accumulated statistics about the previous input. Huffman coding and self-information are ways to formalize this and are thus a form of rational analysis or computational theory, specifying the optimal solution to the computational problem facing the observer (Anderson, 1990; Marr, 1982). Interestingly, we find that observers closely approximate this optimum. However, Huffman coding and self-information are not meant as serious candidates for the psychological mechanism people use for implementing such compression. Indeed, it is a different level of analysis to understand what psychological algorithms and representations are actually responsible for allowing more items to be encoded in memory when those items are redundant. For instance, is the nature of the compression graded over time, or all-or-none?

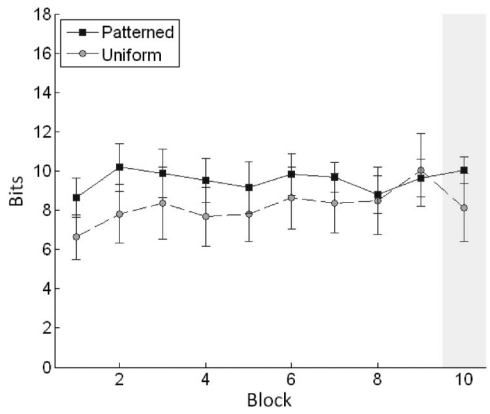

Figure 6. The size of memory estimated in bits, rather than number of colors (using the Huffman coding model). Block 10 is shaded to emphasize that in this block the regularities were removed, so all colors were drawn from a uniform distribution for both groups of observers. Error bars represent ±1 *SEM*.

In the information theoretic models (Huffman coding, self-information), the "cost" of encoding each color pair is equal to the log of the chance of seeing that pair relative to the chance of seeing any other pair. This is the optimal cost for encoding items if they appear with a given probability, and provides for graded compression of a color pair as the items' probability of co-occurrence increases. However, the actual psychological mechanism that people use to remember more items could be either graded as in the rational analysis, or could function as a discrete approximation to this optimum by sometimes encoding highly associated items into a single representation. Chunking models are one way of approaching this kind of discrete approximation (e.g., Cowan et al., 2004). They show increased memory capacity for highly associated items but convert the compressibility to a discrete form: Either a single chunk is encoded or the two colors are separately encoded into two chunks. This distinction between graded compression and all-or-none compression is important because it predicts what is actually encoded by an observer in a single trial. The current results do not address this distinction directly, however, because we do not examine the representational format of the color pairs on each trial. However, a broad literature expresses a preference for viewing compression in working memory as based on discrete chunking (e.g., Chase & Simon, 1973; Cowan, 2005; Miller, 1956; however, see Alvarez & Cavanagh, 2004, Bays & Husain, 2008, and Wilken & Ma, 2004, for support for a graded view). Thus, we sought to examine whether our data could be accurately modeled using this kind of approximation to the information theoretic analysis presented above.

To implement a simple chunking model, one needs to determine a threshold at which associated items become a chunk. The most naïve chunking model is one in which observers reach some fixed threshold of learning that a pair of colors co-occur and treat them

as a chunk thereafter (perhaps after this new chunk enters long-term memory). However, this simple model provides a poor fit to the current data. In such a model, each subject will have a strong step-like function in his or her graph, and the graded form of the group data will arise from averaging across observers. However, in the present data, single observers showed a graded increase in performance by block, suggesting this kind of model does not accurately represent the data.

A more sophisticated class of chunking models have a probabilistic threshold, allowing for single observers to treat each color pair as one chunk more often if they strongly believe it is a chunk, and less often if they are unsure if it is a chunk. In the case where the chance of chunking in such a model is logarithmically proportional to the association between the items, this chunking model is exactly equivalent to a thresholded version of the information theoretic compression model and therefore makes the same predictions across large numbers of trials. However, a chunking model could also assume that the possibility of chunking is linearly proportional to the association between the items, $p_{\text{chunk}(i,\,j)} = \beta \times \theta_{i,\,j}$, in which case it would be possible that the chunking model's fit would differ significantly from that of the more ideal compression algorithms. We did not find this to be the case for the current experiment.

The graph from the best fit linear chunking model is shown in Figure 7. The best fit constant of proportionality was 15, which provided a fit to the data of $r = -.90$ (e.g., for each pair, the chance of being chunked on any given trial was equal to $15 \times \theta_{i,\,j}$, such that once the probability of seeing a given color pair was greater than 1/15th, that color pair was always encoded as a single chunk). Interestingly, this constant of proportionality—because it causes such a steep increase in the chance of chunking even at very low associations and plateaus at a 100% chance of chunking by the time the association reaches 1/15, or 0.067—approximates the shape of a logarithmic curve. The correlation between the probability of chunking under this linear model and the optimal cost function derived via information theory (using self-information) is therefore approximately $r = -.73$. This model thus provides an excellent approximation to the ideal compression algorithm as well.

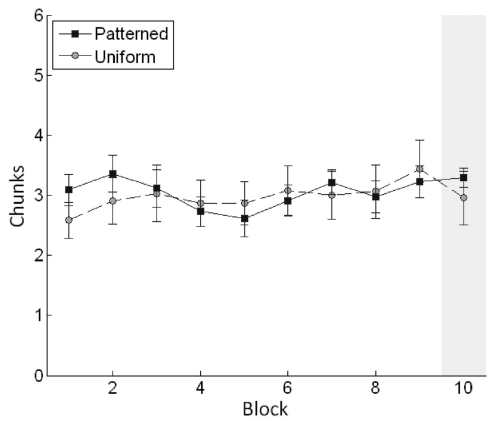

Figure 7. The size of memory (in chunks) for Experiment 1 estimated using the probabilistic linear chunking model. Block 10 is shaded to emphasize that in this block the regularities were removed, so all colors were drawn from a uniform distribution for both groups of observers. Error bars represent ±1 *SEM*.

