

An Ideal Observer Model of Visual Short-Term Memory Predicts Human Capacity–Precision Tradeoffs

Chris R. Sims Robert A. Jacobs David C. Knill
(csims@cvs.rochester.edu) (robbie@bcs.rochester.edu) (knill@cvs.rochester.edu)
Department of Brain and Cognitive Sciences & Center for Visual Science
University of Rochester, Rochester, NY

Abstract

We develop an ideal observer model of human visual short-term memory. Compared to previous models that have posited constraints on memory performance intended solely to account for observed phenomena, in the present research we derive the expected behavior of an optimally performing, but limited-capacity memory system. We develop our model using rate–distortion theory, a branch of information theory that provides optimal bounds on the accuracy of information transmission subject to a fixed capacity. The resulting model provides a task-independent and theoretically motivated definition of visual memory capacity and yields novel predictions regarding human performance. These predictions are quantitatively evaluated in an empirical study. We also demonstrate that our ideal observer model encompasses two existing, but competing accounts of VSTM as special cases.

Keywords: Ideal observer analysis, VSTM, information theory, rate–distortion theory

Introduction

Visual short-term memory (VSTM)—defined as the ability to store task-relevant visual information in a rapidly accessible and easily manipulated form—is a central component of nearly all human activities. Given its importance, it is perhaps surprising that the capacity of this system is severely limited. Numerous investigations of VSTM performance have revealed that we can only accurately store a surprisingly small number of visual objects or features (for a review, see Luck, 2008).

In recent years there have been numerous attempts to define and quantify what is meant by VSTM capacity. Until recently, the predominant view has held that capacity is limited to a small, fixed number of visual objects (typically assumed to be 4) stored in discrete “slots” (Awh, Barton, & Vogel, 2007; Luck & Vogel, 1997; Vogel, Woodman, & Luck, 2001; Cowan & Rouder, 2009). Taking a different approach, Bays and colleagues (Bays, Catalao, & Husain, 2009; Bays & Husain, 2008) explored how the *precision* of features encoded in visual working memory may change as a function of the number of features that are concurrently stored. Based on the finding that memory precision appears to decrease even when as few as 2 items are encoded, the authors proposed that VSTM capacity consists of a single, continuous resource that must be divided among items stored in working memory.

While both the discrete slot and continuous resource models are able to account for a number of empirical findings, both are ultimately unsatisfactory as complete theories of human VSTM. First, in the case of both models the nature of the capacity limit is somewhat arbitrary: the hypothesized capacity limit does not emerge from a principled theoretical basis,

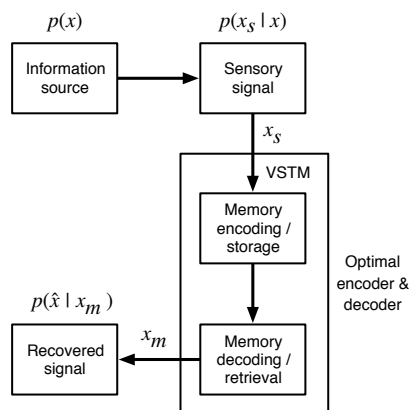


Figure 1: A schematic diagram of the ideal observer model of VSTM.

but rather only serves to account for observed empirical phenomena. Second, the definition of capacity appears largely task-dependent, and it is therefore difficult to form predictions for human performance in different tasks or conditions. In the case of the discrete slot model, there is no strong theoretical justification for determining what visual features or objects can or cannot occupy a single slot, and in the continuous resource model, the nature of the resource that is being divided is only abstractly specified.

The Ideal Observer Model of VSTM

In this section we derive an ideal observer model of visual short-term memory. We show that from an information-theoretic perspective, our ideal observer model is optimally efficient in that it minimizes a particular measure of memory distortion subject to a fixed capacity limit.

The resulting model makes several important contributions to the literature on VSTM. First, whereas previous models have postulated abstract or task-dependent definitions of visual memory capacity, the ideal observer model provides a quantitative definition of capacity that is task-independent and easily interpreted. This enables results obtained in one experiment to generate predictions for performance in another. Second, since our model exhibits provably optimal performance subject to a fixed capacity, it can be used to obtain an assumption-free estimate of the minimum capacity of human VSTM. Finally, we demonstrate that our ideal observer model subsumes two existing models of VSTM: a recent version of the discrete slot model (Cowan & Rouder, 2009), and the continuous resource model (Bays & Husain, 2008).

