Optimal Attentional Allocation in the Presence of Capacity Constraints in Visual Search

Christopher J. Bates (cjbates@ur.rochester.edu)

Robert Jacobs (rjacobs@ur.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY

Abstract

There is large agreement among vision scientists that biological perception is capacity-limited and that attentional mechanisms control how that capacity is allocated. Despite the fact that Bayesian models generally do not include capacity limits, many researchers model perceptual attention as the result of optimal Bayesian inference. This inconsistency arises because vision science currently lacks a feasible and principled computational framework for characterizing optimal attentional allocation in the presence of capacity constraints. Here, we introduce such a framework based on rate-distortion theory (RDT), a theory of optimal lossy compression developed in the engineering literature. Our approach defines Bayes-optimal performance when an upper limit on information processing rate is imposed. Here, we compare Bayesian and RDT accounts in a visual search task, and highlight a typical shortcoming of unlimited-capacity Bayesian models that is not shared by RDT models, namely that they often over-estimate task-performance when information-processing demands are increased. In this study, we asked human subjects to find either one or two targets in a collection of distractors in a single-fixation search task. We predicted relative performance between one- and two-target conditions based on both RDT and Bayesian models. Performance differed between conditions in a way that was well accounted for by the capacity-limited RDT model but not by the capacity-unlimited Bayesian model.

Keywords: Visual attention, visual search, rate-distortion theory, resource rationality, information theory, Bayesian modeling, computational modeling

Introduction

Most publications on perceptual attention contend that attentional mechanisms exist as a way to allocate a limited computational resource. There is ample evidence, for example, that neural tuning curves can change (e.g., in response to a cue) so as to afford higher signal-to-noise ratios in some receptive fields at the cost of lower signal-to-noise ratios for other receptive fields (Reynolds, Pasternak, & Desimone, 2000; Desimone & Duncan, 1995; Spitzer, Desimone, & Moran, 1988). Experiments have also demonstrated that people can voluntarily change the spatial range of their focus of attention, and that an increase in spatial range comes at the cost of lower resolution (Carrasco, 2011). At the same time, however, there has been a debate within the visual search community about whether search times and detection accuracy are best described by a noisy but unlimited-capacity process ("data-limited") versus a limited capacity process that allocates more resource to some parts of a display. Results have been mixed, with some experiments finding stronger evidence for limited capacity, and others finding the reverse (Shimozaki, Schoonveld, & Eckstein, 2012; Eckstein, Shimozaki, & Abbey, 2002; Eckstein, Drescher, & Shimozaki, 2006; Davis, Shikano, Peterson, & Michel, 2003; Palmer, Fencsik, Flusberg, Horowitz, & Wolfe, 2011; Eckstein, 2011, 2017). In additional experiments, data are found to be well-explained by Bayesian or signal-detection models but these models are not compared to capacity-limited models (Ma, Navalpakkam, Beck, Van Den Berg, & Pouget, 2011; Eckstein, Thomas, Palmer, & Shimozaki, 2000; Eckstein, 1998; Schoonveld, Shimozaki, & Eckstein, 2007). Similarly, there are some experiments that support a role for capacity limits and compression of information (e.g., Rosenholtz, Huang, Raj, Balas, & Ilie, 2012).

Data from cueing-paradigm tasks (highly related to visual search) are often explained in terms of capacity limits. In these tasks, subjects are given a cue before the stimulus appears as to which of N locations is likely to contain the target object. Then they report some aspect of that stimulus. Usually the cue indicates the correct location, but on some trials it does not. It is generally found that the cue is helpful when it is correct and hinders when it is incorrect. These results have been explained by some form of attentional allocation, e.g., lowering neural noise at the cued location while increasing it at the uncued location(s). However, they have also been explained without appealing to capacity limits, by stipulating that the cue is incorporated as a Bayesian prior. An ideal observer can treat a cue as indicating a high prior probability of a target appearing at that location, and multiply this probability by the likelihood of the target given the sensory measurement. There are, however, clearly acknowledged limits to the Bayesian account of cueing effects (Shimozaki et al., 2012). For instance, a Bayesian account cannot explain "attentional capture" (Folk, Remington, & Johnston, 1992), in which subjects are unable to ignore cues to object locations even when they are random.