Thus, we find that the best chunk-model matches the data well and generates a flat estimate of the number of chunks needed across the entire experiment. Importantly, however, the expected probability of chunking in this model closely matches the optimal information-theoretic cost function (higher cost = lower probability of chunking). This is to be expected because the information theoretic model predicted 92% of the variance in the behavioral data. This suggests that chunking can be usefully thought of as a discrete approximation to an ideal compression algorithm and therefore can be thought of as a possible psychological implementation of compression.

It is important to note that, despite the assumptions we make in this modeling section, it unlikely that the degree of association between items determines when they form chunks in long-term memory. Instead, it may be that human chunk learning depends on how useful a particular chunk would be in describing the world while avoiding "suspicious coincidences" (see, e.g., Orbán et al., 2008, who provided an elegant Bayesian analysis of this problem). Our analysis of chunking here is meant only as a proof of the concept that chunking models in general implement a form of compression that approximates the true information theoretic optimum.

Discussion

The modeling work we present illustrates two main conclusions: First, compression of redundancies must be taken into account when quantifying human visual working memory capacity; second, this compression can be modeled either in a graded fashion, or in an all-or-none fashion (*probabilistic chunking*), which closely approximates ideal compression algorithms.

The fact that the estimate of the amount of information observers are able to store is constant across the entire experiment, whereas the estimate in terms of number of colors varies a great deal, suggests that compression of redundancies must be taken into account when quantifying human visual working memory capacity. In addition, it is important to note that fitting our information theoretic model by minimizing the correlation to the data is not guaranteed to provide a fit that results in a flat line in terms of the total information remembered. In fact, in most instances a negative correlation will not lead to a flat estimate across the experiment, since a flat line additionally depends on a proportional amount of decrease at each step. The information theoretic modeling results provide significant evidence that the capacity of working memory is a fixed amount of information. Because the chunking model is the discrete version of an optimal compression scheme, this model leads to a fixed capacity measured in discrete units (chunks) just as the information theoretic model let to a fixed capacity measured in continuous information (bits).

Although our model suggests a working memory capacity of 10 bits, this number should not be taken as indicative of limits on human performance. The exact number—10 bits—depends critically on assumptions about how the colors are encoded (three bits per color in our model, given the eight possible color choices). It is important, however, that the estimate of memory size be constant across the experiment and across conditions not depend on our choice of encoding scheme but only on the redundancy inher-

ent in the associations between colors. If observers actually required 100 bits to encode each color then our estimate of capacity in bits would change to about 300 bits—but the estimate would still remain consistent across the experiment, because each color still provides the same proportional amount of information about each other color. Thus, it is safe to conclude that our results are compatible with a fixed amount of information limiting memory performance, but it is difficult to quantify the exact number of bits without specifying the true coding model (see the General Discussion for further discussion of the problem of specifying an encoding scheme).

Experiment 2: Regularities Between Objects

The aim of Experiment 2 was to examine if compression can affect encoding across objects as well as within objects. This experiment was very similar to Experiment 1, with the only difference being how the colors were presented on the display. In Experiment 2, colors were presented side-by-side as separate objects, in close proximity but not spatially contiguous.

While there are many possible definitions of *object*, we use the term to refer specifically to a spatiotemporally contiguous collection of visual features in mid-level vision (Scholl, 2001; Spelke, 1990). This definition is motivated by both neuropsychological and behavioral evidence (Behrmann & Tipper, 1994; Egly, Driver, & Rafal, 1994; Mattingley, Davis, & Driver, 1997; Scholl, Pylyshyn, & Feldman, 2001; Watson & Kramer, 1999). For example, simply connecting two circles with a line to form a dumbbell can induce *object-based neglect*, in which the left half of the dumbbell is neglected regardless of the half of the visual field in which it is presented (Behrmann & Tipper, 1994). If these two circles are not connected, neglect does not operate in an object-based manner. Thus, based on this definition of what counts as an object, the displays of Experiment 1 contained four objects whereas the displays of Experiment 2 contained eight objects (see Figure 8).

In the present experiment, we examined whether or not working memory capacity can take advantage of the statistics between objects. If visual working memory capacity limits are *object-based*, that is, if capacity is constrained by mid-level visual objects, then observers will not be able to take advantage of regularities across objects. However, if multiple visual objects can be stored together (akin to "chunks" of letters, as in *FBI–CIA*), then people will be able to remember more colors from the display as they learn the statistics of the input.

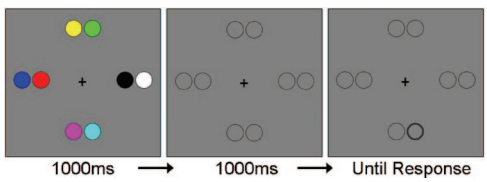

Figure 8. A sample trial from Experiment 2. Eight colors were presented and then disappeared, and after 1 s, one location was cued. Observers had to indicate what color was at the cued location.

Method

Observers. Twenty naïve observers were recruited from the MIT participant pool (age range 18–35) and received $10 for their participation. All gave informed consent.