A schematic illustration of our model is given in Figure 1. The starting point for our analysis is the recognition that human memory can be viewed as a physical communications channel for storing and transmitting information. We assume that there is some information source in the world, labeled x . This signal represents a physical feature of the world—for example, the gaze-relative location of an object that we intend to reach and grasp—and in general, this signal follows some probability distribution given by $p(x)$. This distribution has some mean and variance (e.g., the variance of target locations in a particular environment). This signal is encoded by the human sensory system, resulting in a transformed signal x_s , which can be described by the conditional distribution $p(x_s | x)$. This sensory signal must be stored, and then retrieved from memory. If VSTM is viewed as a communication channel, then the input to this channel is the sensory signal x_s , and the output is the retrieved ‘memory’ signal x_m . The conditional distribution $p(x_m | x_s)$ is referred to as the *channel distribution*. Given the output of this channel, the agent forms an estimate \hat{x} of the original signal in the environment.

If we view VSTM as a physical communications channel, then there are two important properties of this system that are relevant for characterizing its performance. The first is the *distortion* of the ideal observer model. Intuitively, if the channel input x_s and output x_m are on average, very similar to each other, then VSTM has a low distortion. In practice, a common distortion measure is the squared-error difference between the channel input and output:

$$\begin{aligned} D &= \text{E}[(x_m - x_s)^2] \\ &= \iint (x_m - x_s)^2 p(x_m | x_s) p(x_s) dx_m dx_s. \end{aligned} \quad (1)$$

The second property that we derive in this section is the *information rate* of the system, defined as the average quantity of information, measured in bits, that can be stored or transmitted through the channel. We assume that VSTM is capacity-limited, but is an otherwise optimal system. If the input to VSTM is the probabilistic sensory signal given by $p(x_s)$ and the channel distribution is given by $p(x_m | x_s)$, then the rate R is defined mathematically as the average mutual information of x_m and x_s :

$$R = \iint p(x_s) p(x_m | x_s) \log \left[\frac{p(x_m | x_s)}{p(x_m)} \right] dx_m dx_s. \quad (2)$$

If the natural logarithm is used in (2), the information rate is measured in *nats*, or the natural logarithm equivalent of binary bits. In this paper all quantities are reported in terms of the more familiar quantity bits. These two quantities, information rate and distortion, are linked by an important equation known (not surprisingly) as a rate–distortion equation. For a given system with a known information rate, we might wish to know the best-possible performance obtainable by this system: this corresponds to finding the minimum possible distortion subject to a fixed memory capacity. Alternatively, we may observe the performance of a system, and wish to infer

the minimum capacity that the system must necessarily possess to achieve this level of performance. Previously derived results in information theory (Berger, 1971) have shown that if the input to a communications channel follows a Gaussian distribution with arbitrary mean and variance given by σ^2 , then the provably optimal rate–distortion bound is given by

$$R(D) = \frac{1}{2} \max \left(0, \log \frac{\sigma^2}{D} \right). \quad (3)$$

That is to say, for a prescribed level of memory error D , there cannot exist any physical system that achieves this performance using fewer than $R(D)$ bits on average. Similarly, for a fixed capacity R , this equation can be used to derive the minimum distortion achievable by any physical system.

In order to apply the model to human data, it is necessary to specify parametric forms for the various distributions given in Figure 1. We define the information source in the world, $p(x)$, to be Gaussian with mean μ and variance σ_w^2 . We assume that the distribution of incoming sensory signals $p(x_s)$ also follows a Gaussian distribution, with mean μ and variance $\sigma_w^2 + \sigma_s^2$ (the combined variance of the information source and additive Gaussian sensory noise). We do not choose the channel distribution $p(x_m | x_s)$ arbitrarily, but rather define the model so that it achieves the theoretical bounds given by rate–distortion theory. In particular, optimal performance is achieved by designing the channel distribution to be Gaussian with mean and variance given by

$$\begin{aligned} p(x_m | x_s) &= \text{Normal}(\mu_m, \sigma_m^2) \\ \mu_m &= x_s + e^{-2R}(\mu - x_s), \\ \sigma_m^2 &= e^{-4R}(e^{2R} - 1)(\sigma_w^2 + \sigma_s^2). \end{aligned} \quad (4)$$

Computing the distortion of this channel using (1), one obtains the theoretical rate–distortion bound given in (3) for a signal of variance $\sigma_w^2 + \sigma_s^2$, thus verifying that our model of VSTM is optimal. That is to say, there cannot exist any physical system that achieves better performance using the same or fewer bits, on average, to encode information.