The evidence just presented for a possible lack of capacity limits in visual search would seem to fly in the face of other well-established results within the attention literature, which argue that capacity limits almost surely play a role. For example, it is well-established that people can simultaneously track only a limited number of moving objects, and that attending to the moving objects makes it harder to detect changes in other parts of the display (Alvarez & Franconeri, 2007; Alvarez & Oliva, 2008, 2009; Tombu & Seiffert, 2008). In this

paper, we present data suggesting previously unexplored limitations to the capacity-unlimited approach to visual search. But at the same time, we argue that a different kind of ideal observer model, based on rate-distortion theory (RDT), can reconcile the simultaneous successes of Bayesian ideal observers and capacity limits in explaining performance in perceptual tasks.

RDT models often make similar predictions to Bayesian models, because both are the result of optimizing accuracy given noise and uncertainty. For example, both models tend to predict a "regression to the mean" effect, whereby responses are biased toward the mean of a prior distribution over stimuli (Huttenlocher, Hedges, & Vevea, 2000). However, RDT models are also constrained by a limit on mutual information between stimulus and neural response (a measure of how well one can predict one quantity, such as a stimulus, from the other quantity, neural response), which causes them to make very different predictions as information processing demands are increased. Performance in a particular experimental condition can often be fit just as easily by a Bayesian model (with some level of noisiness given by σ) or an RDT model (with a capacity parameter, $\mathcal{C}$). However, studies in visual working memory illustrate an important way in which they diverge. For instance, Bates, Lerch, Sims, and Jacobs (2019) showed that performance in a change-detection task increased as the entropy (roughly, uncertainty) of the stimulus distribution decreased. They found that a single value of capacity in an RDT model was sufficient to explain the data across all conditions of the experiment, which varied along stimulus entropy. By contrast, an alternative Bayesian model could not explain the data without allowing sensory noise to vary with stimulus entropy. To give another example, Orhan and Jacobs (2013) found it necessary to allow sensory noise to increase with set-size in their Bayesian clustering model of visual working memory.

In the experiment presented here, we reasoned that if people adaptively allocate their perceptual capacity, their performance in an attentional task should be limited by the entropy of the stimulus distribution they are exposed to. While most visual search tasks use a single target, here we varied the number of targets (either one or two). Subjects had to report the direction of tilt away from vertical on all targets in the display. We designed our stimuli such that the RDT and Bayesian accounts made widely divergent predictions in the one- versus two-target conditions so we could easily distinguish between them. As was the case in the memory experiments mentioned above, the two models make different predictions because the RDT model is more sensitive to stimulus entropy than the Bayesian model.

Rate-Distortion Theory and Lossy Compression

This section provides an intuitive overview of RDT and lossy compression in the context of visual perception. Readers seeking additional information should see Bates and Jacobs (2020), Bates et al. (2019), Sims (2016, 2018), and Sims, Jacobs, and Knill (2012).

Consider the problem of communicating a signal or message, denoted x. For instance, a signal might be a visual image. In nearly all applications, one does not communicate x directly. Rather one communicates a code for x, denoted $\hat{x}$. For example, a code might be a neural code such as a pattern of neural activities. (In this case, the mapping $x \rightarrow \hat{x}$ is known as neural coding, and the mapping $\hat{x} \rightarrow x$ is neural decoding.) Ideally, one might set $\hat{x} = x$ so that a code conveys all the information about the signal, including all its fine details. That might be what one would do if there were no capacity constraints on a communication channel.

But physically-realized channels, such as neural circuits, always have limited capacity. It is therefore desirable to find an "efficient" coding system that is both compressed (i.e., on average, codes contain a small number of bits) and informative about messages (i.e., reconstructions of messages based on codes are reasonably accurate). Importantly, compressed codes can be "lossy". For example, a lossy code for an image might convey the coarse structure of the image, but not its fine details. More relevant to the topic of attention, a lossy code might convey the detailed structure of one portion of an image (a portion within an agent's focus of attention), but convey only the coarse structure of other portions (portions outside the focus of attention). A loss function quantifies the penalties for mismatches between signals and their reconstructions based on codes. RDT was developed in the engineering literature to characterize the trade-off between rate (or capacity) and distortion (or loss).