Procedure. Observers were presented with displays consisting of eight objects arranged in four pairs around the fixation point (see a sample display in Figure 8). Each object was made up of only one colored circle. Here the two associated colors appeared on separate objects, but we provided a grouping cue in order to not significantly increase the difficulty of the learning problem. All other aspects of the stimuli and procedure were identical to those of Experiment 1.

Results

Performance across groups. Observers in the uniform condition remembered $K = 3.4$ colors on average throughout the experiment (see Figure 9), consistent with previous results on the capacity of visual working memory for colors (Vogel & Awh, 2008) and the results of Experiment 1.

We found that observers in the patterned condition could successfully remember $K = 5.4$ colors after learning the regularities in the displays (Block 9). This memory capacity is significantly higher than the $K = 3.3$ colors they were able to remember when the displays were changed to be uniformly distributed in Block 10 (see Figure 9), $t(9) = 9.72$, $p < .0001$. In addition, capacity increased significantly across the first nine blocks of the experiment, $F(8, 72) = 7.68$, $p < .0001$. There was a significant interaction across blocks between capacity in the uniform condition and capacity in the patterned condition, with observers in the patterned condition remembering more colors over time, $F(8, 144) = 2.27$, $p = .025$.

Eight of 10 observers reported noticing regular patterns in the display. The magnitude of the decrease in memory performance from Block 9 to 10 was the same for observers who explicitly noticed the regularities ($M = 22\%$) and those who did not ($M = 23\%$), and 9 of 10 observers showed such decreases (mean decrease across all observers = 23%). Three of 10 observers in the uniform condition reported noticing regular patterns in the configuration of colors, although no such patterns were present.

We once again separated out trials where the tested item was from a high-probability pair from those where the tested item was from a low-probability pair. When we examined only the low-probability trials, we still found that capacity in Block 9 was significantly higher than in Block 10, with 4.9 colors remembered in Block 9 and 3.4 colors remembered in Block 10, $t(9) = 4.84$, $p = .0009$. Thus, as with Experiment 1, we did not find evidence that people were remembering more items from the display by using postperceptual inference.

Performance across experiments. We compared the first nine blocks in the patterned condition of this experiment to the first nine blocks in the patterned condition of Experiment 1. There were no main effects or interactions (all Fs < 1). Furthermore, we compared the drop in performance between Block 9 and Block 10 across the two experiments. The size of the drop was not significantly different, $t(9) = 0.58$, $p = .58$, suggesting that learning was of a comparable magnitude in both experiments.

Verbal interference. One potential concern is that observers could have used some verbal memory capacity to augment their

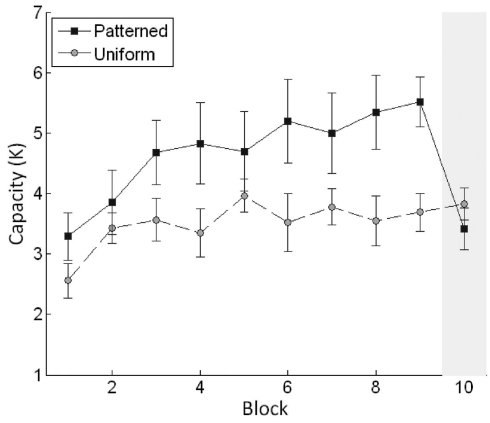

Figure 9. Results of Experiment 2. In the patterned condition some colors appeared on adjacent objects more frequently than did others. Observers' performance increased over time as they learned these regularities. In Block 10 (shaded for emphasis) the regularities were removed, so that all of the colors were drawn from a uniform distribution for both groups of observers. Error bars correspond to ±1 *SEM*.

visual working memory in either the current experiment or Experiment 1. Many past studies have found that estimates of visual working memory capacity are similar with and without verbal interference (e.g., Luck & Vogel, 1997; Vogel et al., 2001). However, because of the added element of learning regularities in our experiments, we decided to test the effects of verbal interference on our paradigm. Because of the similarities between Experiment 1 and Experiment 2, we ran a control experiment using only the paradigm of Experiment 2.

We conducted this control experiment with 7 observers using an identical paradigm to Experiment 2's patterned condition, but with the addition of a verbal interference task (remembering four consonants throughout the duration of the trial, with a new set of four consonants every 10 trials). Observers successfully performed both the verbal interference task and the visual working memory task, with a capacity of 4.5 colors in Block 9 but only 3.2 colors in Block 10, $t(6) = 2.10$, $p = .08$. Capacity in Block 9 under verbal interference was not significantly different than that obtained in Block 9 of Experiment 2, $t(9) = 1.07$, $p = .31$. These data show that observers are still capable of learning the regularities to remember more colors when subject to verbal interference in a challenging dual-task setting.

Modeling. We once again modeled these results to see if they were compatible with a model in which compression is explained via information theory. The maximum fit between the Huffman code model and the human data occurred at $\alpha = 31$ where r, the correlation coefficient between the human and model data, is –.96 ($p < .0001$). This large negative correlation means that when the model predicts there should be long bit strings necessary to encode the stimuli, observers' memory capacity in terms of the number of colors remembered is low. This is exactly what one would expect if visual working memory had a fixed size in bits and took advantage of a compression scheme to eliminate redundant information.