So far, we have considered the properties of the block labeled VSTM in Figure 1. We designed this VSTM system to be optimal subject to a fixed capacity limit, where the optimality criterion was preserving and transmitting an incoming sensory signal with minimum distortion. For an agent performing a task, the signal in the world, x , may be of more interest than its noisy sensory encoding x_s . If the agent possesses knowledge of the statistics of the information source and the noise characteristics of its sensory system, it is straightforward to extend the model to compute a least-squares estimate of x given the capacity-limited memory signal x_m . In particular, this least squares estimate \hat{x} is given by

$$\hat{x} = \frac{\mu \sigma_s^2 + x_m \sigma_w^2}{\sigma_w^2 + \sigma_s^2}. \quad (5)$$

Finally, we may also extend the model to the case where several items have to be stored in VSTM simultaneously. If memory capacity is given by R bits and there are n items

to encode in VSTM, then a simple approach would be to evenly distribute R/n bits among each item. This model would correspond to the continuous resource model (Bays & Husain, 2008), which evenly distributes a continuous pool of resources (in this case bits) among visual features in a scene. By contrast, the discrete slot model (Cowan & Rouder, 2009) instead assumes that VSTM consists of a set of encoding slots (typically assumed to be 4), each of which has some fixed encoding precision. This could also be implemented in our model by focusing $R/4$ bits on a subset of 4 items in the scene.

More generally, our ideal observer model of VSTM can encompass both of these competing theories. If a scene contains n items, we define a distribution over the probability of encoding 0 through n items in VSTM. This distribution allows a subject to encode a different number of items on different trials (for example, sometimes encoding 2 items, and other times encoding 4 items, etc.). According to the continuous resource hypothesis, n items will be encoded with probability 1.0, whereas the discrete slot model predicts that 4 items will be encoded with probability 1.0. Our ideal observer model actually defines a continuum of possible memory encoding strategies, with the continuous resource and discrete slot models but two special cases out of this continuum.

Application of the model to human data

In the previous section, we derived our ideal observer model of human VSTM and proved its optimality from an information-theoretic perspective. Given that any physical system that transmits information must have a finite capacity, we designed our model of VSTM to be maximally efficient subject to this constraint. This property enables an important application of our model to empirical data. By measuring human memory performance, the model enables us to infer the minimum memory capacity of VSTM necessary to achieve the observed level of performance.

Our model also yields a novel prediction regarding human performance. The predicted variance of memory (given by equation 4) depends not only on the capacity of memory and the magnitude of sensory noise, but also on the distribution of information in the environment. If humans have a fixed memory capacity and use that capacity in a nearly optimal manner to encode information with a known distribution, the model predicts that performance should be worse as the variance of the information source σ_w^2 increases. A corollary to this prediction is that although performance should increase as the signal variance decreases, the allocated memory capacity (e.g., the number of bits) should remain constant.

In the next section, we describe an empirical study designed to explore the formal properties and predictions of our model. In applying our model to human data, we had three goals in mind. First, we sought to estimate the capacity of human VSTM in terms of the information-theoretic quantity of bits. Second, we sought to test the novel prediction of our model regarding the relationship between memory performance and the variance of information encoded in memory. Finally, we sought to uncover evidence regarding the encoding strategy

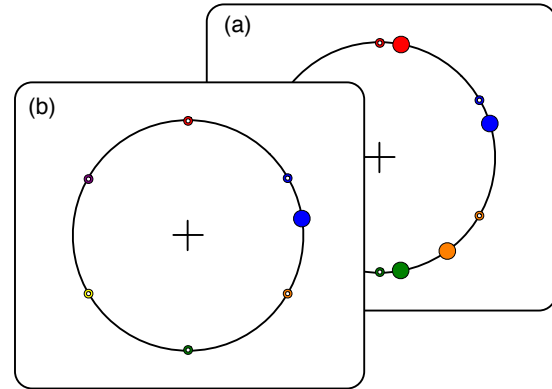


Figure 2: Illustration of the task stimuli. (a) Subjects are shown a circular array of six small ‘home locations’, and a varying number of larger colored dots (in this case 4). (b) After a 500 ms blank retention interval, subjects are shown a display containing just one of the original dots. Subjects must decide if the dot has been perturbed clockwise or counterclockwise from its previous location.

used by humans, as a means of exploring the ongoing debate between competing models of the allocation of VSTM resources.

