Adaptive allocation of human visual working memory capacity during statistical and categorical learning

Christopher J. Bates
Department of Brain & Cognitive Sciences,
University of Rochester, Rochester, NY, USA

Rachel A. Lerch
Cognitive Science Department,
Rensselaer Polytechnic Institute, Troy, NY, USA

Chris R. Sims
Cognitive Science Department,
Rensselaer Polytechnic Institute, Troy, NY, USA

Robert A. Jacobs
Department of Brain & Cognitive Sciences,
University of Rochester, Rochester, NY, USA

Human brains are finite, and thus have bounded capacity. An efficient strategy for a capacity-limited agent is to continuously adapt by dynamically reallocating capacity in a task-dependent manner. Here we study this strategy in the context of visual working memory (VWM). People use their VWM stores to remember visual information over seconds or minutes. However, their memory performances are often error-prone, presumably due to VWM capacity limits. We hypothesize that people attempt to be flexible and robust by strategically reallocating their limited VWM capacity based on two factors: (a) the statistical regularities (e.g., stimulus feature means and variances) of the to-be-remembered items, and (b) the requirements of the task that they are attempting to perform. The latter specifies, for example, which types of errors are costly versus irrelevant for task performance. These hypotheses are formalized within a normative computational modeling framework based on rate-distortion theory, an extension of conventional Bayesian approaches that uses information theory to study rate-limited (or capacity-limited) processes. Using images of plants that are naturalistic and precisely controlled, we carried out two sets of experiments. Experiment 1 found that when a stimulus dimension (the widths of plants' leaves) was assigned a distribution, subjects adapted their VWM performances based on this distribution. Experiment 2 found that when one stimulus dimension (e.g., leaf width) was relevant for distinguishing plant categories but another dimension (leaf angle) was irrelevant, subjects' responses in a memory task became relatively more sensitive to the relevant stimulus dimension. Together, these results illustrate the task-dependent robustness of VWM, thereby highlighting the dependence of memory on learning.

Introduction

In the field of information theory, it is widely known that efficient signal communication depends on the statistical regularities of the information to be communicated (Cover & Thomas, 1991; MacKay, 2003; Shannon & Weaver, 1949). For example, digital audio stored in the MP3 file format can often be reduced in size by a factor of 10 relative to the size of uncompressed audio. This compression is possible, in part, because most audio files are not purely random frequencies, but rather possess rich statistical structure—some frequencies are more common than others, and transitions between frequencies are often predictable. In addition to exploiting statistical regularities, MP3 files compactly store audio signals because they do not attempt to encode the signal perfectly. To the human ear certain frequencies are less discriminable than others, and hence it is less important to encode those frequencies exactly. This example demonstrates that the designers of the MP3 format used two sources of knowledge—knowledge of statistical regularities, and of the kinds of errors that are permissible or less costly—to address the problem of "bit allocation," an instance of the broader problem of distributing a limited or scarce resource to achieve the maximum benefit (Gersho & Gray, 1992). Through bit allocation,

Citation: Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2):11, 1–23, https://doi.org/10.1167/19.2.11.

https://doi.org/10.1167/19.2.11 Received October 4, 2017; published February 25, 2019 ISSN 1534-7362 Copyright 2019 The Authors

engineered systems can store, process, or communicate information in a highly efficient manner.

Here, we are concerned not with the design of engineered systems, but rather with understanding information processing in a particular natural system: human visual working memory (VWM). Since the publication of Miller's famous paper on the "magic number seven, plus or minus two" (Miller, 1956), it has been known that working memory is limited in capacity, and that this capacity can be measured using constructs from information theory. However, a question that has lingered unanswered is the extent to which working memory can be limited yet *efficient*, in the formal sense of making optimal use of its available capacity. This would require a system that is adapted to the statistical regularities of the to-be-remembered items, and also adapted to the importance of storing different dimensions more or less accurately. In other words, this would require a system that has successfully addressed the "bit allocation" problem.

In this paper, we analyze people's VWM performances via rate-distortion theory (Berger, 1971), a branch of statistical decision theory that extends conventional Bayesian approaches through the use of information theory to study rate-limited (or capacity-limited) processes (Sims, 2015, 2016; Sims, Jacobs, & Knill, 2012). Rate-distortion theory provides a normative framework for understanding biological perception, much like theories of perception constructed around Bayesian inference (Knill & Richards, 1996). However, rate-distortion theory extends conventional Bayesian approaches by incorporating a strong theory of capacity limits on information processing. Hence, one should view the current framework as an extension of, rather than replacement for, conventional Bayesian approaches to perception. As discussed below, this framework allows us to quantitatively estimate aspects of people's perception and cognition in a rigorous and principled manner. In particular, we are interested in estimating peoples' VWM capacity, their VWM sensitivity to the distributions of stimulus features, and their VWM sensitivity to the nature of a task, such as which features are more or less important for performing the task.

Our experiments used a novel set of stimuli created using a computer animation software package. The stimuli were images of artificial plants with varying leaf widths and leaf angles. In Experiment 1, subjects performed a change-detection task, a common test of VWM performance. Here, leaf widths were assigned a distribution (e.g., a normal distribution), and we studied whether, over time, subjects' VWM performances would adapt to this distribution. The data indicate that, yes, subjects did indeed adapt to the properties of this stimulus distribution. In Experiment 2, we initially trained subjects to categorize plants when one stimulus dimension (e.g., leaf width) was relevant for distinguishing plant categories whereas another dimension (leaf angle) was irrelevant. Next, subjects performed trials using a change-detection task. Analyses indicate that subjects' responses were more sensitive to the stimulus dimension that was relevant to the categorization task relative to the task-irrelevant dimension.

Taken together, these experiments illustrate the task-dependent robustness of VWM, thus highlighting the dependence of memory on learning. A person who is a poor learner (or a person who is new to a visual environment or task) will have poor knowledge of the statistical regularities of items in the visual environment and poor knowledge of the requirements of visual tasks. Such a person will, inevitably, show poor VWM performances. In contrast, an excellent learner (or a person who is highly familiar with a visual environment or task) will have good knowledge of the statistical regularities of to-be-remembered items and good knowledge of task requirements. As illustrated by the empirical and theoretical findings reported here, the acquisition of this knowledge makes it possible for this person to engage in bit allocation—that is, to allocate resources so as to maximize benefits—and thus to show good VWM performances.[1]

Background literature

A growing body of work points to an important role for knowledge of statistical regularities in VWM, and suggests that use of statistical regularities allows for more efficient memory (Bae, Olkkonen, Allred, & Flombaum, 2015; Brady & Alvarez, 2011; Brady, Konkle, & Alvarez, 2009; Brady & Tenenbaum, 2013; Corbett, 2016; Huttenlocher, Hedges, & Vevea, 2000; Orhan & Jacobs, 2013; Sanocki, Sellers, Mittelstadt, & Sulman, 2010; Sims et al., 2012; Swan, Collins, & Wyble, 2016; Victor & Conte, 2004). For example, Brady and Alvarez (2011) and Corbett (2016) showed that subjects' memories for items in a display are biased toward items' summary statistics, meaning statistical regularities averaged over multiple items in the display. Brady and Tenenbaum (2013) showed that subjects can capitalize on regularities in the spatial arrangements of objects to improve performances in a memory task.

While the above work establishes that VWM can leverage statistical regularities, it does not address whether VWM is dynamically adaptive or, if so, how this adaptation might work. Towards this end, we seek to connect VWM to the phenomenon of implicit *statistical learning* in human visual perception (Fiser &

Aslin, 2001, 2002a, 2002b; Orbán, Fiser, Aslin, & Lengyel, 2008). For example, Orbán et al. (2008) used scenes containing novel shapes arranged in a grid, where the shapes were drawn from a finite set. The arrangements of shapes contained "chunks," where a chunk was a group of shapes that often appeared together in a particular spatial configuration. The authors demonstrated that subjects implicitly learned these chunks, and that their learning was best accounted for by a hierarchical Bayesian model that inferred the most likely chunks given the arrangements of objects observed. We propose that, like visual perception, VWM may be similarly adaptive, if it relies on general statistical learning mechanisms. Thus VWM could quickly tune itself based on the statistical properties of the visual environment.

Prior studies have begun to demonstrate a link between statistical learning and VWM. Brady et al. (2009) showed subjects displays with pairs of concentric circles. Subjects' memories for the colors of these circles improved over a few hundred trials when the colors of circles forming a pair were correlated, but not when they were uncorrelated. When subjects were initially exposed to circles with correlated colors but the correlations were removed in the final block of trials, memory performance dropped to the same level as that of a control group for which colors were always uncorrelated. This result suggests that the improvement in memory performance during the initial phase of the experiment was due to the acquisition of knowledge of the color correlations, as opposed to an increase in capacity.

Huttenlocher et al. (2000) conducted an experiment in which, on each trial, a subject was briefly shown a fish-shaped silhouette whose size was drawn from a size distribution and, after a short delay, the subject was required to reproduce the stimulus from memory by adjusting the size of a probe silhouette. It was found that, over time, subjects showed a bias in their reproductions toward the mean of the size distribution. The authors conceptualized the size distribution as representing a category which the subjects implicitly learned. To account for the "perceptual magnet effect" (i.e., the finding that judgments for stimuli that are clearly within a category are often biased toward the category mean), they used a Bayesian model that computed a weighted sum between the category distribution and a noisy, unbiased memory distribution.

A distinguishing feature of the research presented here is that we analyze our experimental data and interpret our results using a normative information-theoretic framework based on rate-distortion theory, thereby improving on conventional Bayesian approaches by taking into account the fact that VWM is capacity limited (Berger, 1971). This framework allows us to make inferences about subjects' perceptual and cognitive processes at a much finer scale than has previously been possible. For instance, it allows us to separately infer subjects' knowledge of statistical regularities about visual stimuli (which may differ from the true regularities) and knowledge of which types of memory errors are task-relevant versus irrelevant. Maintaining the distinction between these two types of knowledge is important when, for example, a stimulus feature has a statistical regularity, but this feature is irrelevant for the task that the subject is attempting to perform. The framework also allows us to infer subjects' VWM capacities using mathematically meaningful units (e.g., bits) as opposed to the ad hoc or domain-specific units (e.g., slots, items, etc.) that have been used previously in the scientific literature.

Theoretical framework

Sims (2016) provided a tutorial on rate-distortion theory and its application to visual perception and memory. Here, we briefly overview this framework, focusing on the predictions it generates for the current experiments.

The basis for our approach is the assumption that VWM approximates an efficient communication channel. Abstractly, an information channel can be conceived of as an information processing system that takes as input a signal x drawn from some probability distribution $p(x)$, and produces a possibly different signal $\hat{x}$ as output, according to a conditional probability distribution $p(\hat{x}|x)$. As applied to the study of VWM, the channel input consists of afferent sensory information, and the channel output is the perceptual representation held in, or retrieved from VWM. A conceptual diagram is given in Figure 1.

An efficient channel must possess three formal properties (Cover & Thomas, 1991; MacKay, 2003; Shannon & Weaver, 1949). First, the channel must be sensitive to the statistics of the signal conveyed over the channel. That is, an efficient channel must possess and exploit knowledge of the sensory signal distribution $p(x)$. In conventional Bayesian models, knowledge of the sensory distribution $p(x)$ is referred to as the observer's "prior" distribution, and this distribution sets the background knowledge or "probabilistic context" for processing sensory signals. In rate-distortion theory, knowledge about $p(x)$ plays the same role. When applied to the study of VWM, rate-distortion theory predicts that memory performances for a given visual item will differ across contexts, such as when the item is sampled from a uniform distribution versus a normal distribution across possi-

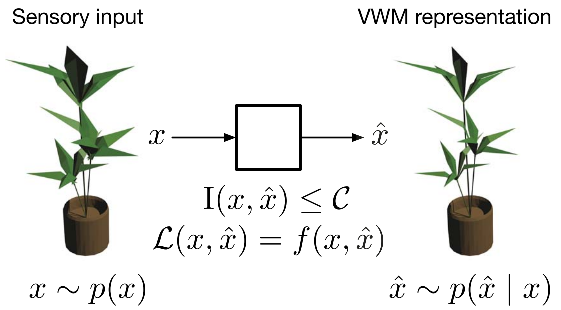
Sensory input
VWM representation
x
$\hat{x}$
$\mathrm{I}(x,\hat{x}) \le \mathcal{C}$
$\mathcal{L}(x,\hat{x}) = f(x,\hat{x})$
$x \sim p(x)$
$\hat{x} \sim p(\hat{x} \mid x)$

Figure 1. VWM is modeled as an information-theoretic communication channel. Sensory signals (in this example, images of houseplants) are input to the channel; the output is the (error-prone) VWM representation. The channel has finite capacity, meaning that the mutual information between channel input and output, $\mathrm{I}(x,\hat{x})$, must be less than or equal to some finite capacity $\mathcal{C}$. Lastly, to be an efficient channel, it is optimized to minimize a specified cost function, $\mathcal{L}(x,\hat{x})$. For an efficient channel, the cost function is defined by the task the observer is performing.

ble items (assuming the memory system has been given time to learn a particular distribution). The idea that neural systems adapt to their afferent signal statistics has received extensive empirical support in sensory neuroscience, where it is known as the efficient coding hypothesis (Barlow, 1961).