In addition, this model allows us to once again examine if there is a fixed-bit limit on memory capacity. The Huffman codes gives a measure of average bits per object, and the memory performance gives a measure in number of objects remembered. As in Experiment 1, multiplying the average size of the Huffman code times the number of items remembered gives us an estimate of the number of bits of information a given set of observers recalled in a given block (see Figure 10). Notice that once again both the groups of observers in the uniform condition and the patterned condition showed the same total capacity in bits, despite the overall difference in the number of items remembered between the groups. Second, the total bit estimate remained remarkably constant between Block 9 and Block 10 in the patterned group, even though the memory performance measured in number of items showed a significant cost when the statistical regularities were removed.

One interesting prediction of the model is that the patterned group should actually be worse at Block 10 than the uniform group, since the patterned group now has a set of statistics in mind that are no longer optimal for the displays. Indeed, the pattern in the behavioral data trends this way, but the difference between both groups in Block 10 was not significant (see Figure 9), $t(9) = 0.64$, $p = .47$. One possible explanation for why performance for the patterned group does not fall completely below the uniform group is that observers notice that their model has become inappropriate after several trials in Block 10 and begin using a relatively local estimate of the probability distribution (e.g., across the last few trials) or revert to a uniform model. This suggests a good deal of flexibility in the model observers use to encode the display.

In addition, we modeled these results using a probabilistic chunking model where the chance of chunking was linearly proportional to the probability of the color pair. Using the same parameters as in Experiment 1, this model too provided a good fit to the data ($r = .94$; see Figure 11), and it produced an almost flat estimate of the number of chunks over time in both groups.

Discussion

Observers in the patterned group were able to successfully take advantage of the redundancy in the displays, as their capacity

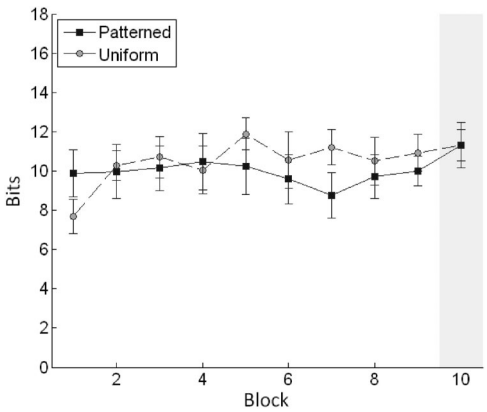

Figure 10. The size of memory estimated in bits, rather than number of objects (using the Huffman coding model). Block 10 is shaded to emphasize that in this block the regularities were removed, so all colors were drawn from a uniform distribution for both groups of observers. Error bars represent ±1 *SEM*.

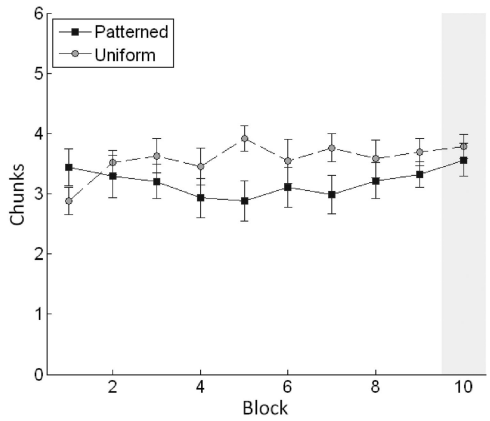

Figure 11. The size of memory (in chunks) for Experiment 2 estimated using the probabilistic linear chunking model. Block 10 is shaded to emphasize that in this block the regularities were removed, so all colors were drawn from a uniform distribution for both groups of observers. Error bars represent ±1 *SEM*.

increased significantly over time. These data, as well as the estimated capacity in bits from the modeling, reveal strikingly similar patterns between Experiment 1 and Experiment 2. This suggests that observers were equally able to take advantage of the redundancy in the displays when the redundancies were present between adjacent mid-level visual objects rather than within such objects.

This experiment has some implications for the classic slot model of visual working memory (Luck & Vogel, 1997; Zhang & Luck, 2008). Specifically, a strict interpretation that one slot can hold only one mid-level visual object from the display does not account for the present data. The patterned group was able to remember more objects over time, so the capacity of working memory cannot be a fixed number of mid-level visual objects. However, if multiple objects can fit into one slot, then the present data can be accounted for. Indeed, this suggests that slots in visual working memory should be viewed similarly to chunks in verbal working memory (Cowan, 2001). Thus, in the present experiment *visual chunks* could be formed that consist of pairs of colored objects (see also Orbán et al., 2008, for evidence of statistical learning of chunks of multiple objects). Of course, another possibility is that working memory capacity should be thought of as graded, rather than based on chunks or slots at all (e.g., Alvarez & Cavanagh, 2004; Bays & Husain, 2008, 2009). This would make it more closely approximate the information theoretic ideal and would account for the present data directly.

An important factor in the present experiment is that we provided a grouping cue for observers by putting the two colors that will covary in closer proximity to each other than to the other colors. We expect that learning would still be possible even if the items were not specifically grouped, as others have demonstrated that statistical learning can operate across objects, even in cases when the display is unparsed, and that such learning results in the formation of visual chunks (Fiser & Aslin, 2001, 2005; Orbán et al., 2008; see also Baker, Olson, & Behrmann, 2004). However, our aim in this experiment was not to create a difficult learning situation; rather, our aim was to demonstrate that visual working memory can take advantage of these learned statistics to remember more of the display even when the statistics relate the co-occurrence of different objects, as in the work of Fiser and Aslin (2001). It is an avenue of future research to explore what kinds of statistics can be gleaned from the input and where the statistical learning mechanisms fail. It will also be important to discover if situations exist in which observers can successfully learn statistical regularities but are unable to take advantage of those regularities to efficiently store items in memory.