Experiment

Method

Participants Eight volunteers from the University of Rochester participated in the experiment. All subjects had normal or corrected-to-normal vision.

Apparatus Subjects were seated 40 cm from an LCD monitor set to a resolution of 1280×1024 pixels. Subjects’ heads were kept in a fixed location using a chin rest. During the experiment, subjects wore a head-mounted eye tracker (EyeLink II; SR Research). The experimental software was written to ensure that subjects maintained stable fixation for the duration of each trial. Trials where eye movements were detected were repeated.

Procedure At the start of each trial, subjects were shown a screen containing a large ring and a array of six evenly spaced small circles (see Figure 2). These circles are subsequently referred to as the ‘home locations’. Each home location was a different color, and their location did not change from trial to trial. The display also contained a fixation cross, located at the center of the display. Subjects were instructed to maintain fixation on this cross for the duration of each trial. After 250 ms, a stimulus array was displayed (Figure 2a). This display contained the original home locations, but also contained a varying number of larger colored dots. These larger dots were randomly located around the circle, where each location was drawn from a Gaussian distribution with the color-corresponding home location as the mean location. On different trials, the number of larger dots (the set size) varied, using a set size of 1, 2, 4, or 6 items. Subjects were instructed to

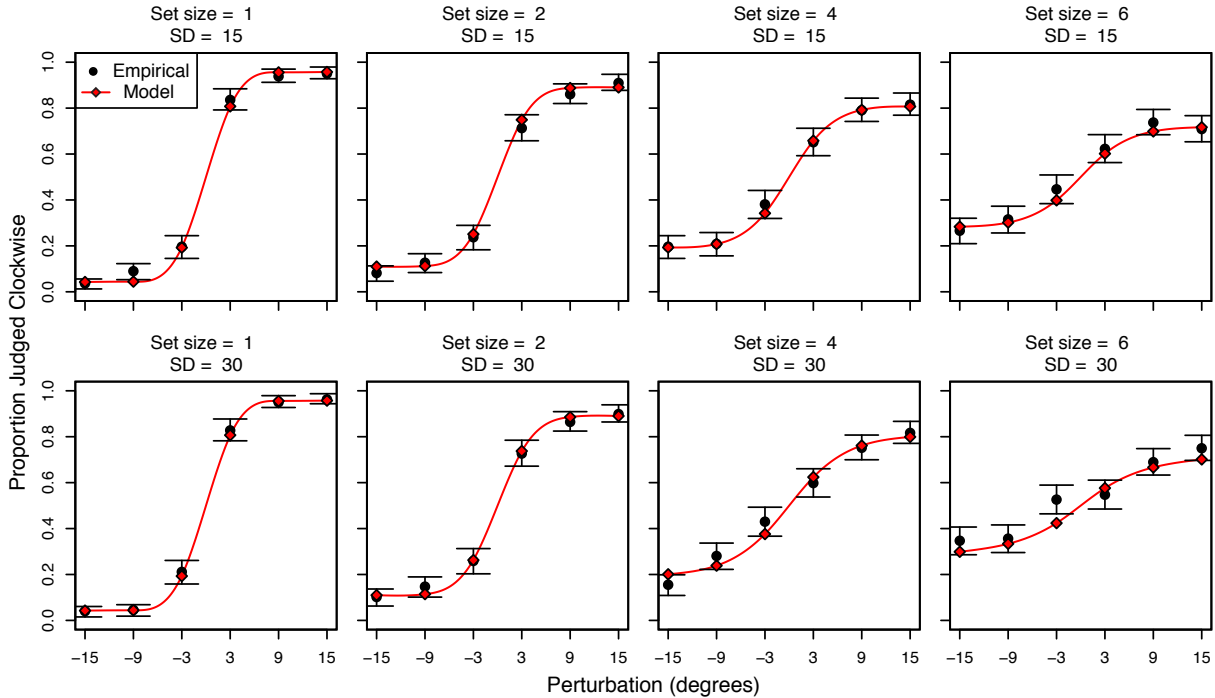


Figure 3: Probability of responding clockwise for each condition of the experiment. Black circular markers indicate human performance, red diamond-shaped markers indicate the posterior predictive mean data for the ideal observer model. The smooth curve was generated by fitting a spline to the model predictions. Error bars are for human data, and correspond the 95% highest density credible interval.

memorize the location of each larger dot relative to its home location. This stimulus presentation duration lasted for 1 second, after which the display was blanked for an interval lasting 500 ms.