Second, a channel must have sufficient *channel capacity* to achieve the desired level of performance. When channel capacity is lower than the statistical complexity of the information source (measured by its information entropy; Cover & Thomas, 1991; MacKay, 2003), it is certain that errors must occur in the course of information transmission. While capacity limits present a major topic of study in VWM (for a review, see Ma, Husain, & Bays, 2014), information theory contributes a principled measure of limited-capacity perceptual systems.

Lastly, for a communication channel to be optimal, it is necessary to quantify the criterion for performance in terms of a *cost function*. While limits in capacity necessitate errors in information processing, an optimal system should seek to minimize the cost of perceptual error. The cost function $\mathcal{L}(x,\hat{x})$ specifies the goals for the channel in terms of the cost of reproducing signal x as a possibly different signal $\hat{x}$. For unitary stimuli, a simple cost function might be the squared error $(\hat{x} - x)^2$. For an optimally efficient channel, the cost function is defined by the behavioral task that the organism seeks to perform. Note that the organism's implicit task may not necessarily agree perfectly with the experimenter-defined task.

With these three properties characterized, it is possible to define an optimally efficient communication channel as a channel that minimizes expected (or average) cost for the given cost function, subject to a constraint on available channel capacity. This can be stated as follows:

$$\text{Goal: Minimize } \mathrm{E}[\mathcal{L}(x,\hat{x})] \text{ with respect to } p(\hat{x}|x), \\ \text{subject to } \mathrm{I}(x,\hat{x}) \le \mathcal{C}. \quad (1)$$

The term $\mathrm{I}(x,\hat{x})$ refers to the mutual information between the channel input x and output $\hat{x}$ (it is a measure of the amount of information conveyed by the channel). This quantity is constrained by the limit on information rate for the channel, $\mathcal{C}$. In the current work we use an efficient numerical algorithm for solving this constrained optimization problem (see Blahut, 1972 and Sims, 2016).

Although rate-distortion theory and conventional Bayesian approaches often make similar predictions—not unsurprising given that rate-distortion theory makes extensive use of Bayesian statistics—these predictions are not always identical. Differences in their predictions stem from the fact that rate-distortion theory assumes that the processes under study are information rate- or capacity-limited, whereas conventional Bayesian approaches do not make assumptions about capacity limits.

An important difference, relevant to Experiment 1 discussed below, arises when one considers how an ideal observer's memory performance should change with changes in the distribution of visual stimuli. This situation is extensively discussed in Appendix A where it is shown that, given modest mathematical assumptions, rate-distortion theory predicts that an ideal observer's memory performance should steadily decline with increases in the standard deviation of the stimulus distribution, whereas a conventional Bayesian approach predicts that this performance will degrade slowly (see Figure 2). At an intuitive level, these differences in predictions are expected. Rate-distortion theory assumes that the ideal observer allocates its limited capacity to cover the entire stimulus range, and thus this capacity is "spread thinner" as the size of the stimulus range increases. In contrast, a conventional Bayesian model assumes that an ideal observer does not have a capacity limit, and thus the observer's memory performance can be robust to increases in the size of the stimulus range. As reported below, the results of Experiment 1 are qualitatively consistent with the predictions of rate-distortion theory.

Additional important predictions of the rate-distortion framework are illustrated in Figure 3. In this figure, sensory signals are assumed to vary along a unidimensional distribution $p(x)$. In Figure 3a, a stimulus x_0 (indicated by the vertical line) is sampled from a probability distribution $p(x)$, shown in blue. The stimulus x_0 is stored in a capacity-limited VWM. Rate-distortion theory predicts that the precision of the distribution of VWM representations $[p(\hat{x}|x = x_0)$, referred to as the memory distribution and shown in

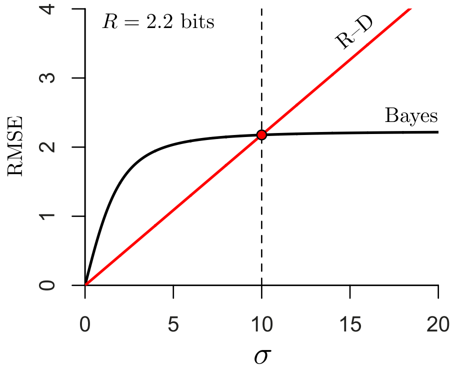

Figure 2. The horizontal axis plots the standard deviation of the stimulus distribution, and the vertical axis plots the root mean squared error in an ideal observer's memory performance. Given modest mathematical assumptions (see Appendix A), rate-distortion theory predicts that memory performance will rapidly diminish with increases in the standard deviation σ of the stimulus distribution (red line). In contrast, a conventional Bayesian approach predicts that performance will degrade slowly. Both models are calibrated to produce identical performance at $\sigma = 10$. For the rate-distortion model, information capacity was fixed at 2.2 bits, close to the value estimated from subjects' data in Experiment 1.

orange] depends on the capacity of VWM. The left panel assumes an available capacity of 3 bits, whereas the right panel has a capacity of 1 bit. This is reflected in the figure, because the orange distribution is narrower (more precise) in the left panel compared to the right.

Figure 3b illustrates rate-distortion theory's prediction that the precision of a VWM memory distribution for a given stimulus depends strongly on the entropy (approximately, the width) of the stimulus distribution. The examples in the left and right panels have the same available channel capacity (1 bit), but memory precision is greater when stimuli are sampled from a narrower (lower entropy) distribution. As discussed above (see Figure 2), this effect is not predicted by conventional Bayesian theories of perception (see Appendix A).

As demonstrated in Figure 3c, limitations in capacity also provide an elegant explanation for "set size" effects in VWM (i.e., VWM performance decreases as the number of visual items in a display increases). The left panel shows the predicted VWM memory distribution as a function of the number of items stored, assuming that VWM's total capacity is evenly divided among items. For example, for set-size 6, the plotted curve represents the memory distribution for just one of the six items, but curves for the five other items would have the same shape. The right panel shows the predicted recall standard deviation as a function of set size (that is, the standard deviation of the memory distributions plotted in the left panel). A model based on rate-distortion theory has been shown to provide a close quantitative account of empirical set size effects (Sims et al., 2012). In contrast, conventional Bayesian models of VWM do not provide a satisfying account of set size effects. These models typically assume that memory representations are corrupted by additive noise. To account for set size effects, the variance of this noise must increase with set size (e.g., Orhan & Jacobs, 2013). This ad hoc assumption made for the

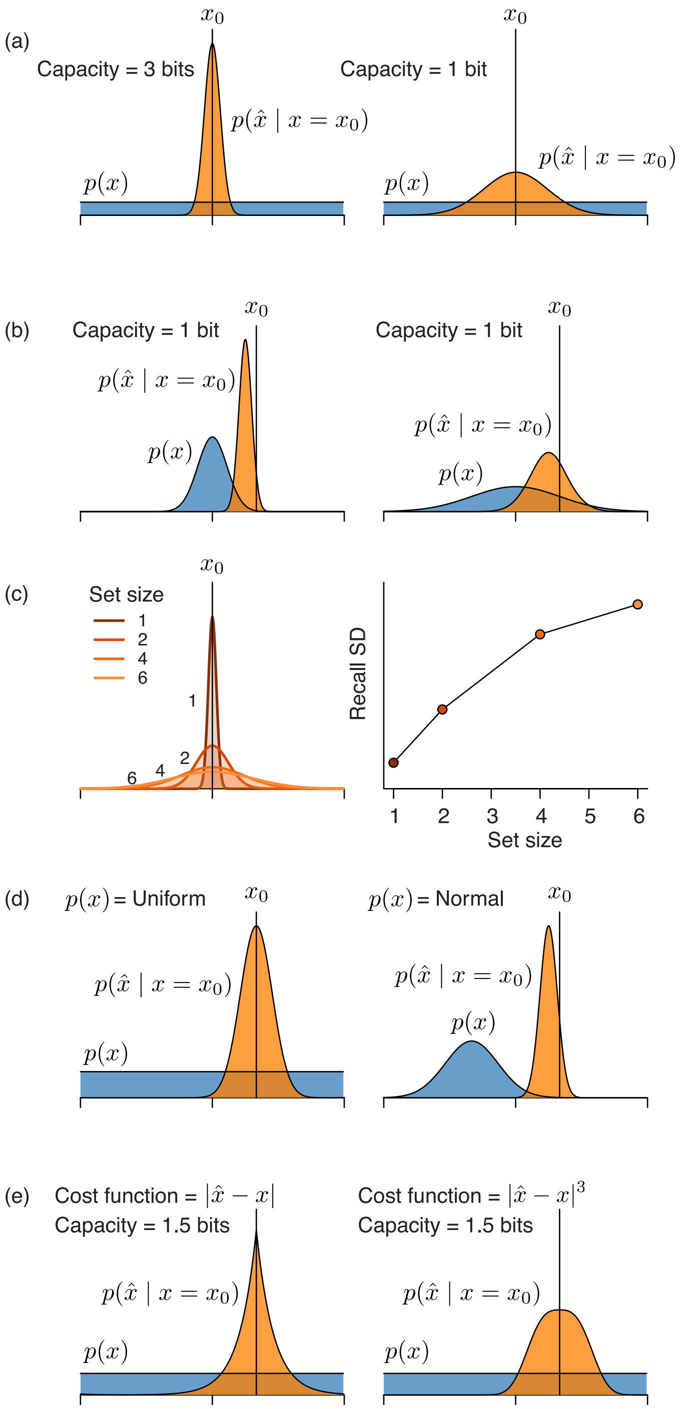

Figure 3. Illustrations of some predictions of the rate-distortion framework. See text for explanations.

purpose of "curve fitting" experimental data is necessary because these models do not inherently include a notion of rate-limited (or capacity-limited) processing.

Figure 3d demonstrates that when stimuli are drawn from a non-uniform distribution, rate-distortion theory (like Bayesian inference) predicts that the VWM memory distribution will be biased towards the mean of the stimulus distribution. In the left panel, the stimulus distribution is uniform, whereas this distribution is normal in the right panel. Consequently, the VWM memory distribution in the right panel is biased towards the mean of the stimulus distribution, whereas the distribution in the left panel is not.

Rate-distortion theory also predicts that the shape of a VWM memory distribution depends critically on the cost function that VWM seeks to optimize, as demonstrated in Figure 3e. The VWM memory distributions in the left and right panels differ because the left panel assumes an absolute error criterion, whereas the right panel assumes a cubic error function. (All other panels in Figure 3 assume a squared error cost function.)

To validate the application of rate-distortion theory to the study of VWM, all of the theory's predictions need to be experimentally evaluated. In this paper, we focus on two predictions. First, to be efficient, VWM should be sensitive to the statistics of sensory information. This would be demonstrated by changes in VWM memory distributions with changing stimulus distributions. Second, VWM should be sensitive to the costs of memory error. Different tasks that impose different demands and costs of error should lead to different distributions of errors produced in memory. To our knowledge, these predictions are not shared by competing theories of VWM (for an overview of alternative theories of VWM, see Ma et al., 2014). While previous research on VWM has uncovered evidence of both statistical adaptation effects and adaptation to the demands of a task, rate-distortion theory uniquely unifies these findings within a single coherent and normative framework that is specialized for the study of rate-limited processes.