Finally, as in Experiment 1, the modeling showed that even though people are remembering more items, their working memory capacity is actually constant when quantified by the amount of information remembered (or the number of chunks stored). In general, this suggests that the tools of information theory combined with Bayesian learning models enable us to take into account the compressibility of the input information and provide clear, testable predictions for how many items observers can remember. This suggests that compression must be central to our understanding of visual working memory capacity.

General Discussion

We presented two experiments contrasting memory capacity for displays where colors were presented in random pairs versus in recurring patterns. In the first experiment, the colors that formed a pattern were presented as part of the same object. In the second experiment, the colors that formed a pattern were presented on two different but spatially adjacent objects. For both experiments we found that observers were successfully able to remember more colors from the displays in which regularities were present. The data indicate that this is not due to postperceptual inference but reflects an efficient encoding. We proposed a quantitative model of how learning the statistics of the input would allow observers to form more efficient representations of the displays and used a compression algorithm (Huffman coding) to demonstrate that observers' performance approaches what would be optimal if their memory had a fixed capacity in bits. In addition, we illustrated that a discrete model of chunking also fits our data. The degree of compression possible from the display was highly correlated with behavior, suggesting that people optimally take advantage of statistical regularities to remember more information in working memory.

We thus show that information theory can accurately describe observers' working memory capacity for simple colors that are associated with each other, since such a capacity depends on compression. By using a statistical learning paradigm, we control the statistics of the input, allowing us to measure the possible compression in this simple task. Since in the world almost all items we wish to remember are associated with other objects in the environment (Bar, 2004), using information theory to quantify the limits of working memory capacity is likely of more utility for natural viewing conditions than measuring the number of independent items that people can remember.

Resolution Versus Number

One interesting factor to consider is whether the increase in percent correct we observed during training in the patterned group could be due to an increase in the resolution at which observers store the colors, rather than an increase in the number of colors

remembered per se (similar to the claims of Awh, Barton, & Vogel, 2007).

We believe several factors speak against such an account. In particular, if a fixed number of items are remembered and only the resolution of storage is increasing, then the fixed number of items remembered would have to be at least six (the number of colors remembered by the patterned group in the ninth block of trials). This seems very unlikely, given that previous estimates of the fixed number are on the order of three or four (Cowan, 2001; Luck & Vogel, 1997), even for studies that explicitly address this issue of the resolution with which items are stored (Awh, Barton, & Vogel, 2007; Zhang & Luck, 2008). In addition, while Awh et al. (2007) provided some evidence that for complex objects there may be a difference in resolution between different object classes, both Rouder et al. (2008) and Zhang and Luck (2008) have recently argued for discrete fixed-resolution representations in the domain of color (although see Bays & Husain, 2009, for a critique of this work). These articles provide evidence that for simple features like color, the colors are either remembered or not remembered, rather than varying in resolution. Finally, it is not clear why the covariance introduced in the colors would affect the resolution of a single color, and what the proper relationship would be between the resolution and the association strength. For these reasons we believe it is unlikely that the current data reflect changes in the resolution of the items rather than the quantity of items stored.

Relationship Between Slots and Objects

Much of the work on visual working memory has emphasized the privileged role of objects in memory capacity. For example, there is often an advantage to representing two features from the same object as opposed to two of the same features from two different objects (Luck & Vogel, 1997; Xu, 2002). In fact, visual working memory is often conceptualized as containing 3–4 *slots*, in which one object, and all its features, can be stored in one slot (e.g., Luck & Vogel, 1997; Zhang & Luck, 2008) with some degree of fidelity. In this literature, *objects* typically are assumed be the units of mid-level vision, specifically a spatiotemporally contiguous collection of features.

Our data suggest that at least the simplest version of an object-based capacity limit, in which one object in the world is stored in one slot in the mind, is not sufficient. If observers have a fixed working memory capacity of 3–4 objects on average, then both the uniform and patterned groups should show the same memory performance in Experiment 2. Similarly, if observers can remember at most 3–4 values along a single feature dimension (like color), then both the uniform and patterned groups should show the same memory performance in Experiment 1. However, in both Experiment 1 and Experiment 2, the patterned groups were able to remember almost twice as many objects by the end of the experiment. Thus, if there are slots in the mind, they must be able to hold more than one mid-level visual object, much like how chunks can contain multiple digits or words in the verbal working memory literature. The critical point here is that visual working memory should not be said to hold only 3–4 mid-level visual objects or 3–4 values along a single feature dimension, but instead needs to allow for visual chunking. Alternatively, visual working memory capacity may be characterized in a more graded fashion rather than using slots or chunks as a unit of measure (Alvarez & Cavanagh, 2004; Bays & Husain, 2008; Wilken & Ma, 2004).

Chunking

The current behavioral data cannot directly address whether the proper way to characterize the capacity of the system is in terms of a continuous measure of information or in terms of a model in which items are stored discretely in chunks or slots (Cowan, 2001; Luck & Vogel, 1997; Miller, 1956; Simon, 1974). Our information theoretic analysis puts a theoretical limit on how compressible this information should be to a learner. However, exactly *how* this compression is implemented in psychological constructs remains an open question. One possibility is that associated items become more and more compressible over time (i.e., start to take up less space in memory). Another possibility is that pairs of items take up either two chunks or one, depending on a probabilistic chunking threshold. Importantly, a continuous model of compression can be closely approximated by a discrete model, as long as the threshold for forming a chunk is related to the cost of the items in information theoretic terms. In fact, any chunking model that will account for our data will need to form chunks in a way that is compatible with our information theoretic analysis. In this sense, information theory allows us to constrain chunking models significantly and has the potential to break us out of the circular dilemma of determining what ought to count as a single chunk (Simon, 1974).