After the memory retention interval, subjects were shown a display containing the original home locations, but just one of the original larger dots (Figure 2b). The location of this larger dot was always different from its previous location, and on different trials was perturbed by an amount drawn from the set $\{-15, -9, -3, 3, 9, 15\}$ degrees. The task for the subject was to decide whether the dot had been perturbed clockwise or counterclockwise relative to its previous location. Subjects responded by pressing one of two keys, depending on the direction of the perturbation. Subjects were then given feedback regarding the correctness of their choice.

The variance of the Gaussian distribution governing the position of the dots was also manipulated as a within-subject condition. Subjects completed two sessions of the experiment on separate days. On one of the days, the dots were drawn from a Gaussian distribution with low variance ($SD = 15$ degrees), while on the other session the Gaussian distribution used a high variance ($SD = 30$ degrees). In both conditions the mean of the Gaussian distribution was always centered at the dot's home location. The order of the two sessions was counterbalanced across subjects.

In all, subjects completed 30 trials in each of 48 conditions (4 set sizes \times 6 perturbations \times 2 variance conditions) over the course of two 1-hour sessions.

Results

In Figure 3, the proportion of clockwise responses is plotted for each condition of the experiment (the black circular markers indicate human data). Separate columns in the figure show performance at each set size ($N = 1, 2, 4,$ or 6 items) and the two rows of the figure show performance in the low variance (top row) and high variance (bottom row) conditions. Error bars correspond to the Bayesian 95% highest density credible interval, assuming a uniform prior for correct response rate.

To examine human performance in greater detail, we fit Gaussian cumulative density functions (CDFs) to the data in Figure 3. For each panel in the figure we estimated a separate mean and precision (inverse of variance) parameter. We also estimated a separate lapse rate for each set size, where a lapse trial corresponds to a guessed response due to the probed item not being encoded in memory. The Gaussian CDFs were fit using a Monte Carlo Bayesian parameter estimation procedure (Kruschke, 2011), using broad priors for each parameter¹. According to the continuous resource model (Bays & Husain, 2008), memory precision should decrease with increasing set size, while the lapse rate should remain constant. By contrast, according to the discrete slot model (Cowan & Rouder, 2009) the lapse rate should increase with set size, but

¹We used a uniform(0, 1) prior for the lapse rate, a uniform(0, 100) prior for the standard deviation of the Gaussian CDF, and a Gaussian prior with mean = 0 and precision = 0.001 for the mean of the Gaussian CDF. We collected 100,000 Monte Carlo samples from the posterior for each parameter.