Experiment 1

Suppose a person views a set of plant leaves, and then attempts to recall the leaves' colors. The person may make memory errors but, intuitively, these errors will not be uniformly distributed. Instead the errors should exhibit a systematic pattern in which some errors are more likely than others. For instance, it is unlikely that the person will make an error by recalling the color of a leaf as blue, because the person has learned over time that the probability of a blue leaf is very small. By contrast, the same may not be true if the item to be remembered is someone's shirt, instead of a leaf. The goal of Experiment 1 was to examine whether and, if so, how subjects' VWM performances adapt to the distribution of stimulus features.

Subjects

The experimental study was approved by the Research Subjects Review Board at the University of Rochester. Three hundred twenty-four subjects (81 subjects in each of four conditions) participated in the experiment over the world wide web via the Amazon Mechanical Turk (MTurk) crowd-sourcing marketplace. Interfacing with MTurk was facilitated through the use of the psiTurk programming platform (Gureckis et al., 2016). psiTurk was configured so that only individuals based in the United States could participate in the experiment. Subjects stated that they were at least 18 years old. It took approximately 30 min to complete the experiment, and each subject received $3.25 for his or her participation.

Stimuli

Visual stimuli were computer-rendered images of artificial plants created using Blender, a computer graphics package for three-dimensional (3-D) modeling and animation. A 3-D model of a plant was read into Blender, and "shape keys" were defined characterizing the width of a plant's leaves and the angle at which a plant's leaves sag. Using these shape keys, morphs of a plant were created and rendered. As illustrated in Figure 4, the collection of plants can be regarded as residing in a 2-D space where leaf width varies along one dimension and leaf angle varies along the other dimension. Experiment 1 used plants with varying leaf width but fixed leaf angle. Leaf width was discretized to 101 values. Experimental stimuli did not exceed the edges of the stimulus space, meaning that stimuli ranged from the plant with the narrowest leaves, referred to as plant-0, to the plant with the widest leaves, referred to as plant-100.

These stimuli are an advance over stimuli used in the majority of previous experiments studying VWM because they are both naturalistic and easily controlled in a precise manner. Previous experiments have tended to use simple (perhaps simplistic) stimuli, such as oriented bars or colored squares. The advantage of these simple stimuli is that they can be precisely controlled. The disadvantage of using these stimuli is that there are empirical and theoretical reasons to question whether VWM performances with these stimuli are representative of performances in more

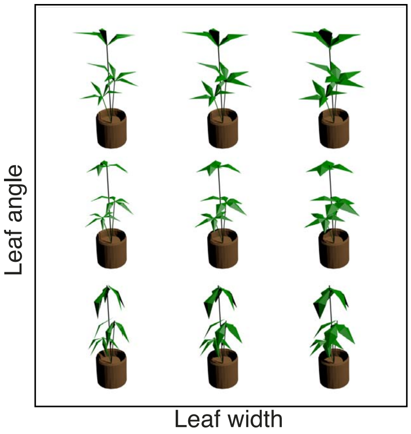

Figure 4. Two-dimensional space of artificial plants. The width of a plant's leaves varies along the horizontal axis from narrow to wide. The angle of a plant's leaves varies along the vertical axis from an angle in which leaves sag downward to an angle in which leaves are horizontal. In Experiment 1, the leaf-width dimension was discretized to 101 values, and thus the experiment used 101 images. In Experiment 2, the leaf-width and leaf-angle dimensions were each discretized to 11 units, and thus the experiment used 121 images. For illustrative purposes, this figure shows nine images.

natural settings (Brady, Störmer, & Alvarez, 2016; Endress & Potter, 2014; Orhan & Jacobs, 2014). Experiments with natural stimuli are infrequent because these stimuli are difficult to control (Rust & Movshon, 2005). The stimuli used in our experiment strike a good balance between naturalness and control.

Procedure

After providing informed consent, reading the instructions, and successfully completing a multiple-choice quiz testing their understanding of the instructions, subjects performed a change-detection task. On each trial, a subject viewed a fixation cross for 750 ms, a target image for 2000 ms, a blank screen for 1000 ms, and then a probe image that stayed on the screen until the subject responded. The probe image was randomly displaced to the left or right by 15 pixels relative to the target image. If the subject believed that the plants depicted in the target and probe images were the same, the subject was instructed to press the "s" (same) key. Otherwise the subject should press the "d" (different) key. Following the response, the subject was provided with feedback indicating whether the response was correct. Each subject performed two practice trials followed by 220 experimental trials. Half of the experimental trials were "same" trials and the remaining were "different" or "change" trials. On a change trial, the leaf widths of the plants depicted in the target and probe images differed by an amount that was drawn randomly from the set $\{-2\Delta, -\Delta, \Delta, 2\Delta\}$, where Δ was 0.08 times the maximum possible difference in leaf width (i.e., the leaf width for plant-100 minus the leaf width for plant-0).

Each subject participated in one of four possible experimental conditions. Different conditions used different stimulus distributions. In one condition, plants depicted in target images were randomly sampled from a uniform distribution over the set of possible plants. In the remaining three conditions, target plants were sampled from normal distributions whose means were plant-30, plant-50, and plant-75, respectively (standard deviations were set to 10).

Data analysis: Rate-distortion theory

To apply rate-distortion theory to the current experimental data, it is necessary to specify the three properties discussed above: the subject's knowledge of the stimulus distribution $p(x)$, the capacity of the channel $\mathcal{C}$, and the cost function that is minimized $\mathcal{L}(x,\hat{x})$. We assume that subjects' belief about the stimulus statistics, $\tilde{p}(x)$, follows a normal distribution, $\tilde{p}(x) = \mathcal{N}(\mu,\sigma)$, and we infer μ and σ. Since the channel capacity of VWM is not known in advance, it is estimated from the experimental data as a free parameter in the model. In Experiment 1, the cost function, which specifies the cost of each possible memory error $\hat{x} \neq x$, is assumed to be squared error.

Rate-distortion theory provides predictions for an optimal, but capacity-limited VWM system. Presented with a particular plant stimulus x, the model generates predictions for the distribution of possible memory errors, $p(\hat{x}|x)$. To apply this model to our experimental task, it is necessary to specify how noisy memory representations are mapped onto binary responses in a change-detection task. In keeping with our goal of deriving a normative model, we assume that subjects compute the posterior probability that a change has occurred, given a noisy memory representation $\hat{x}$ and a probe stimulus y, via Bayes' theorem:

$$p(\text{“change”}|\hat{x}, y) \propto p(\hat{x}, y|\text{“change”})p(\text{“change”}). \quad (2)$$

This introduces one additional parameter into the model, namely the prior probability that a given trial will be a change trial, $p(\text{“change”}) = p_{\text{change}}$. To allow for noise in responses, we assume that subjects exhibit probability matching (Vulkan, 2000). That is, the probability that a subject responds "different" (or "change") on a given trial is equal to the computed probability that a change has occurred (see Keshvari, van den Berg, & Ma, 2013, for a similar application of this idea in modeling VWM). A schematic of the VWM model and the decision rule is shown in Figure 5. A

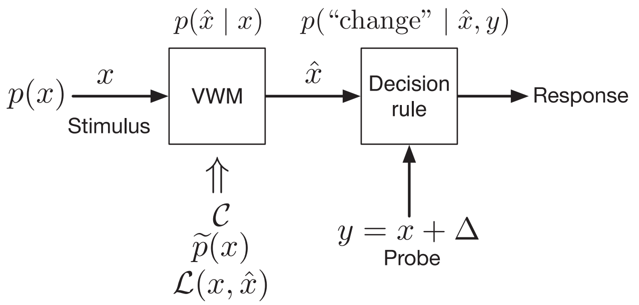

Figure 5. Illustration of an information-theoretic model of VWM for change detection. Information theory states that an efficient communication channel must possess three properties: knowledge of the relevant signal statistics ($\tilde{p}(x)$), a sufficient channel capacity ($\mathcal{C}$) to achieve the desired level of performance, and a cost function ($\mathcal{L}(x, \hat{x})$) that defines the cost of errors in signal communication. A decision rule infers the probability that a change has occurred using Bayes' theorem, given the error-prone memory representation ($\hat{x}$) and the probe stimulus (y).

derivation of the Bayesian decision rule is provided in Appendix B.

In summary, the model for Experiment 1 requires estimating four parameters: channel capacity $\mathcal{C}$, the prior probability of a change trial p_{change}, μ, and σ. We fit the model to the aggregated data from all subjects in each condition (220 trials × 81 subjects = 17,820 trials). However, we chose to make our model hierarchical by fixing p_{change} and capacity across subjects, as we did not expect these two parameters to be condition-dependent (see below). Model parameters were estimated by means of maximum likelihood estimation.[2]

	Mean 30	Mean 50	Mean 75	Uniform	Global
μ	31.4	49.3	58.6	34.6	
σ	20.2	18.8	21.9	32.7	
$\mathcal{C}$					2.26
p_{change}					0.40

Table 1. Maximum likelihood estimates of model parameters in each of the four conditions of Experiment 1. *Notes*: The last column (Global) gives values of parameters that were shared across conditions. Capacity $\mathcal{C}$ is measured in bits.

Results

Because reliable parameter estimates require large numbers of data items, we aggregated the responses of all subjects in each experimental condition. Results are shown in Figure 6 and Table 1. In the left graph of Figure 6, the horizontal axis gives the trial number and the vertical axis gives the average percent correct for each condition. If VWM adapts to the stimulus statistics in order to improve performance, we should expect better performance in the three conditions using a normal distribution than in the condition using a uniform distribution, since a uniform distribution has greater entropy (demonstrated in Figures 2 and 3b). The left graph of Figure 6 shows that this prediction was born out: Performance increased as the entropy of the stimulus distribution decreased ($p < 0.01$ for each pairwise comparison between performances in the uniform condition versus a normal condition, where hypothesis testing was conducted via bootstrapping). This outcome is broadly consistent with the predictions of rate-distortion theory, and does not support

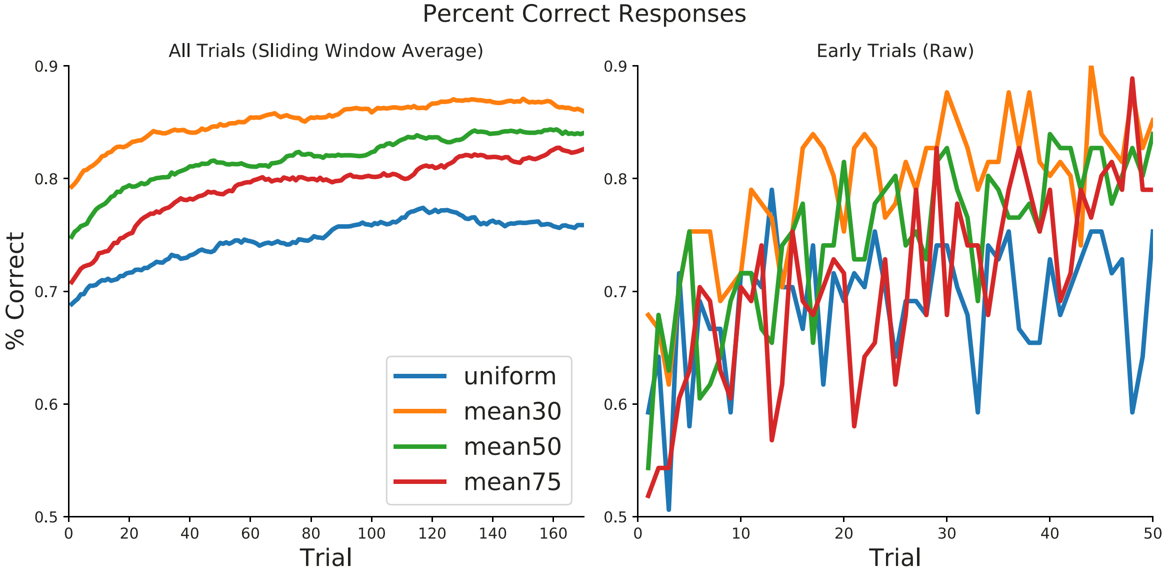

Figure 6. (Left) Average percent correct for aggregated subject data as a function of trial number for each experimental condition. Values were computed using a 50-trial sliding window with an interval of 1 (i.e., Trials 1–50, 2–51, etc.). (Right) Raw (i.e., unsmoothed) average percent correct for aggregated subject data for Trials 1–50.

conventional Bayesian approaches that predict that memory performance should be largely insensitive to changes in the widths of stimulus distributions (see discussion above, Figure 2, and Appendix A). A potential problem with the left graph of Figure 6 is that it suggests that performance differences between conditions existed at the start of the experiment. This problem arises because we smoothed the data across trials to prepare this graph. The graph on the right shows the raw (i.e., non-smoothed) data for the first 50 trials. It indicates that performance differences did not exist at the start of the experiment, though they appear to have arisen due to learning relatively early during training.