Coding Model

It is important to emphasize that compression must be defined with respect to a coding model. Naïve information theoretic models (e.g., Kleinberg & Kaufman, 1971), which simply assume that all items are coded with respect to the possible choices for a particular task, are not adequate ways of characterizing the capacity of the memory system. For example, using such a coding scheme it takes 1 bit to represent a binary digit and 3.3 bits to represent a decimal digit. However, as described clearly in Miller (1956), if observers can remember a fixed amount of information, then on the basis of the number of decimal digits they can remember, they ought to be able to remember many more binary digits.

Some hints at what the psychological coding model might be like in this case come from evidence that shows observers tend to store digits phonetically (Baddeley, 1986). Thus, perhaps a proper information theoretic model would encode both binary and decimal digits with respect to the entire set of phonemes. Of course, even the phonetic coding scheme is not sufficient for capturing how much information is in a string, as the conceptual content matters a great deal. For example, people can remember many more words if the words make a coherent sentence than if they are randomly drawn from the lexicon (Simon, 1974). This is also true in memory for visual information: Sparse cartoon drawings are remembered better when given a meaningful interpretation (Bower, Karlin, & Dueck, 1975; see also Wiseman & Neisser, 1974). Presumably abstract line drawings have much longer *coding strings* than when those same line drawings can be encoded with respect to existing knowledge.

In the current experiment, we specifically avoided having to discover and specify the true coding model. By exploring compression within the domain of associations between elements (col-

ors in the current study), we need to specify only the information present in their covariance. Specifying how long the bit string is for a display of eight colored circles would require a complete model of the visual system and how it encodes the dimensions of colored circles. Because the true coding model is likely based in part on the natural statistics of the visual input, and given the frequency of a gray screen with eight colored circles on it in our everyday visual experience, the bit strings for such a display are likely quite long. Instead we used a paradigm that builds associations between elements over time, allowing us to control the coding model that could be learned from the regularities in the displays. This method avoids many of the pitfalls traditionally associated with information theoretic models (e.g., those examined by Miller, 1956). Importantly, our results demonstrate that in this simplified world of associated colors, visual working memory is sensitive to the incoming statistics of the input. This approach opens the door for future work to apply information theoretic models to human cognition without first solving for the perceptual coding schemes used by the brain.

Moving beyond simple pairwise associations between colors, for more complex stimuli and in more real-world situations, observers can bring to bear rich conceptual structures in long-term memory and thus achieve much greater memory performance (e.g., Ericsson, Chase, & Faloon, 1980). These conceptual structures act as internal models of the world and therefore provide likelihoods of different items appearing in the world together. For example, observers know that computer monitors tend to appear on desks, that verbs tend to follow subjects, and that kitchens tend to be near dining rooms. Importantly, our information theoretic framework can, at least in principle, scale up to these more difficult problems, since it is embedded in a broader Bayesian framework which can make use of structured knowledge representations (Kemp & Tenenbaum, 2008; Tenenbaum, Griffiths, & Kemp, 2006).

Relation to Learning and Long-Term Memory

Compressibility and chunking are rarely formalized outside the literature on expertise (e.g., chunking models; Gobet et al., 2001), and thus the relation between visual working memory capacity and the learning of relations between items has received little attention in the literature (although see Cowan et al., 2004, for an analysis in the verbal domain). However, there are several interesting data points about the role of learned knowledge in working memory capacity more broadly; for example, adults have a greater working memory capacity than do children (Simon, 1974). In addition, there is a large literature on expertise and chunking (Chase & Simon, 1973; Gobet et al., 2001), where there is significant appreciation of the fact that long-term knowledge is a significant factor in working memory capacity (see also Curby, Glazek, & Gauthier, 2009; Olsson & Poom, 2005; Scolari, Vogel, & Awh, 2008).

By relating working memory capacity and chunking strongly to information theory, our results suggest a broad purpose for a particular kind of long-term knowledge acquisition: statistical learning. In particular, a great deal of recent work has focused on a set of statistical learning mechanisms that are capable of extracting many different regularities with only minutes of exposure and appear to be relatively ubiquitous, occurring in the auditory, tactile, and visual domains and in infants, adults, and monkeys (Brady & Oliva, 2008; Conway & Christiansen, 2005; Fiser & Aslin, 2002; Hauser, Newport, & Aslin, 2001; Kirkham, Slemmer, & Johnson, 2002; Saffran, Aslin, & Newport, 1996; Turk-Browne, Jungé, & Scholl, 2005). The present results suggest that one of the primary reasons for being sensitive to such regularities might be that it allows us to remember more in working memory by eliminating redundancy in our representations. They also emphasize how quickly such long-term memories can be built and can start to influence capacity measures—observers in the present studies demonstrated significant improvements in working memory capacity by Block 2, only a few minutes into the experiment. In addition, it is important to keep in mind that statistical learning mechanisms need not be limited to learning simple associations between items. Both the learning process and the representations that are learned can be, and likely are, much richer than simple associations (see, for example, Orbán et al., 2008, and Frank, Goldwater, Mansinghka, Griffiths, & Tenenbaum, 2007).