RDT defines a constrained optimization problem. It seeks a probability distribution over codes given signals, denoted $p(\hat{x}|x)$, that minimizes the expected value of a loss function. However, the mutual information between codes and signals (i.e., the average amount of information $\hat{x}$ conveys about x) cannot exceed the capacity of the communication channel. Formally, this constrained optimization problem is stated as follows:

$$p^*(\hat{x}|x) = \underset{p(\hat{x}|x)}{\arg\min}\, \mathbb{E}_{p(x,\hat{x})}\mathcal{L}(x,\hat{x}) \quad \text{subject to } \mathrm{MI}(x;\hat{x}) \leq \mathcal{C}. \tag{1}$$

where $\mathcal{C}$ denotes the channel's capacity (in bits) and $\mathcal{L}(x,\hat{x})$ denotes the loss function. The expected value of the loss function is taken with respect to the joint distribution $p(x,\hat{x})$. Because $p(x,\hat{x}) = p(x)\, p(\hat{x}|x)$, one typically specifies a prior distribution $p(x)$ over messages, sometimes referred to as an input or stimulus distribution. A maximum likelihood solution to the constrained optimization problem can be found using the Blahut algorithm (see Sims, 2016). Crucially, as indicated above, it is through the constrained optimization problem that RDT models find lossy codes implementing optimal attentional allocation in the presence of capacity constraints.

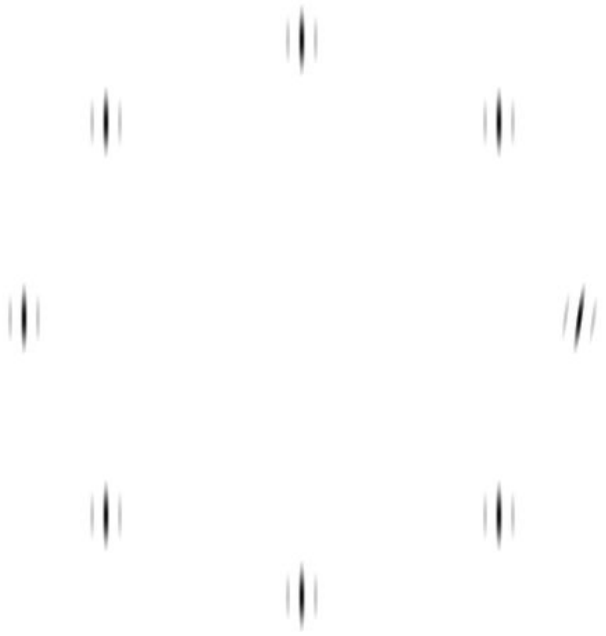

Figure 1: Example stimulus from the one-target condition. In the two-target condition, the two targets were always 180° apart, and the direction of the tilt for each target was chosen randomly.

Stimuli and Procedure

We gave subjects a visual search task with either one or two search targets. Displays consisted of $N = 8$ objects evenly spaced on a circle centered relative to a fixation cross. 'Distractors' were vertically-oriented Gabor-like objects[1], whereas targets were tilted a small fixed amount in the clockwise or counter-clockwise direction relative to vertical.

The stimulus on a given trial was generated as follows. First, a set of tilt values and locations was chosen for each target (either one or two). If there was just one target, the location was picked at random over the N possible locations. If there were two targets, the location of the first target was picked at random over all N locations, but the second target was constrained to be 180° apart on the circle. Thus, there were $8 \times 2 = 16$ equi-probable unique stimuli in the one-target case, and in the two-target case there were also $4 \times 4 = 16$ equi-probable unique stimuli. Figure 1 shows a sample stimulus from the one-target condition.

In the two-target case, the placement constraint for the second target was introduced to disincentivize anticipatory saccades away from the stimulus center and towards the ring of objects. Placement of objects along the circle ensured that visual acuity was approximately equal for all eight objects (assuming subjects maintained fixation at the stimulus center).

Amazon Mechanical Turk subjects were randomly assigned to the one-target condition (40 subjects) or the two-target condition (41 subjects). In both conditions, subjects were instructed to fixate on the cross in the center of the screen, which came on for 500 ms prior to the stimulus. The stimulus remained on the screen for 150 ms, and was followed by a response screen, where subjects used the mouse to select the location(s) and tilt(s) of the target(s). In both conditions, subjects were paid $6.00 to complete 500 trials. Most subjects took approximately 20-30 minutes to complete the task. Below we analyze only the last 200 trials.