Comparing the three conditions using a normal distribution, it seems that subjects performed best when plants tended to have narrow leaves. Note, however, that leaves tended to be narrower in the condition using a uniform distribution than in the condition using a normal distribution with a mean of 75 but, despite this, subjects performed better in the latter condition. Thus, although the task was easiest when leaves were narrow, this fact cannot explain subjects' relative poor performance in the uniform condition.

Table 1 summarizes our modeling results from Experiment 1. We used maximum likelihood estimation to infer values of our model's parameters based on subjects' responses. When doing so, we inferred parameter values for all experimental conditions simultaneously in a hierarchical fashion: The parameters that we did not expect to depend on condition were shared across conditions ($\mathcal{C}$ and p_{change}), while the remaining parameters were allowed to vary by condition (μ and σ). We report values using only subjects' last 100 trials. (We found that a trial window of about 100 was necessary for stable estimates.)

The stimulus mean μ and standard deviation σ estimates are sensible, and follow the expected qualitative trend. In the uniform condition, σ was higher than in all others, indicating that subjects in this condition learned a higher variance stimulus distribution. In the normal conditions with means of 30, 50, and 75 (standard deviations set to 10 in all cases), μ was about 30, 50, and 60, respectively, while σ was always about 20. We find these outcomes to be impressive. With relatively little exposure, subjects in each condition learned roughly the correct stimulus distribution and used this information when making VWM judgments. Taken as a whole, these results support our rate-distortion theory framework because they strongly suggest that subjects adapted their VWMs based on the stimulus distributions in order to improve memory performances.

One might speculate that performances differed in the various conditions simply because the conditions used different sets of stimuli on average, and that changes to some stimuli (such as plants with broad leaves) were more detectable than others. To rule out this possibility, we conducted the following analysis. From the set of trials from the uniform stimulus condition, we subsampled three different subsets. In these subsets, the distribution of stimuli matched exactly the distributions used for the plant-30, plant-50, and plant-75 conditions. If performance depended on the physical stimuli, and *not* adaptation to statistical context, performance in the subsampled trials should be identical to those in the corresponding normal conditions.

The results strongly reject this explanation. On the three subsets with stimulus means of 30, 50, and 75, subjects responded correctly on 75%, 71%, and 73% of the trials. The corresponding performances by subjects in the conditions using the normal stimulus distributions were 84%, 81%, and 78%. Binomial tests found significant differences for all three cases ($p < 1 \times 10^{-10}$). These results are general in the sense that they held for 50 replications where each replication used different subsampled subsets.

In summary, the experimental data indicate that subjects in each condition learned (roughly) the correct stimulus distribution to improve their performances on the change-detection task. Moreover, they did this relatively quickly. Perhaps counter-intuitively, memory performance depended not just on the stimuli that were to be remembered, but the probabilistic context from which the stimuli were sampled. These results are consistent with the hypothesis that good memory requires good learning.

Lastly, we also used maximum likelihood estimation to fit a Bayesian observer model to subjects' responses. As expected based on our discussion above, our capacity-limited rate-distortion model provides a comparatively better fit to the experimental data than the Bayesian model. Further details about the comparisons of the two models can be found in Appendix A.

Experiment 2

In Experiment 2, we used category-learning as a way to explore how people dynamically reallocate VWM capacity. Arguably, the most important statistical property of everyday environments is that objects tend to fall into groupings or categories (Mervis & Rosch, 1981; Pothos & Wills, 2011; Smith & Medin, 1981). Animals can be categorized as fish, reptiles, birds, or mammals; fruits can be categorized as apples, pears, peaches, or bananas; and mushrooms or herbs can be categorized as poisonous or safe. Unsurprisingly, the ability to categorize objects is fundamental to human cognition.

An interesting aspect of many categorization tasks is that some feature dimensions are relevant for identifying category membership, whereas other dimensions are not. For instance, body shape is relevant for deciding whether a car is a Ford Mustang or a Chevrolet Camaro, but body color is not. To experts in a domain, it is obvious which feature dimensions are relevant for understanding the categorical structure of items in the domain. To novices, determining which dimensions are important can be a difficult challenge.

Given this challenge, a possible approach might be to perceive, remember, and process *all* feature values of observed items. Unfortunately, people are not always able to do so. Because people's perceptual and memory systems have information processing limits, they cannot simultaneously perceive all features of all objects in an environment and, even if they could, they could not remember and process all this information.

It would seem to be efficient if a person with limited capacity deploys that capacity in a strategic manner, by allocating memory resources according to the relative importance of remembering different features accurately. We hypothesize that people dynamically reallocate their limited VWM capacity based on the task-defined importance of remembering different features, particularly as defined by the categorical structure of remembered items. We predict that people allocate more capacity toward remembering the feature values of objects when those features are relevant for determining objects' category memberships.

Experiment 2 was conducted in the same manner as Experiment 1 with exceptions noted below.

Subjects

One hundred and one subjects participated in Experiment 2. As in Experiment 1, subjects were recruited and completed the experiment via Amazon Mechanical Turk (MTurk).

Stimuli

Visual stimuli depicted artificial plants residing in the 2-D stimulus space illustrated in Figure 4. The leaf-width and leaf-angle dimensions were each discretized to 11 units, and thus the experiment used 121 images.

Procedure

The experiment contained two stages. During the first stage, subjects performed categorization trials. By defining a (deterministic) boundary in the stimulus space, plants were assigned to one of two categories, labeled "taxiforma alpha" and "taxiforma beta" (fictional plant subspecies). For one group of subjects ($n = 50$), the widths of a plant's leaves determined the plant's category membership (leaf angle was an irrelevant feature dimension). For instance, plants with narrow leaves may have been members of taxiforma alpha, whereas plants with wide leaves were members of taxiforma beta. For the remaining subjects ($n = 51$), the angles of a plant's leaves determined the plant's category membership (leaf width was an irrelevant dimension). On each of 64 categorization trials, a plant, referred to as the target, was randomly selected (with the constraint that the target could not lie on the category boundary). The subject viewed a fixation cross (displayed for 500 ms) and then an image of the target (displayed for 1000 ms). Then the subject judged the target's category. A feedback message indicated whether the subject's response was correct.

The second stage consisted of 88 trials. At random, half of these were categorization trials and half were change-detection trials. A trial started by displaying a fixation cross followed by an image of a target plant. If the trial was a categorization trial, the subject was then asked to judge the target's category. If the trial was a change-detection trial, the image of the target was followed by a blank screen (displayed for 1000 ms), which was then followed by the displays of a fixation cross (500 ms) and an image of a second plant, referred to as the probe, which remained on the screen until the subject responded. The position of the image of the probe (and the preceding fixation cross) was randomly jittered to the left or right relative to the position of the image of the target. At random, the probe was identical to the target on half the change-detection trials. When the probe differed from the target, it differed by either 0, 1, 2, or 3 units along the leaf-width dimension, the leaf-angle dimension, or both (with the constraint that it could not differ by 0 units along both dimensions, and with the constraint that the probe had to lie within the 2-D space). After the image of the probe appeared, a subject judged whether the images of the target and probe were visually identical ("same" response) or not ("different" response). Feedback was not provided on change-detection trials.

Results

We analyzed the experimental data using mixed-effects logistic regression and using rate-distortion theory.

Mixed-effects logistic regression

Subjects' responses on the change-detection trials were analyzed using mixed-effects logistic regression models. In general, logistic regression is an extension of linear regression that is suitable for modeling binary response data (e.g., "same" and "different" responses; McCullagh & Nelder, 1989). A mixed-effects logistic regression is an extension of a logistic regression that describes a relationship between a response variable and independent variables whose coefficients can vary with respect to one or more grouping variables (e.g., one subset of responses was produced by one subject, whereas another subset of responses was produced by another subject). Consequently, it can represent the covariance structure related to the grouping of data (Faraway, 2016; Gelman & Hill, 2007; Knoblauch & Maloney, 2012; Moscatelli, Mezzetti, & Lacquaniti, 2012).

Each data item contained five numbers describing a trial: (a) the response variable indicated whether a subject responded "same" or "different"; (b) the independent variable delta-leaf-width was set to the absolute value of the difference in leaf width between the first and second plants; (c) delta-leaf-angle was set to the analogous value for leaf angle; (d) condition indicated whether a subject was trained in the categorization condition for which leaf width was the relevant dimension or the condition for which leaf angle was the relevant dimension; and (e) subject gave the subject number. Delta-leaf-width, delta-leaf-angle, and condition were normalized to have a mean of zero and variance of one. The regression was performed using the "lme4" library (Bates, Mächler, Bolker, & Walker, 2015) of the R statistical computing environment (R Core Team, 2015).

We defined three nested models where Model $\mathcal{M}_1$ is the most complex, $\mathcal{M}_2$ is nested within $\mathcal{M}_1$, and $\mathcal{M}_3$ is nested within $\mathcal{M}_2$. These models reflect different assumptions regarding the influence of the training condition on change detection performance. In $\mathcal{M}_1$, population-level effects (so-called *fixed effects*) were modeled using an intercept and independent variables delta-leaf-width, delta-leaf-angle, and condition, as well as interactions between delta-leaf-width and condition and between delta-leaf-angle and condition. The intercept and the coefficients on delta-leaf-width and delta-leaf-angle were allowed to vary by subject (*random effects*). Model $\mathcal{M}_2$ was identical to $\mathcal{M}_1$ except that it omitted the interaction terms in the set of independent variables. Model $\mathcal{M}_3$ was identical to $\mathcal{M}_2$ except that it omitted the condition variable. If categorization training led subjects to dynamically reallocate VWM capacity by assigning more capacity to the feature dimension that was relevant during training, then the interaction terms (i.e., Delta-Leaf-Width × Condition and Delta-Leaf-Angle × Condition) should be essential. That is, $\mathcal{M}_1$ should outperform $\mathcal{M}_2$ and $\mathcal{M}_3$. Otherwise, the performances of the models should not significantly differ.

As predicted, $\mathcal{M}_1$ (AIC score = 3461.9) performed better than $\mathcal{M}_2$ (AIC score = 3467.4), with the difference in their performances being statistically significant based on a likelihood ratio test, $\chi^2(2) = 9.46$, $p = 0.009$. $\mathcal{M}_1$ also performed significantly better than $\mathcal{M}_3$ (AIC score = 3465.7), and $\chi^2(3) = 9.8242$, $p = 0.02$. The performances of $\mathcal{M}_2$ and $\mathcal{M}_3$ did not significantly differ ($p = 0.55$). Consistent with our hypothesis, these results indicate that subjects allocated more VWM capacity toward remembering the feature values of objects when those features were relevant for determining objects' category membership. In other words, subjects dynamically reallocated their limited VWM capacity based on the categorical structure of the to-be-remembered items.

Rate-distortion analysis

We further analyzed our data using the same rate-distortion model as used in Experiment 1. However, in contrast to the previous experiment, here stimuli were chosen at random from a 2-D space ($p[x] = 1\ /\ N$ for all stimuli x). Furthermore, we assumed that subjects adopted a perfect model of the stimulus statistics, hence $\tilde{p}(x) = p(x) = 1/121$ (there were 121 possible stimuli). Since our goal in this experiment was to measure changes in subjects' cost functions, the cost function was also estimated from the data (rather than assuming a squared error cost function as in Experiment 1). For each plant stimulus x, the cost function specifies the cost of each possible memory error $\hat{x} \neq x$. In the general case, fully specifying such a cost function requires $121 \times 120 = 14{,}520$ parameters. Thus we made two simplifying assumptions.