Conclusion

The information we can hold in working memory is surprisingly limited. However, in the real world there are strong associations and regularities in the input, and our brain is tuned to these regularities in both perception and memory (Anderson, 1990; Field, 1987). In an information theoretic sense, such regularities introduce redundancies that make the input more compressible.

We have shown that observers can take advantage of these redundancies, enabling them to remember more colors in visual working memory. In addition, while we showed this using simple associations between colors, the Bayesian modeling framework we used has the potential to scale up to learning over more complex representations. Thus, we believe that the tools of probabilistic modeling and information theory can help in understanding how observers form long-term memory representations and use them in working memory. More generally, our data support the view that perceptual encoding rapidly takes advantage of redundancy to form efficient codes.

References

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

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

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

Anderson, J. R., & Schooler, L. J. (2000). The adaptive nature of memory. In E. Tulving (Ed.), *The Oxford handbook of memory* (pp. 557–570). New York: Oxford University Press.

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

Baddeley, A. D. (1986). *Working memory.* New York: Oxford University Press.

Baker, C., Olson, C., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science, 15,* 460–466.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5,* 617–629.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation, 1,* 295–311.

Bays, P. M., & Husain, M. (2008, August 8). Dynamic shifts of limited working memory resources in human vision. *Science, 321,* 851–854.

Bays, P. M., & Husain, M. (2009, February 13). Response to comment on "Dynamic Shifts of Limited Working Memory Resources in Human Vision." *Science, 323,* 877.

Behrmann, M., & Tipper, S. P. (1994). Object-based visual attention: Evidence from unilateral neglect. In C. Umilta & M. Moscovitch (Eds.), *Attention and Performance XV: Conscious and Nonconscious Information Processing* (pp. 351–376). Cambridge, MA: MIT Press.

Bower, G. H., Karlin, M. B., & Dueck, A. (1975). Comprehension and memory for pictures. *Memory & Cognition, 3,* 216–220.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. *Psychological Science, 19,* 678–685.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436.

Broadbent, D. E. (1958). *Perception and communication.* London: Pergamon Press.

Burton, G. J., & Moorehead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics, 26,* 157–170.

Chandler, D. M., & Field, D. J. (2007). Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America, 24*(A), 922–941.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science, 3,* 57–65.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 24–39.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York: Wiley.

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

Cowan, N. (2005). *Working memory capacity.* Hove, United Kingdom: Psychology Press.

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science, 15,* 634–640.

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 94–107.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General, 123,* 161–177.

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980, June 6). Acquisition of a memory skill. *Science, 208,* 1181–1182.

Field, D. J. (1987). Relations between the statistics of natural images and response properties of cortical cells. *Journal of the Optical Society of America, 4*(A), 2379–2394.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scene. *Psychological Science, 12,* 499–504.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 458–467.

Fiser, J., & Aslin, R. N. (2005). Encoding multielement scenes: Statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General, 134,* 521–537.

Frank, M. C., Goldwater, S., Mansinghka, V., Griffiths, T., & Tenenbaum, J. (2007). Modeling human performance in statistical word segmentation. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 281–286). Austin, TX: Cognitive Science Society.

Frazor, R. A., & Geisler, W. S. (2006). Local luminance and contrast in natural images. *Vision Research, 46,* 1585–1598.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Science, 5,* 236–243.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition, 78,* B53–B64.

Huffman, D. A. (1952). A method for construction of minimum-redundancy codes. *Proceedings of the IRE, 40,* 1098–1101.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, USA, 105,* 10687–10692.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83,* 35–42.

Kleinberg, J., & Kaufman, H. (1971). Constancy in short-term memory: Bits and chunks. *Journal of Experimental Psychology, 90,* 326–333.

Luck, S. J., & Vogel, E. K. (1997, November 20). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281.

Magnussen, S., Greenlee, M. W., & Thomas, J. P. (1996). Parallel processing in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 202–212.

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

Mattingley, J. B., Davis, G., & Driver, J. (1997, January 31). Preattentive filling-in of visual surfaces in parietal extinction. *Science, 275,* 671–674.

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

Neisser, U. (1967). *Cognitive psychology.* Englewood Cliffs, NJ: Prentice-Hall.

Olshausen, B. A., & Field, D. J. (1996, June 13). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607–609.

Olson, I. R., & Jiang, Y. (2002). Is visual short-term memory object based? Rejection of the "strong-object" hypothesis. *Perception & Psychophysics, 64,* 1055–1067.

Olson, I. R., & Jiang, Y. (2004). Visual short-term memory is not improved by training. *Memory & Cognition, 32,* 1326–1332.

Olson, I. R., Jiang, Y., & Moore, K. S. (2005). Associative learning improves visual working memory performance. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 889–900.

Olsson, H., & Poom, L. (2005). Visual memory needs categories. *Proceedings of the National Academy of Sciences, USA, 102,* 8776–8780.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences, USA, 105,* 2745–2750.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences, USA, 105,* 5975–5979.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996, December 13). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80,* 1–46.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition, 80,* 159–177.

Scolari, M., Vogel, E. K., & Awh, E. (2008). Perceptual expertise enhances the resolution but not the number of representations in working memory. *Psychonomic Bulletin & Review, 15,* 215–222.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379–423.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). Oxford, United Kingdom: Oxford University Press.

Simon, H. A. (1974, February 8). How big is a chunk? *Science, 183,* 482–488.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14,* 29–56.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74,* 1–29.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science, 10,* 309–318.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134,* 552–564.

Vogel, E., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science, 17,* 171–176.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 92–114.