[1]Gabors patches were generated using a standard 2-D Gabor filter that was rectified so that values did not go below zero.

Models

We compared two classes of models to subjects' responses: RDT and Bayesian. Both model classes shared the same (optimal) decision rule, but differed in how they calculated the sensory response. The Bayesian models assumed sensory responses were drawn from a von Mises (circular Gaussian) likelihood given the stimulus, while the RDT models assumed sensory responses were the outputs of an optimal lossy information channel (see Sims, 2016). We assumed subjects had exact knowledge of how their own sensory responses were produced given a stimulus, and that they had accurate knowledge of the stimulus prior distribution in the task when making a decision.

Decision rule: For both RDT and Bayesian models, the decision rule is given by:

$$p(y_\theta, y_{loc}|\hat{x}) = \sum_x p(y_\theta, y_{loc}, x, \hat{x})/p(\hat{x}) \tag{2}$$

where

- y_θ is either a scalar (in the one-target condition) or two-element vector (in the two-target condition) indicating the angle(s) of the target(s);
- y_{loc} indicates the location index (or indices) of the target(s);
- x represents the visual stimulus. Because the stimulus consists of eight Gabor patches, x is a vector with eight elements. Each element indicates the angle of its corresponding patch; and
- $\hat{x}$ is a model's sensory response, code, or representation of x;

The joint distribution can be factorized as $p(y_\theta, y_{loc}, x, \hat{x}) = p(y_\theta)p(y_{loc})p(x|y_\theta, y_{loc})p(\hat{x}|x)$. Note that $p(x|y_\theta, y_{loc})$ is deterministic, since the stimulus was always identical given values of target angle(s) and location(s).

RDT models: For the RDT models, $p(\hat{x}|x)$ was given by the solution to the RDT constrained optimization problem defined above (Equation 1). Optimal solutions were found using the `RateDistortion` package in R (see Sims, 2016). The exact form of the loss function (penalizing mismatches between x and $\hat{x}$) is described below.

Bayesian models: For the Bayesian models, $p(\hat{x}|x)$ was given by

$$p(x_s|x) = \frac{\prod_i^N e^{\frac{1}{\sigma}\cos(x_s^{(i)} - x^{(i)})}}{\sum_{x'} \prod_i^N e^{\frac{1}{\sigma}\cos(x'^{(i)} - x^{(i)})}} \tag{3}$$

where i indexes over items in the display, and

$$\hat{x} = \arg\min_{x'} \mathbb{E}_{p(x|x_s)} \mathcal{L}(x, x'). \quad (4)$$

That is, in the Bayesian models, sensory measurement x_s has a discretized von Mises distribution (discretized because x takes one of 16 possible values in both one- and two-target conditions), and $\hat{x}$ is chosen to minimize the expected loss given x_s.

One-parameter models: We first tried modeling experimental data with simple, single-parameter models: capacity $\mathcal{C}$ for RDT models (Equation 1) and σ for Bayesian models (Equation 3). The loss function for both was given by:

$$\mathcal{L}(x, \hat{x}) = \|\hat{x} - x\|^2. \quad (5)$$

However, neither of these models provided good fits with our experimental data. Consequently, we extended the models with two additional free parameters.

Full (three-parameter) models: First, in the two-target condition, it seems plausible that subjects cognitively understood that the two targets were 180° apart, but that this understanding did not influence their low-level sensory responses. In the models, we implemented this intuition by using the 180°-apart constraint in the decision-making part of a model (Equation 2; for example, the constraint was used when calculating $p(y_{loc})$). However, the full (three-parameter) models did not use this constraint in the sensory part of a model. Calculating Equations 1 or 4 requires consideration of a prior distribution over sensory displays. A "legal" display is one in which the two targets are 180° apart, and an "illegal" display violates this constraint. In the full models, we set the prior probability of an illegal display, $p_{illegal}(x)$, to be based on a value denoted τ. This was implemented so that if $\tau = 0$, then no probability mass was assigned to illegal values (corresponding to use of the 180°-apart constraint), and if $\tau = 1$, then the distribution over all displays (illegal and legal) was a uniform distribution.