First, we assumed that costs must satisfy a metric space. This approach builds directly on Shepard's universal law of generalization (Shepard, 1987). In our model, the cost for misremembering stimulus x as a different stimulus $\hat{x}$ defines the Euclidean distance between the two stimuli within a 2-D "perceptual" space. In addition, we assumed that the costs for each stimulus dimension are independent so that, for example, the cost for misremembering a wide leaf for a narrow one is independent of the leaf angle. With these assumptions, the cost function can be compactly specified by $10 + 10$ interstimulus distances.

As indicated here and discussed below, these parameters define the "perceptual coordinates" of each plant stimulus. If the cost of misremembering one stimulus as another is small, then the two stimuli will be close together in a perceptual space (i.e., the stimuli will be regarded as perceptually similar). In contrast, if the cost of confusing two stimuli is large, then the stimuli will be far apart in this space. We hypothesized that category training would induce changes in this perceptual space, such that category-relevant features

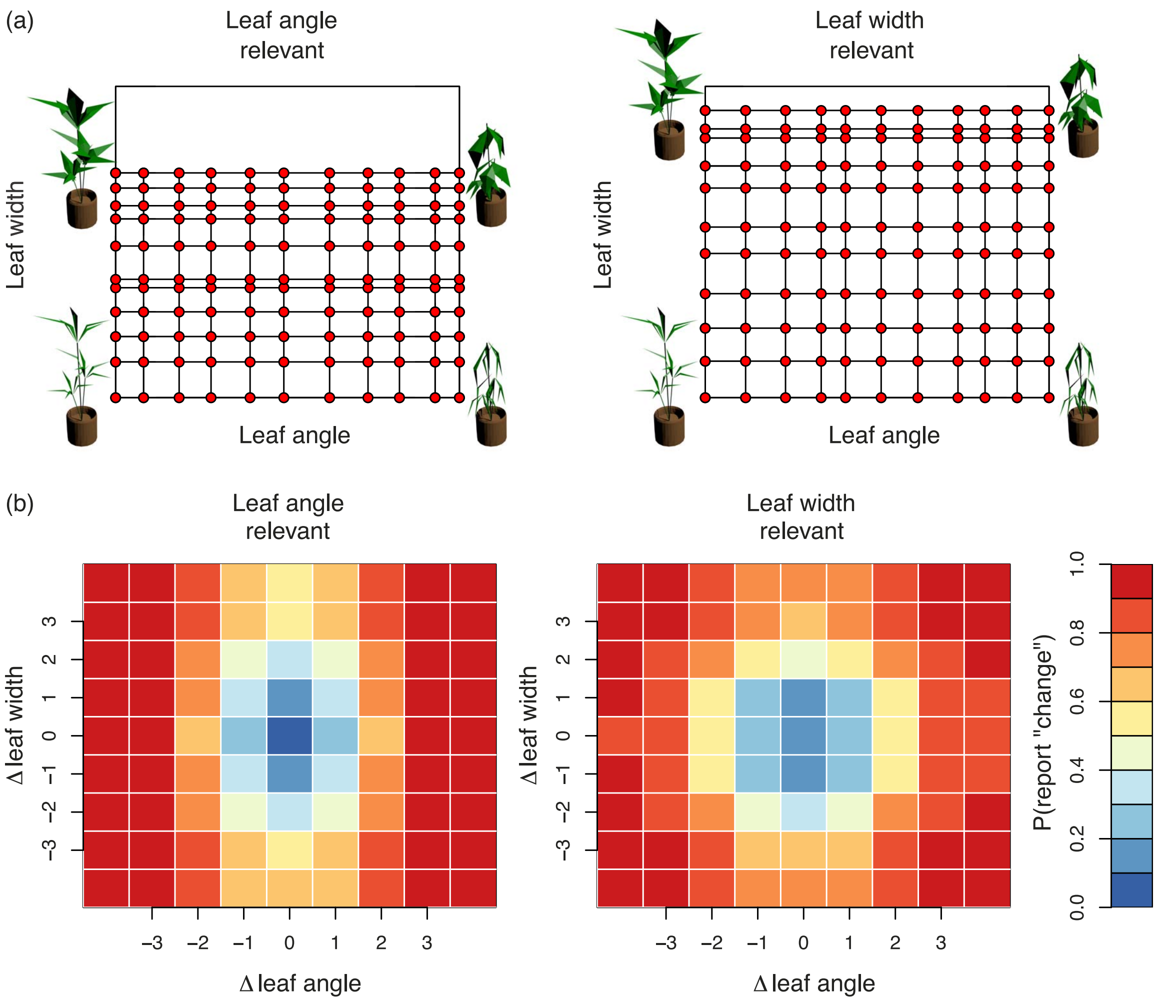

Figure 7. (a) Changes in perceptual space induced by category training. Each point in the grid represents a stimulus. The cost function for VWM error, $\mathcal{L}(x, \hat{x})$, defines the Euclidean distance between points in this space. Training in the condition where leaf width is relevant (right panel) leads to an expansion in perceptual space along this dimension such that memory errors in leaf width are more costly relative to the condition where leaf angle is relevant. (b) Probability of reporting "different" or "change" as a function of changes to leaf angle and leaf width, estimated according to the model. The figure illustrates that observers are more sensitive to changes in leaf angle in the leaf-angle relevant condition (left panel). This adaptation is accompanied by a decrease in sensitivity to changes in leaf width when compared to the data from the leaf-width relevant condition (right panel).

would become more distinct. The model in this experiment required estimating 22 parameters: channel capacity $\mathcal{C}$, the prior probability of a change trial p_{change}, and the 20 parameters characterizing the cost function. We fit the model to the aggregated data from all subjects in each condition (a minimum of 2,200 trials). Model parameters were inferred using maximum likelihood estimation.

The estimated channel capacity for the leaf-angle relevant condition was 4.16 bits, while the capacity for the leaf-width relevant condition was 4.06 bits. Given these highly similar values, the model did not account for differences between the conditions by assuming that total memory capacity differed. The parameter p_{change} was also nearly identical between conditions: 0.21 in the leaf-angle relevant condition, and 0.20 in the leaf-width relevant condition. Hence, according to the model, performance in the two conditions differed only in terms of the implicit cost function for memory error.

The estimated "perceptual spaces" for the stimuli are shown in Figure 7a. Each point in this grid represents the location of a specific plant stimulus in psychological space. Points that are closer together are more perceptually (or mnemonically) similar, and hence more likely to be confused. The points are arranged with leaf angle varying along the horizontal axis and leaf width varying along the vertical axis. The cost function for the model $\mathcal{L}(x, \hat{x})$ is defined as the Euclidean distance between stimuli x and $\hat{x}$ in this space. The left panel shows the estimated perceptual space for the leaf-angle relevant condition, while the right panel shows the leaf-width condition. When leaf

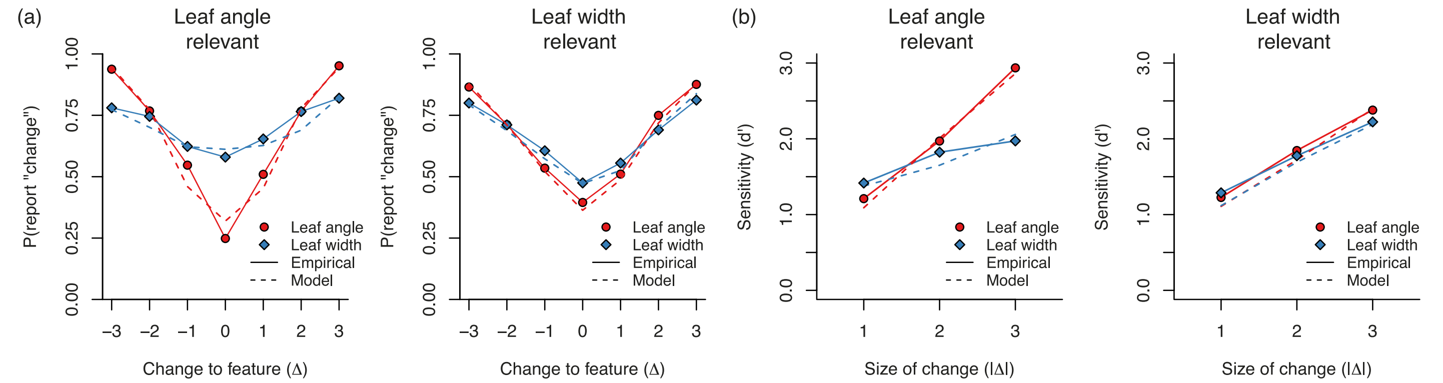

Figure 8. (a) Probability of reporting "change" as a function of the size of the change to each stimulus dimension. The *x*-axis indicates the size of the change in discrete stimulus units. Probabilities for a given feature dimension are obtained by averaging over all change magnitudes to the other dimension. (b) Corresponding average sensitivity (d') to changes in each stimulus dimension and in each condition. As predicted by our model, performance in the leaf-width relevant condition (right panel) is characterized by higher sensitivity to changes in leaf width, accompanied by reduced sensitivity to changes in leaf angle, compared to performance in the leaf-angle condition.

angle is task-relevant, distances along this dimension are expanded (by a factor of about 1.25×) relative to distances along the dimension of leaf width. When leaf width is the relevant dimension, perceptual distances along both dimensions are nearly equal (the perceptual space is more nearly square).

It seems that in both conditions leaf angle is the more salient or discriminable dimension. This could be a low-level psychophysical property of the plant stimuli. Alternatively, it may be that leaf angle is more meaningful semantically (any keeper of house plants will affirm that drooping leaves are an ominous sign). The more important point is that whatever a priori bias exists, it is substantially modified by task experience in an adaptive manner.[3]

Figure 7b shows how these differences in cost functions impact memory performance. Each panel shows a plot of the probability that the model reports "change" as a function of the size of the perturbation to leaf angle (x-axis) and leaf width (y-axis). Intuitively, large changes are very likely to be detected. More interesting is that the "just noticeable difference" in stimulus space differs between the two conditions (compare left and right panels). As hypothesized, smaller changes to leaf angle are more noticeable when subjects are trained to categorize plants based on their leaf angle. This adaptation is accompanied by a decrease in sensitivity to changes in leaf width when compared to the data from the leaf-width relevant condition.

Importantly, the difference between the left and right panels in Figure 7b is not due to differences in overall memory capacity. Rather, the model demonstrates that subjects possess a fixed memory capacity, and flexibly allocate memory resources to different feature dimensions in an adaptive manner.

To further validate our model, Figure 8a and b provide quantitative comparisons of human and model performances (solid vs. dashed lines). Figure 8a plots the probability that the model and human subjects report "change" as a function of the size of change made to a given feature dimension, averaging over all nonzero change magnitudes for the other feature dimension. As expected, changes in leaf angle are more likely to be reported when subjects are trained in the leaf-angle relevant condition (left panel). Figure 8b adopts a signal detection framework analysis (Macmillan & Creelman, 2004), and reports sensitivity (d') to each level of stimulus change. As predicted by rate-distortion theory, increases in sensitivity to changes in leaf angle were accompanied by decreases in sensitivity to changes in leaf width.

The analyses above were conducted based on group-level data obtained by aggregating responses from all experimental subjects. We also fit a simplified version of the model to the data from each individual subject. Since each subject completed only 44 change-detection trials, it was necessary to restrict the number of free parameters in the simplified model. The perceptual spaces illustrated in Figure 7a suggest that the primary effect of the category training was to scale the perceptual space in the task-relevant dimension. Hence, our simplified model assumed a cost function where distances along the leaf-angle dimension were evenly spaced, and distances along the leaf-width dimension were a scalar multiple r:1 of that distance, where the parameter r is estimated by the model. In addition, the channel capacity, and prior probability of change, p_{change}, were estimated from the data. The log of the aspect ratio parameter, $\log r$, was compared between the two conditions using a t test. A log transform was applied because ratios are linear on a logarithmic scale (i.e., following a log transformation, the difference between a ratio of 1/2:1 and 1:1 equals the difference between a ratio of 2:1 and 1:1). The results of this comparison indicate that the aspect ratio was significantly larger in the leaf-width relevant condition, $t(99) = 2.15$, $p = 0.034$. In other words, on a subject-by-subject basis, leaf widths were psychologically more distinct in the condition where subjects were trained to categorize based on leaf width. Estimates of subjects' capacity and p_{change} were highly similar to estimates based on the aggregated data, and did not differ between conditions (mean capacity $= 4.02$, 4.11 bits in the leaf-angle and leaf-width relevant conditions, respectively; mean $p_{\text{change}} = 0.24$, 0.22).