Watson, S. E., & Kramer, A. F. (1999). Object-based visual selective attention and perceptual organization. *Perception & Psychophysics, 61,* 31–49.

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

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4,* 1120–1135.

Wiseman, S., & Neisser, U. (1974). Perceptual organization as a determinant of visual recognition memory. *The American Journal of Psychology, 87,* 675–681.

Xu, Y. (2002). Limitations in object-based feature encoding in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 458–468.

Zhang, W., & Luck, S. J. (2008, May 8). Discrete fixed-resolution representations in visual working memory. *Nature, 452,* 233–235.

(*Appendix follows*)

Appendix

Model Results Using Self-Information

The self-information at seeing a given item, *i*, expressed in bits, is

$$S = -\log_2(p_i).$$

More bits of information (S) are gained by seeing items that are of low probability (small p_i) than items that are of high probability (large p_i). The number of bits of self-information is the mathematical optimum for how many bits must be required to encode particular stimuli from a given distribution (Shannon, 1948).

In practice, it is difficult or impossible to achieve codes that are exactly equal in length to the self-information for an item, simply because codes must be discrete. Hence, throughout the article we focused on a particular coding scheme—Huffman coding—that is both simple and approximates optimal compression. However, it is worthwhile to ask whether we find similar results looking not at the length of the Huffman codes for all the items in a given block, but instead at the number of bits of surprise for those items. Thus, we modeled our experiment using surprise to calculate the number of bits for each item rather than the length of the code generated by Huffman coding.

We used the same values for the priors as the Huffman code results in the main text: $\alpha = 34$ and $\alpha = 31$, respectively, for the two experiments. The number of bits of self-information correlate at $r = -.94$ (Experiment 1) and $r = -.95$ (Experiment 2) with human memory performance. Figures A1 and A2 show the results of multiplying the number of bits of surprise with the number of colors remembered by observers for Experiments 1 and 2, respectively. The results once again support the idea of compression as a major factor in visual working memory: Observers are able to remember an approximately fixed number of bits, remembering more colors when the items are more redundant.

Derivation of K Formula

In an eight-alternative forced choice, observers may choose the correct answer for one of two reasons: (a) they may know the correct answer, or (b) they may guess the correct answer by chance. In order to estimate capacity (the number of items remembered out of the eight items in the display), we need an estimate of the first kind of correct answer (knowing the colors), discounting the second kind of correct answer (guesses).

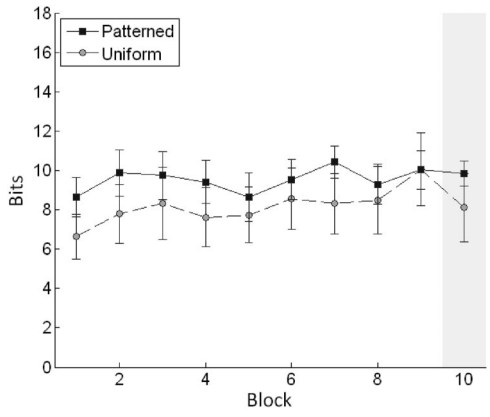

Figure A1. The size of memory for Experiment 1 estimated using self-information. Block 10 is shaded to emphasize that in this block the regularities were removed, so all colors were drawn from a uniform distribution for both groups of observers. Error bars represent ±1 *SEM*.

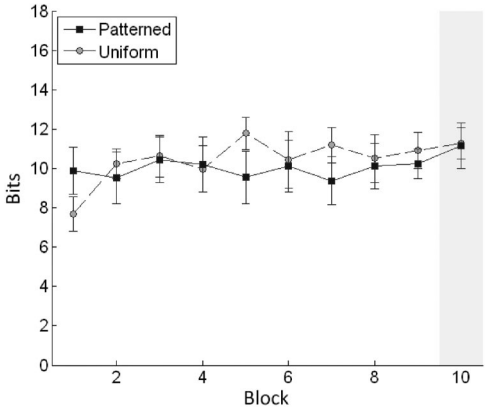

Figure A2. The size of memory for Experiment 2 estimated using self-information. Block 10 is shaded to emphasize that in this block the regularities were removed, so all colors were drawn from a uniform distribution for both groups of observers. Error bars represent ±1 *SEM*.

To begin deriving such a formula we write percent correct (PC) as a function of the two different kinds of answer—answers for those items that observers remember, which they get right 100% of the time, and answers for those items that observers do not remember, which they get right 1/8th of the time. If observers successfully remember K items from a display of eight items, PC may thus be formulated as:

$$\mathrm{PC} = [(K/8) \times 1] + \{[(8 - K)/8] \times 1/8\},$$

where the first term accounts for items correctly remembered and the second term accounts for items on which the observer guesses. For example, if an observer remembers two items ($K = 2$), then for 2/8ths of the items he or she chooses the right answer 100% of the time, whereas the other 6/8ths of the time, he or she guesses and chooses the right answer 1/8th of the time. Simplifying and solving for K, we get

$$(\mathrm{PC} \times 8 \times 8) = (8 \times K) + 8 - K$$

$$(\mathrm{PC} \times 8 \times 8) - 8 = (8 \times K) - K$$

$$(\mathrm{PC} \times 8 \times 8) - 8 = K \times (8 - 1)$$

$$K = [(\mathrm{PC} \times 8 \times 8) - 8]/7.$$

This equation then allows us to directly calculate the capacity of an observer (K) as a function of percent correct (PC).

Received August 1, 2008
Revision received May 6, 2009
Accepted May 11, 2009 ■