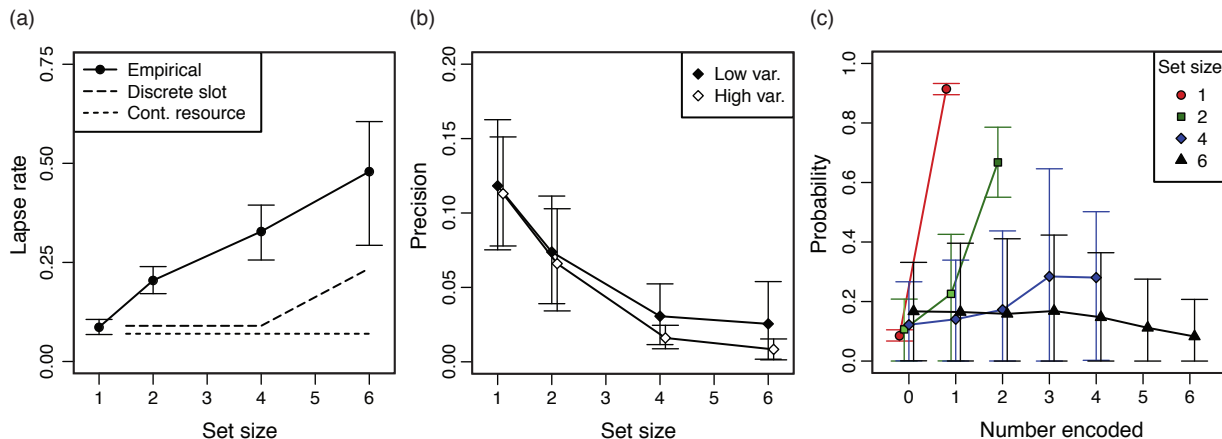


Figure 4: (a) Estimated lapse rate for a Gaussian CDF fit to human data. Dashed lines indicate hypothetical predictions of the discrete slot and continuous resource models. (b) Estimated memory precision (inverse of variance) for a Gaussian CDF. (c) Estimated probability of encoding $n = i$ items, for $i = 1$ to N for the ideal observer model. All error bars indicate 95% highest density credible intervals.

only once the set size increases beyond the available number of slots (typically assumed to be ~ 4).

The estimated lapse rate is shown in Figure 4a, while the estimated precision is plotted in Figure 4b. The results of our analysis suggest that neither the discrete slot nor continuous resource models provide an adequate account of our data. The continuous resource model predicts a constant lapse rate, whereas the discrete slot model predicts that lapse rate should only increase as the set size becomes greater than the available number of slots. The dashed lines in Figure 4a plot hypothetical predictions for these two models. In our data, precision was found to decrease, in accord with a model that must distribute a continuous resource among items. However, lapse rate was observed to increase with increasing set size, and the change in lapse rate was significant even comparing set sizes of 1 and 2 items.

A novel prediction of our ideal observer model was that memory variance should be higher for the high variance condition of the experiment (or equivalently, memory precision should be lower). Estimated memory precision was not significantly different for the smallest two set sizes (the 95% credible interval for the difference in precision included zero), whereas precision was significantly lower in the high variance condition for the two largest set sizes, as can be observed in Figure 4b. Thus, our empirical data qualitatively support the prediction of our ideal observer model. In the next section, we apply our model to evaluate whether it can also quantitatively account for the specific pattern of results obtained in the experiment.

Modeling Results

For each trial of the experiment, the model encoded the angular position of a subset of stimulus items in memory, using the VSTM mechanism illustrated in Figure 1. Memory capacity was evenly distributed among each encoded item. During the test portion of each trial, the model observed a sensory noise corrupted version of the probe item, and compared it

to the memorized location of the original item. If the probed item was not encoded during the stimulus presentation, the model randomly chose its response. Otherwise, if the probe item was perceived to be clockwise relative to the remembered location, the model responded that the perturbation was clockwise, and otherwise it responded counterclockwise.

The model contains three types of parameters: the memory capacity (R), sensory noise (σ_s), and the probability of encoding $n = i$ items for $i = 0$ to N , where N is the set size on the current trial. As noted previously, the continuous resource model (Bays & Husain, 2008) and the discrete slot model (Cowan & Roudier, 2009) can be viewed as special cases of our model by setting $n = N$ with probability 1.0 (continuous resource model), or $n = 4$ with probability 1.0 (discrete slots model). By placing a distribution over the number of items encoded in VSTM, our ideal observer model encompasses a range of possible models, where the continuous resource and discrete slot models are two special cases.

Rather than attempting to fit the human data by searching over the parameter space, we placed broad prior distributions over each parameter², and used Bayesian inference to infer distributions over probable parameter values (see Kruschke, 2011, for an introduction to Bayesian data analysis techniques). Recall that our model predicts that memory capacity should remain constant across the two variance conditions of the experiment, even though memory precision decreased in the high variance condition. Because of this prediction, we estimated separate capacity parameters, R_{Low} and R_{High} , for each variance condition. If the inferred capacity is the same or similar in both conditions, this would confirm our model's ability to parsimoniously account for the human data.

²For the memory capacity, we used a uniform prior in the range (0, 100) bits. For the sensory noise σ_s , we used a uniform(0, 100) prior, and for the encoding probabilities, we used a flat Dirichlet(1, 1, ..., 1) distribution to define a prior over the probability of encoding 0 through N items for each set size. We collected 100,000 samples from the posterior distribution for each parameter.