Discussion

Here, we have described our normative framework for the study of VWM based on rate-distortion theory, an extension of conventional Bayesian approaches that uses information theory to characterize rate-limited (or capacity-limited) processing. We demonstrated that VWM is adaptive in two ways predicted by our framework: It reallocates its capacity based on the statistical regularities of the to-be-remembered items and it reallocates its capacity based on the demands of the task. Importantly, our results are framed not in terms of ad hoc theories of cognitive processing or measures of memory capacity, but rather derive from foundational principles of information theory (Cover & Thomas, 1991; MacKay, 2003; Shannon & Weaver, 1949). Given a constraint on information processing, VWM seeks to minimize the cost of (inevitable) memory errors. As the costs of memory error are fundamentally task-specific, this requires VWM to be an adaptive information processing system.

An intuitive understanding of the results in Experiment 2 is that subjects adaptively altered the shape of their error distribution in memory (Figure 7b). We find an intriguing parallel between this result and the concept of an "uncontrolled manifold" in motor control (Scholz & Schöner, 1999; Todorov & Jordan, 2002), whereby people are able to reduce motor variability in task-relevant dimensions, at the expense of increasing motor variability along irrelevant dimensions. At the least, both examples demonstrate a system that adapts to an external task by adapting to and exploiting its own limitations.

The reader may note that the estimated capacities in Experiment 2 were larger than those in Experiment 1. If we only probe a subset of stimulus dimensions, however, we should expect to underestimate the true VWM capacity. For example, Experiment 1 only probed the information that subjects stored about leaf width, even though subjects likely also stored (at least partial) information about leaf angle, color, etc. Therefore, it should be expected that Experiment 2, which probed both leaf width and leaf angle, should be closer to the true capacity in its estimates. Since it is impossible to simultaneously probe all knowledge stored in a memory representation (including knowledge that is not relevant to the experimental task), it may be difficult to estimate the full capacity of VWM in a task-independent manner.

Previous work has demonstrated that VWM can adaptively reallocate its resources in a task-dependent manner that may be linked to visual attention (Bays, 2014; Bays, Gorgoraptis, Wee, Marshall, & Husain, 2011; Bays & Husain, 2008; Gorgoraptis, Catalao, Bays, & Husain, 2011; Melcher & Piazza, 2011). For example, Bays, Gorgoraptis, et al. (2011) found that attentional cuing of one item in a display led to better recall of that item and that this recall advantage was maintained only if the item was task-relevant. Bays and Husain (2008) reported data indicating that VWM resources can be reallocated based on selective attention and toward targets of upcoming eye movements.

Our data from Experiment 2 are consistent with these earlier works, suggesting a role for attention in accounting for our results. For instance, subjects trained in the leaf-angle relevant condition may have learned that leaf angle is the relevant feature for the experimental task, and that leaf width can be ignored. Of course, this interpretation is entirely consistent with the rate-distortion account provided here in which VWM allocates its capacity in a task-dependent manner. Both VWM and visual attention involve prioritization of information in the presence of competing signals, as well as sustaining relevant perceptual information across time. Although visual attention and VWM have traditionally been regarded as distinct cognitive processes, a growing body of work (Awh, Vogel, & Oh, 2006; Chun, 2011; Kiyonaga & Egner, 2013; Pashler, Johnston, & Ruthruff, 2001) in the past decade has revealed widespread overlap between these two systems. Pashler et al. (2001) argued that attentional mechanisms evolved out of necessity to leverage limited processing capacity to relevant information relating to ongoing behavioral goals.

Consequently, visual attention may not be distinct from the adaptive allocation of cognitive resources, and rate-distortion theory provides a plausible computational account of this process. Given that memory and attention are both rate-limited, future work could focus on a unified rate-distortion framework providing accounts of shared aspects of visual memory and visual attention. We speculate that the work presented here is an early step in this direction.[4]

Some research has assumed that VWM has a single pool of memory resources shared by all stimulus dimensions. In contrast, several investigators have recently considered the hypothesis that VWM has separate and independent pools of resources for different visual features, such as orientation and color (Bays, Wu, & Husain, 2011; Fougnie, Asplund, & Marois, 2010; Shin & Ma, 2017; Wheeler & Treisman, 2002). The experimental findings regarding this hypothesis are inconclusive. Consequently, researchers have begun considering hybrid accounts in which VWM has separate pools of resources for different features, but these pools may be linked and thus are not independent (Brady, Konkle, & Alvarez, 2011; Shin & Ma, 2017). In the work reported here, it did not seem plausible that VWM has independent capacities for leaf width and leaf angle. We therefore modeled VWM as having a single capacity for both features, and found that this modeling assumption was sufficient to explain our empirical data. Future work on VWM will need to better define which features share memory resources and which features have their own resource pools and capacities.

An emerging body of research on VWM has demonstrated that memory is sensitive to the statistics of visual information (Brady & Alvarez, 2011; Brady & Tenenbaum, 2013; Corbett, 2016; Huttenlocher et al., 2000; Orbán et al., 2008; Orhan & Jacobs, 2013; Sanocki et al., 2010; Victor & Conte, 2004), and some work has attempted to tie this phenomenon to fundamental principles of information theory (Brady et al., 2009; Sims et al., 2012; Victor & Conte, 2004). Other research has shown that the precision of VWM is task-specific (Fougnie et al., 2010; Sims, 2015; Swan et al., 2016), with greater memory precision for features that are task-relevant. A closely related line of research has also explored the influence of categories on perception and VWM (Bae et al., 2015; Goldstone, 1994; Huttenlocher et al., 2000; Lipinski, Simmering, Johnson, & Spencer, 2010; Nosofsky, 1987; Persaud & Hemmer, 2016). In particular, Goldstone (1994) described the phenomenon of acquired distinctiveness, or increased perceptual distinctiveness for items that reside in different categories. To explain this and related findings, Persaud and Hemmer (2016) proposed that prior knowledge in the form of categories and uncertain working memory representations may be combined through Bayesian inference. While conventional Bayesian inference provides a normative theory for reasoning about and acting on uncertainty, it is the subfield of rate-distortion theory that is specialized for providing a theory of processing when resources are limited, and of how resources can be allocated in a manner that approaches the theoretical bounds on efficiency defined by information theory. Our approach uniquely serves as a normative and principled computational framework for characterizing VWM as a limited but efficient processing system.

Although the application of rate-distortion theory is relatively recent in the context of VWM (Sims et al., 2012), the general idea that cognition is adapted to the properties of the task environment is not new. In defining bounded rationality, Herbert Simon wrote, "The human mind is an adaptive system. It chooses behaviors in the light of its goals, and as appropriate to the particular context in which it is working" (Simon, 1992). More recently, this concept has formed the basis of "computational rationality" (Gershman, Horvitz, & Tenenbaum, 2015), or the idea that an optimal cognitive system is one that is optimized with respect to both the external task, as well as its own processing limitations. Our analysis of VWM falls neatly within this general approach.

Lastly, our results have implications for understanding the nature of perceptual expertise. Prior research has shown that perceptual experts (such as individuals who are experts at recognizing cars or birds) demonstrate a range of behavioral (Bukach, Gauthier, & Tarr, 2006; Curby & Gauthier, 2010) as well as neural differences (Herzmann & Curran, 2011). Rate-distortion theory provides a new vocabulary for understanding these differences. In particular, it suggests that an expert might have superior VWM performance compared to a novice for any combination of three reasons: greater channel capacity, greater knowledge of statistical regularities, or a cost function that is more finely tuned to the demands of a task. The current results provide positive evidence for the latter two. However, the present study is limited in terms of working with participants who received only minimal training. The application of this same approach to a truly expert population represents a fruitful area of investigation.

Keywords: visual working memory, visual learning, ideal observer

Acknowledgments

CJB was supported by an National Science Foundation (NSF) National Science Foundation Research Traineeship (NRT) graduate training grant (NRT-1449828) and an NSF Graduate Research Fellowship (DGE-1419118). This work was also supported by NSF research grants (DRL-1560829 and DRL-1561335).

Commercial relationships: none.
Corresponding author: Christopher J. Bates.
Email: cjbates@ur.rochester.edu.
Address: Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY, USA.

Footnotes

[1] Here, “bit allocation” specifically refers to changes in the pattern of memory errors resulting from adapting VWM to the current task. How someone decides to “allocate” their capacity is synonymous with how they decide to distribute their errors, given what they have observed and given limits on their ability to store visual information with perfect fidelity. These decisions are crucial for maximizing performance. A mathematical description of this process is given in the following sections, and in particular the error distribution for memory is given by the channel, $p(\hat{x}|x)$.

[2] The rate-distortion model was implemented using the “RateDistortion” library (Sims, 2016) of the R statistical computing environment. This library contains algorithms for efficiently solving the constrained optimization problem described by Equation 1. A tutorial introduction to this library and its use in modeling human perception is described in Sims (2016). Model parameters were obtained by maximum likelihood estimation, using L-BFGS or Nelder-Mead optimization implemented within R (via “optim”). Complete code for the model is available from the third author’s website.

[3] A reader may note that, in each condition, subjects remembered aspects of the task-irrelevant stimulus dimension, contrary to the predictions of rate-distortion theory. It should be kept in mind that subjects received only a small amount of training (approximately 20–30 min). With additional training, we would expect subjects to show additional adaptation.

[4] It is not our intention to claim that visual attention and visual memory are one and the same. Rather, we believe that the terms “attention” and “memory” are vaguely defined in the cognitive science literature, each potentially covering many phenomena and mechanisms. Because the terms are so vague, they overlap. For instance, the results of many experiments can be accounted for by stating that subjects “allocated more attentional capacity to stimulus dimension *A* than *B*” or, equally validly, by stating that subjects “allocated more VWM capacity to stimulus dimension *A* than *B*.” Which description is used in an article often depends on the personal biases of the article’s authors. Additional work (Chun, 2011) has addressed this issue by working towards an organizing taxonomic framework for differentiating between these two mechanisms.

References

Awh, E., Vogel, E., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201–208.

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, *34*, 3632–3645.

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, *11*(10):6, 1–15, https://doi.org/10.1167/11.10.6. [PubMed] [Article]

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., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*, 1622–1631.

Berger, T. (1971). Rate distortion theory: A mathematical basis for data compression. Englewood Cliffs, NJ: Prentice-Hall.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, *18*(4), 460–473.

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

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502.

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

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences, USA*, *113*, 7459–7464.

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

Bukach, C. M., Gauthier, I., & Tarr, M. J. (2006). Beyond faces and modularity: The power of an expertise framework. *Trends in Cognitive Sciences*, *10*(4), 159–166.

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, *49*(6), 1407–1409.

Corbett, J. E. (2016). The whole warps the sum of its parts gestalt-defined-group mean size biases memory for individual objects. *Psychological Science*, *28*(1), 12–22. Los Angeles, CA: Sage Publications.

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

Curby, K. M., & Gauthier, I. (2010). To the trained eye: Perceptual expertise alters visual processing. *Topics in Cognitive Science*, *2*(2), 189–201.

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General*, *143*, 548–565.

Faraway, J. J. (2016). *Extending the linear model with r*. Boca Raton, FL: CRC Press.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504.

Fiser, J., & Aslin, R. N. (2002a). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458.

Fiser, J., & Aslin, R. N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822–15826.

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

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. New York, NY: Cambridge University Press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015, July 17). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gersho, A., & Gray, R. M. (1992). *Vector quantization and signal compression*. Norwell, MA: Kluwer Academic Publishers.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178.

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience*, *31*, 8502–8511.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*, *48*, 829–842.

Herzmann, G., & Curran, T. (2011). Experts memory: An ERP study of perceptual expertise effects on encoding and recognition. *Memory & Cognition*, *39*(3), 412–432.

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

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology*, *9*(2), e1002927.

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic Bulletin & Review*, *20*(2), 228–242.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in r*. New York, NY: Springer.

Lipinski, J., Simmering, V. R., Johnson, J. S., & Spencer, J. P. (2010). The role of experience in location estimation: Target distributions shift location memory biases. *Cognition*, *115*(1), 147–153.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. New York, NY: Psychology Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.

Melcher, D., & Piazza, M. (2011). The role of attentional priority and saliency in determining capacity limits in enumeration and visual working memory. *PLoS One*, *6*(12), e29296.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.

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

Moscatelli, A., Mezzetti, M., & Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: The generalized linear mixed model. *Journal of Vision*, *12*(11):26, 1–17, https://doi.org/10.1167/12.11.26. [PubMed] [Article]

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87–108.

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*, *105*(7), 2745–2750.

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

Orhan, A. E., & Jacobs, R. A. (2014). Toward ecologically realistic theories in visual shortterm memory research. *Attention, Perception, and Psychophysics*, *76*, 2158–2170.

Pashler, H., Johnston, J. C., & Ruthruff, E. (2001). Attention and performance. *Annual Review of Psychology*, *52*(1), 629–651.

Persaud, K., & Hemmer, P. (2016). The dynamics of fidelity over the time course of long-term memory. *Cognitive Psychology*, *88*, 1–21.

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization*. Cambridge, UK: Cambridge University Press.

R Core Team. (2015). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, *8*, 1647–1650.

Sanocki, T., Sellers, E., Mittelstadt, J., & Sulman, N. (2010). How high is visual short-term memory capacity for object layout? *Attention, Perception, & Psychophysics*, *72*(4), 1097–1109.

Scholz, J. P., & Schöner, G. (1999). The uncontrolled manifold concept: Identifying control variables for a functional task. *Experimental Brain Research*, *126*(3), 289–306.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Shepard, R. N. (1987, September 11). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shin, H., & Ma, W. J. (2017). Visual short-term memory for oriented, colored objects. *Journal of Vision*, *17*(12):9, 1–19, https://doi.org/10.1167/17.12.9. [PubMed] [Article]

Simon, H. A. (1992). What is an explanation of behavior? *Psychological Science*, *3*(3), 150–161.

Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision*, *15*(3):2, 1–27, https://doi.org/10.1167/15.3.2. [PubMed] [Article]

Sims, C. R. (2016). Rate-distortion theory and human perception. *Cognition*, *152*, 181–198.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–830.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Swan, G., Collins, J., & Wyble, B. (2016). Memory for a single object has differently variable precisions for relevant and irrelevant features. *Journal of Vision*, *16*(3):32, 1–12, https://doi.org/10.1167/16.3.32. [PubMed] [Article]

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, *5*(11), 1226–1235.

Victor, J. D., & Conte, M. M. (2004). Visual working memory for image statistics. *Vision Research*, *44*(6), 541–556.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*, 101–118.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*(1), 48–64.

Appendix A

Comparison of a Bayesian ideal observer model to rate-distortion theory

This appendix derives the contrasting predictions of a Bayesian ideal observer and a model based on rate-distortion theory. The Bayesian observer assumes that memory representations are Gaussian-noise corrupted versions of afferent sensory signals. The rate-distortion model is shown to exhibit a similar mathematical form with one critical exception: For an optimal, but capacity-limited observer, the magnitude of encoding noise scales with the width of the stimulus distribution. As a result, the two models generate qualitatively different predictions for performance changes as sensory statistics are manipulated, a scenario which forms the basis of Experiment 1.

Consider a continuous univariate sensory signal, x, sampled from a Gaussian distribution with arbitrary mean and variance such that $x \sim \text{Normal}(\mu, \sigma^2)$. An abstract visual memory task is modeled in which an observer is shown a sample from this distribution and tasked with remembering it as accurately as possible. The task is nontrivial because the observer's memory (and hence response) is related only probabilistically to the sensory signal according to a conditional probability distribution, $p(\hat{x}|x)$, where the details of this distribution are defined according to one of two different models of visual memory performance. The different models considered in this appendix are:

- A Bayesian observer with fixed Gaussian "memory noise," and
- An optimal, but capacity-limited observer based on rate-distortion theory.

For the purposes of this analysis, it is assumed that the observer seeks to produce an estimate $\hat{x}$ that minimizes the squared error between the sensory signal and its reported value. The use of a squared error objective is motivated by its mathematical convenience, but none of the essential features of the analysis depend upon this particular assumption. This objective is the *loss function* for the observer,

$$\mathcal{L}(x, \hat{x}) = (\hat{x} - x)^2. \quad (3)$$

Using this loss function, the performance of the observer is defined as the mean squared error:

$$\begin{aligned} D &= \mathrm{E}_{x,\hat{x}}[\mathcal{L}(x,\hat{x})] = \mathrm{E}_{x,\hat{x}}\left[(\hat{x}-x)^2\right] \\ &= \iint_{-\infty}^{\infty} (\hat{x}-x)^2 p(\hat{x}|x)p(x)\,\mathrm{d}\hat{x}\,\mathrm{d}x. \quad (4) \end{aligned}$$

This quantity also corresponds to the observer's expected variance.

In the following two subsections, we derive the optimal observer according to each model. Of particular interest is the relationship between the mean squared error of the observer, D, and the variance of the prior over the signal distribution, σ^2.

Bayesian observer model

We first consider an observer model in which the sensory signal is corrupted by zero-mean additive Gaussian memory noise during the memory encoding process, resulting in a noisy memory representation x_e. The memory representation is related to the sensory signal according to:

$$p(x_e|x) = \text{Normal}(x, \sigma_e^2) = \frac{e^{-\frac{(x_e-x)^2}{2\sigma_e^2}}}{\sqrt{2\pi}\sigma_e}. \quad (5)$$

The Bayesian posterior probability distribution over the signal x, given the noisy memory representation x_e is given by:

$$p(x|x_e) = \frac{p(x_e|x)p(x)}{\int_{-\infty}^{\infty} p(x_e|x)p(x)\,\mathrm{d}x}. \quad (6)$$

For a given noisy memory representation x_e, a Bayes-optimal observer should produce an estimate $\hat{x}$ of the true signal x, which minimizes the expected loss. For the squared error loss function, this expected loss as a function of the estimate $\hat{x}$ is:

$$\text{Loss}(\hat{x}|x_e) = \int_{-\infty}^{\infty} (\hat{x}-x)^2 p(x|x_e)\,\mathrm{d}x \quad (7)$$

This loss function reaches a minimum at the point where its first derivative is zero. Solving for the minimum mean squared error estimator, one obtains:

$$\hat{x} = \frac{x_e\sigma^2 + \mu\sigma_e^2}{\sigma^2 + \sigma_e^2}. \quad (8)$$

Note that this is the optimal estimator as a function of the latent quantity x_e. The corresponding distribution in terms of the observed sensory signal x, or $p(\hat{x}|x)$, is given by:

$$\begin{aligned} p(\hat{x}|x) &= \int_{-\infty}^{\infty} \delta\left(\hat{x} - \frac{x_e\sigma^2 + \mu\sigma_e^2}{\sigma^2+\sigma_e^2}\right) p(x_e|x)\,\mathrm{d}x_e \\ &= \frac{e^{-\frac{\left(x\sigma^2+\mu\sigma_e^2-\hat{x}(\sigma^2+\sigma_e^2)\right)^2}{2\sigma^4\sigma_e^2}}\left(\sigma^2+\sigma_e^2\right)}{\sqrt{2\pi}\sigma^2\sigma_e}, \quad (9) \end{aligned}$$

where $\delta(\cdot)$ indicates the Dirac delta function. Lastly, the expected cost for this observer is obtained by evaluating the expression in Equation 4, with the distribution $p(\hat{x}|x)$ defined by Equation 9. One obtains for the ideal Bayesian observer,

$$D_{\text{Bayes}} = \mathrm{E}_{x,\hat{x}}[\mathcal{L}(x,\hat{x})] = \frac{\sigma^2\sigma_e^2}{\sigma^2+\sigma_e^2}. \quad (10)$$

Note in particular that when the magnitude of encoding noise is small relative to the signal prior ($\sigma_e < \frac{1}{10}\sigma$), performance of the observer is essentially flat with respect to changes in the variance of the signal distribution. This is illustrated in Figure 9a.

Rate-distortion observer model

We now consider an observer with a constraint placed on the information rate of the perceptual

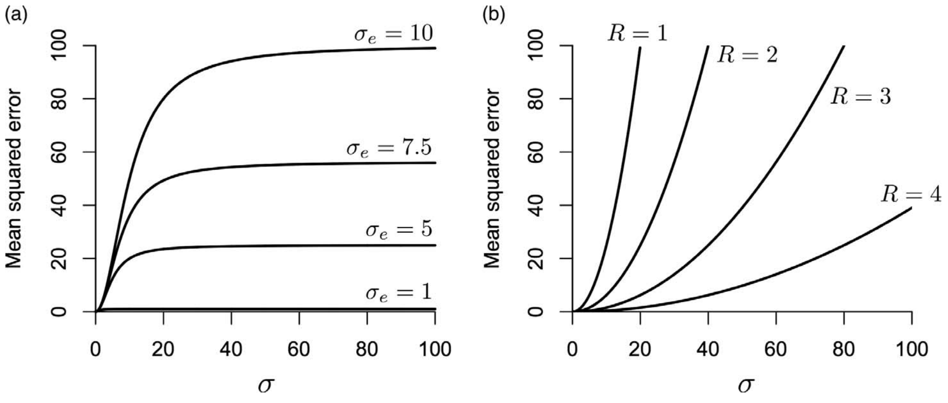

Figure 9. Expected cost (mean squared error) for a Bayesian ideal observer (left) and a model based on rate-distortion theory (right), as a function of the standard deviation of the prior distribution of sensory signals, σ. (a) Performance of the Bayesian observer model. Performance is shown for four levels of the encoding noise, σ_e. (b) Performance of the rate-distortion model. Performance is shown for four levels of the information rate of the channel, R, measured in bits.

channel. The information rate of a channel is defined in terms of the mutual information, $\mathrm{I}(x, \hat{x})$, of the channel's input and output:

$$\mathrm{I}(x, \hat{x}) = \iint_{-\infty}^{\infty} p(\hat{x}|x)p(x) \log\left(\frac{p(\hat{x}|x)p(x)}{p(\hat{x})p(x)}\right) \mathrm{d}x\mathrm{d}\hat{x}. \quad (11)$$

A finite bound is placed on the information rate of the channel, such that $\mathrm{I}(x, \hat{x}) \leq R$ for some finite positive value R. Subject to this bound, the goal of the observer is to minimize expected cost, defined as above in terms of the mean squared error. The optimal channel, $p^{\star}(\hat{x}|x)$ according to this criterion satisfies

$$\begin{aligned} &p^{\star}(\hat{x}|x) = \inf_{p} \mathrm{E}_{x,\hat{x}}\left[(\hat{x} - x)^2\right] \\ &\text{subject to } \mathrm{I}(x, \hat{x}) \leq R. \end{aligned} \quad (12)$$

Equation 12 is a fundamental problem in the field of rate-distortion theory. In general, this equation cannot be solved analytically. However, for the special case considered here (a Gaussian signal prior, and squared error loss function), a closed-form solution can be obtained. For this specific problem, one can derive the so-called rate-distortion curve as the absolute minimum information rate (R) necessary to achieve a specified level of performance (D):

$$R(D) = \frac{1}{2} \log \frac{\sigma^2}{D} \quad (13)$$

It turns out that the optimal channel that minimizes the expected cost and achieves the above rate-distortion bound has the same form as the Bayesian observer model described in the previous section, with the critical difference that the magnitude of the Gaussian encoding noise is no longer a fixed constant, but is rather a function of the variance of the signal distribution.

The proof of this result proceeds as follows. We first assume that the claim is true—namely, that for a Gaussian sensory signal and squared error loss function, the optimal channel in the rate-distortion sense can be defined in terms of Gaussian encoding noise followed by an optimal Bayesian decoder. We then derive the conditions under which this channel achieves the constraint on information rate given in Equation 12, and show that these conditions are achievable. It is then shown that the channel saturates the rate-distortion bound—meaning no other channel could achieve lower expected cost at the same information rate. This approach builds upon the results already derived in the previous section.