The inference process also allowed us to determine the posterior predictive distribution, or a prediction of what the human data should look like under the assumptions encapsulated by the model. If the posterior predictive data appears similar to the human data, this can be interpreted as evidence that the assumptions of the model are a viable explanation for the data. In Figure 3, the diamond-shaped plot markers indicate the posterior predictive mean behavior of the model. As can be seen by inspection of the figure, human and model data are in close quantitative agreement (in nearly all cases, the model data fall within the credible interval for the observed human data).

Figure 4c plots the inferred distributions over the number of items encoded in memory for each set size. In contrast to both the continuous resource model and the discrete slot model, it appears that subjects adopted a rather flexible encoding strategy for storing items in VSTM. For the set size 6 condition, the continuous resource model predicts a sharp spike at $n = 6$, whereas the slot model predicts a sharp spike at $n = 4$. Both of these models appear improbable in light of the data. In contrast, it appears that the number of stimuli encoded varied considerably from trial to trial, among the whole range of 0 to 6 items. A similar trend is observed for the other set size conditions, although as the set size decreases, there is an increasing tendency for subjects to encode the entire stimulus array.

The estimated mean sensory noise parameter σ_s was found to be 2.18 degrees (95% credible interval = 1.90 to 2.45). For the memory capacity parameters R_{Low} and R_{High} , the posterior mean estimate of capacity was 8.44 bits and 8.27 bits, respectively. The 95% credible interval for the difference in capacity between the two conditions ($R_{Low} - R_{High}$) equaled [-3.64, 2.70]; as this range includes zero we conclude that there is no significant difference in allocated memory capacity for the two variance conditions. Thus, our model is able to offer a parsimonious explanation as to why performance was worse in the high variance condition. Such a performance drop is predicted by an optimal memory system that has a single, fixed memory capacity, and must encode information from an information source with increased variance.

Summary & Conclusions

The nature of the capacity limit in human visual short-term memory (VSTM) is rather poorly understood. While previous theories have posited mechanisms intended to account for observed phenomena, in the present research we applied an ideal observer framework to uncover the expected behavior of an optimally performing, but finite capacity memory system. An advantage of our model is that it links a theoretically grounded and task-independent definition of capacity with quantitative predictions for performance in behavioral experiments.

To evaluate the predictions of our model, we conducted an experiment in which subjects must remember visual features (stimulus location) in arrays of varying set size. Based on these data, we were able to infer a quantitative estimate of the

capacity of human VSTM. Importantly, by using an optimal model of VSTM, this estimate represents a theoretical lower bound on human memory capacity.

We demonstrated that by allowing a flexible distribution of memory capacity among stimulus items, our model generalizes two previous, but competing models of VSTM as special cases. The continuous resource (Bays & Husain, 2008) and discrete slot models (Cowan & Rouder, 2009) differ primarily in how memory is divided among elements in the visual scene. The results of our analysis demonstrate that both models are improbable as explanations for human performance. Instead, it appears that humans exhibit tremendous flexibility in how they allocate their memory capacity (on different trials, encoding a widely varying number of items in memory). The ideal observer model was easily able to account for these data, and in fact provided a close quantitative fit to the observed results.

Finally, our model also generated a novel prediction regarding the precision with which humans can encode visual features with high versus low variance. Our model predicted that under a fixed capacity limit, memory precision should decrease with increasing stimulus variance. Human performance was found to closely match predictions from the ideal observer analysis.

Acknowledgments The authors would like to thank Leslie Chylinski for assistance with subject recruitment and data collection. This research was supported by grants NIH R01-EY13319 to David Knill and NSF DRL-0817250 to Robert Jacobs.

References

- Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*(7), 622-8.
- Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 1-11.
- Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851.
- Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression*. Englewood Cliffs, NJ: Prentice-Hall.
- Cowan, N., & Rouder, J. N. (2009). Comment on "dynamic shifts of limited working memory resources in human vision". *Science*, *323*(13), 877c.
- Kruschke, J. K. (2011). *Doing bayesian data analysis: A tutorial with r and bugs*. Academic Press.
- Luck, S. (2008). Visual short-term memory. In S. Luck & A. Hollingworth (Eds.), *Visual memory* (p. 43-85). New York: Oxford University Press.
- Luck, S., & Vogel, E. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279-280.
- Vogel, E., Woodman, G., & Luck, S. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology*, *27*(1), 92-114.