To begin, consider the channel distribution in Equation 9 above. The marginal distribution for this channel is

$$p(\hat{x}) = \int_{-\infty}^{\infty} p(\hat{x}|x)p(x)\,\mathrm{d}x \quad (14)$$

$$= \frac{\mathrm{e}^{-\frac{(\hat{x}-\mu)^2(\sigma^2+\sigma_e^2)}{2\sigma^4}}\sqrt{\sigma^2+\sigma_e^2}}{\sqrt{2\pi}\sigma^2}. \quad (15)$$

With $p(\hat{x}|x)$ and $p(\hat{x})$ defined, the mutual information of the channel determined according to Equation 11 is

$$\mathrm{I}(x, \hat{x}) = \frac{1}{2} \log\left(1 + \frac{\sigma^2}{\sigma_e^2}\right). \quad (16)$$

In order to satisfy the constraint on information rate, it must be the case that $I(x, \hat{x}) \leq R$. This can be achieved by setting

$$\sigma_e = \frac{\sigma}{\sqrt{-1 + e^{2R}}}. \quad (17)$$

Note that according to the model, the encoding noise of the channel depends on both the constraint on information rate, as well as the distribution of the sensory signals (via its dependence on σ). This equation defines the "effective encoding noise" for a channel with information rate R.

Hence, the channel described is achievable by choosing the encoding noise to satisfy the above. It remains to be shown that it is also optimal in the rate-distortion sense. To do so, note that the effective encoding noise can be used to re-parameterize the expected cost for a Bayesian observer (Equation 10, above) in terms of its equivalent information rate:

$$D_{\mathrm{RD}} = \mathrm{E}_{x,\hat{x}}[\mathcal{L}(x,\hat{x})] = \frac{\sigma^2\sigma_e^2}{\sigma^2+\sigma_e^2} = \mathrm{e}^{-2R}\sigma^2. \quad (18)$$

The above equation defines the minimum expected loss for the channel (the channel "distortion") in terms of its information rate. One can also solve this equation for R and obtain the information rate required for the channel to reach a specified level of distortion D:

$$R = \frac{1}{2}\log\frac{\sigma^2}{D_{RD}}. \quad (19)$$

Hence, the channel achieves the rate-distortion bound exactly, and the proof is concluded.

Figure 9b illustrates the performance of the rate-distortion model as the standard deviation of the sensory signal is varied.

Comparison of a Bayesian observer and rate-distortion theory

The results in Figure 9 illustrate a fundamental qualitative difference between a Bayesian observer model and a model based on rate-distortion theory. For the Bayesian model, performance is asymptotically independent of the statistics of the sensory signal. This performance is limited primarily by the magnitude of internal encoding noise. For the rate-distortion model, the magnitude of encoding noise intrinsically depends on the statistics of the sensory signal. This is not an arbitrary assumption of this particular model, but rather reflects the predicted behavior of an optimally efficient communication channel.

To more directly compare the approaches, one can calibrate the rate-distortion model to perform identically to the Bayesian observer model at a particular level of encoding noise and width of the stimulus prior, and compare the performances of the two models as the prior is varied. This comparison is illustrated in Figure 2. (To generate these data, the capacity of the rate-distortion model was set to 2.2 bits, close to the value estimated from subjects' data in Experiment 1.) The red curve in this figure gives the performance of the rate-distortion model, whereas the black curve gives the performance of the Bayesian observer model. The Bayesian model was calibrated to produce identical performance as the rate-distortion model at $\sigma = 10$ (the value used in Experiment 1). This was achieved by ensuring that both models have the same effective encoding noise. As can be seen in the figure, increases to the standard deviation of the stimulus distribution lead to divergent predictions between the two models. For the rate-distortion model, memory performance declines monotonically with increases to the width of the stimulus distribution. In contrast, the Bayesian observer's performance is relatively invariant to such changes.

Notably, the empirical data from Experiment 1 are qualitatively consistent with the rate-distortion model, and inconsistent with the Bayesian observer model that assumes memory encoding noise that is independent of the stimulus distribution. Recall that subjects in the uniform condition (where the stimulus distribution has a large variance) of Experiment 1 showed significantly worse memory performance compared to the performances of the subjects in the other conditions (where the stimulus distributions had smaller variance). This decline in memory performance is directly predicted by rate-distortion theory, but at odds with a standard Bayesian account of visual memory. Hence, the data are more parsimoniously explained by a capacity-limited, but efficient information communication channel.

Comparison of Bayesian and rate-distortion modeling results

Recall that we used maximum likelihood estimation to estimate the values of the rate-distortion model's parameters based on subjects' responses (Table 1). To further compare the rate-distortion and Bayesian observer models, we also used maximum likelihood estimation to fit the Bayesian model to the experimental data. Two versions of the Bayesian model were considered, one in which σ_e and p_{change} were shared across experimental conditions, and one in which only p_{change} was shared across conditions. The decision rule and choice for $p(y|x, C)$ for the Bayesian versions were identical to those of the rate-distortion model. We predicted that (a) when σ_e is shared across conditions, the likelihood should be lower than that of the rate-distortion model; (b) when σ_e is shared across conditions, the model should predict only a small decrement in overall percent correct responses in the uniform condition compared to the Gaussian conditions; and (c) when σ_e is allowed to vary by condition, it should be highest in the uniform condition.

All of our predictions are born out by the results. Tables 3 and 4 report maximum likelihood estimates

	Mean 30	Mean 50	Mean 75	Uniform
Subjects	0.84	0.81	0.78	0.74
Rate-distortion	0.82	0.82	0.78	0.74
Bayesian (σ_e local)	0.83	0.82	0.78	0.74
Bayesian (σ_e global)	0.79	0.80	0.78	0.78

Table 2. Proportion-correct responses predicted by each model, compared to subjects. *Notes*: The Bayesian model with fixed memory noise fails to predict subjects' worse performance in the uniform condition.

for the two versions of the Bayesian model. The log-likelihood of the Bayesian model that shared memory noise across conditions was −14,901, while that of the rate-distortion model was −14,799. When the memory noise was allowed to vary, its variance was higher in the uniform condition than any of the normal conditions. Furthermore, within the normal conditions, memory noise increased in accordance with leaf width, which is consistent with the finding in Experiment 1 that subjects performed better on trials with skinnier leaves. Note that the standard deviation of the memory noise when it was shared (4.21) is very close to the average of its values when it was varied across conditions (4.15), and that these values are close to what is predicted by Equation 17 (assuming $\sigma = 20$ and $R = 2.2$ bits, then $\sigma_e = 4.46$). Also note that the value for p_{change} in both versions is very close to that of the rate-distortion model.

As expected, we found that the Bayesian model with global σ_e predicted only a small decrement in the overall proportion of correct responses in the uniform condition, whereas both the rate-distortion model and the other version of the Bayesian model better matched the subjects' proportion-correct scores. For each of these three models, we simulated responses by sampling from $p(C|x,y)$. These results illustrate that the Bayesian model fails to account for subjects' worse performance in the uniform condition (see Table 2).

We also found some of the estimates of μ and σ in some conditions of the Bayesian models to be less intuitive than their counterparts in the rate-distortion model. For example, in both versions of the model, the estimate of σ was large in the mean-75 condition, and the estimate of μ was low, relative to what we found in the rate-distortion model.

	Mean 30	Mean 50	Mean 75	Uniform	Global
μ	30.7	49.5	16.4	12.2	
σ	16.0	16.2	45.9	35.2	
σ_e					4.21
p_{change}					0.40

Table 3. Maximum likelihood estimates of the Bayesian model's parameters in each of the four experimental conditions of Experiment 1 when both the memory noise, σ_e, and p_{change} are shared across conditions. *Notes*: The last column (Global) gives values of parameters that were shared across conditions.

	Mean 30	Mean 50	Mean 75	Uniform	Global
μ	38.7	48.4	36.8	24.2	
σ	37.6	21.4	38.3	35.4	
σ_e	3.63	3.86	4.18	4.93	
p_{change}					0.40

Table 4. Maximum likelihood estimates of the Bayesian model's parameters in each of the four experimental conditions of Experiment 1 when the memory noise, σ_e, is not shared but rather is allowed to vary across conditions.

Overall, we conclude that the rate-distortion model provides a better account of the experimental data than either version of the Bayesian model. When matched for equal numbers of parameters, the Bayesian model has lower likelihood than the rate-distortion model. And when the memory noise is allowed to vary by condition (which is theoretically less parsimonious), the memory noise varied across conditions in exactly the way predicted by our preceding theoretical analysis. These results suggest capacity limits are an important factor in accounting for people's visual memory performances.

Appendix B

Derivation of the Bayesian decision rule

Figure 5 provides a diagram of the complete information-theoretic model, consisting of a capacity limited channel (VWM) and subsequent Bayesian decision rule. In this appendix we derive the optimal Bayesian decision rule for determining whether a "change" trial has occurred. We define the following random variables: C is a binary random variable indicating whether or not the current trial is a change trial. The variables x and $\hat{x}$ indicate the memory stimulus and its noisy memory representation, respectively. The probe stimulus presented to subjects is indicated by y. A probabilistic graphical model relating these variables is shown in Figure 10.

Based on this graphical model, the joint probability distribution over all variables is given by

$$p(\hat{x}, x, y, C) = p(\hat{x}|x)\, p(y|x, C)\, p(x)\, p(C). \quad (20)$$

From this, the conditional probability of a change trial given the noisy memory representation and the probe stimulus is derived as follows:

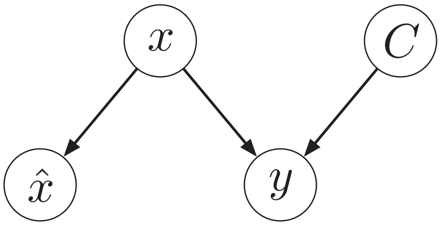

Figure 10. Probabilistic graphical model illustrating the relationship between random variables in the Bayesian change detection decision rule.

$$p(\hat{x}, y, C) = \sum_x p(\hat{x}|x)\, p(y|x, C)\, p(x)\, p(C) \quad (21)$$

$$p(C|\hat{x}, y) = \frac{p(\hat{x}, y, C)}{p(\hat{x}, y)} \quad (22)$$

$$= \frac{p(\hat{x}, y, C)}{\sum_C p(\hat{x}, y, C)} \quad (23)$$

$$= \frac{\sum_x p(\hat{x}|x)\, p(y|x, C)\, p(x)\, p(C)}{\sum_C \sum_x p(\hat{x}|x)\, p(y|x, C)\, p(x)\, p(C)}. \quad (24)$$

To complete the derivation we need to specify the form for each term in Equation 24. As noted in the main text, the prior probability of a change trial, $p(C = 1)$, is treated as a parameter in the model, p_{change}. The distribution of the probe given the memory stimulus and change trial status is:

$$p(y|x, C) = \begin{cases} 1 & (x = y)\ \&\ (C = 0) \\ 0 & (x \neq y)\ \&\ (C = 0) \\ 0 & (x = y)\ \&\ (C = 1) \\ f(x) & (x \neq y)\ \&\ (C = 1) \end{cases}. \quad (25)$$

To produce the values reported in the main text, $f(x)$ was set to the true (experiment-defined) probability of a probe given a target. (We found qualitatively similar results for several other choices of $f[x]$.) The prior distribution over stimuli, $p(x)$, was modeled as a normal distribution in Experiment 1, with parameters μ and σ, normalized over the space of possible stimulus values. In Experiment 2, the prior distribution was uniform, $p(x) = 1/N$. Lastly, the conditional distribution over memory representations, $p(\hat{x}|x)$, was obtained by solving the rate-distortion equation given in the main text (Equation 1). Consequently, the Bayesian decision rule is optimal with respect to the structure of noise and variability in VWM. This equation was solved using the "RateDistortion" package available for the R statistical programming environment, and described in Sims (2016).

Equation 24 defines the probability that a "change" trial has occurred given a noisy memory representation and probe stimulus. The model assumes that subjects exhibit probability matching; hence, this equation also describes the likelihood of a "change" response for a given memory representation and probe stimulus. To fit this model to the data it is necessary to marginalize over the distribution of memory representations for a given memory stimulus (since internal memory representations are not directly observable). Hence,

$$p(C|x, y) = \sum_{\hat{x}} p(C|\hat{x}, y) p(\hat{x}|x). \quad (26)$$

Equation 26 defines the likelihood function for the data. Model parameters were determined by maximum likelihood estimation using this equation.