Second, recall that subjects in our experiment indicated both the target location(s) and direction(s) of tilt on each trial. It seems plausible that subjects may have regarded either target location or tilt-direction as more important than the other. In particular, our data indicated that subjects were more accurate at identifying target location. Define the following two loss functions, denoted $\mathcal{L}_{SE}$ and $\mathcal{L}_{loc}$, as follows:

$$\mathcal{L}_{SE}(x, \hat{x}) = \frac{\|\hat{x} - x\|^2}{\max_{x'} \|x' - x\|^2} \quad (6)$$

$$\mathcal{L}_{loc}(x, \hat{x}) = \frac{\sum_{n=1}^{N} \mathbb{1}(x_n, \hat{x}_n)}{N_{targets}} \quad (7)$$

where n indexes over target locations, $N_{targets}$ is the number of targets, and $\mathbb{1}(x_n, \hat{x}_n)$ is an indicator function that equals one when a subject's response incorrectly identifies the Gabor at location n as a target. $\mathcal{L}_{SE}$ is the square-error loss between x and $\hat{x}$, whereas $\mathcal{L}_{loc}$ measures error based solely on subjects' estimates of target location. The full models used the loss function

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{SE} + \alpha\mathcal{L}_{loc} \quad (8)$$

where α is a parameter governing how much the loss is based on both target location and tilt-direction versus target location alone.

In summary, the full RDT models have three parameters ($\mathcal{C}$, τ, and α), and the full Bayesian models also have three parameters (σ, τ, and α).

Parameter fitting: For each model, we estimated its maximum likelihood parameter values based on trials from (i) the one-target condition, (ii) the two-target condition, and (iii) both conditions combined, using the `optim` function in the R programming environment. The likelihood of a model was given by

$$L(\phi) = \prod_t p_{y_\theta, y_{loc}|x}(x_{resp}^{(t)} | x^{(t)}) \quad (9)$$

where ϕ is the set of model parameters, t indexes over trials, and $x_{resp}^{(t)}$ is a subject's response on trial t. The probability $p_{y_\theta, y_{loc}|x}$ is the probability of the decision under a model, and was given by a probability matching rule (i.e., responses were chosen with frequency proportional to the probability they are correct; Da Silva, Victorino, Caticha, & Baldo, 2017; Wozny, Beierholm, & Shams, 2010; Craig, 1976).

Results

To assess the models, we compared their predicted response accuracies to subjects' response accuracies. We examined overall accuracy (both location and tilt correct), as well as location and tilt accuracies, independently (Figure 2). We found that subjects performed about 20 points worse in the two-target condition in terms of overall (both target location and tilt-direction) accuracy (79% versus 60% correct; see Figure 2, left panel). The full RDT model provides an excellent quantitative fit to this experimental finding. By contrast, the one-parameter RDT predicts identical performance in both conditions (since the stimulus entropy is identical across conditions), and the full and one-parameter Bayesian models predict a large *increase* in accuracy in the two-target condition. Intuitively, this prediction can be understood as a result of the constraint that the second target is fixed relative to the first. Many sensory measurement errors can be "cleaned up" given the constraint on target locations, since measurements that would result in a constraint violation can be ignored. As a result, the Bayesian models incorrectly predict better performance in the two-target condition relative to the one-target case.

We found that the full RDT model gave the best fit to the overall accuracies, predicting subjects' mean performance nearly perfectly. Neither one-parameter model could explain the data very well. The one-parameter RDT model clearly

outperformed the one-parameter Bayesian model when parameter fits were based on all data, though the one-parameter Bayesian model had an advantage in likelihood when parameters were fit separately for each condition. Both models matched overall human performance well when allowed to fit data from each condition separately.

The middle panel of Figure 2 presents the same models as the left panel, except that only location accuracies are presented (that is, the percent of responses that indicated the correct target locations, even if the tilt directions were reported incorrectly). We find that the full RDT model better predicts the location accuracies (compare blue and pink lines for one-parameter and full RDT models, respectively). Because α in the full RDT model was estimated to be greater than zero, it seems that subjects may have been slightly more concerned with locating targets than identifying their tilt directions. Similarly, we find that the tilt accuracies are well-accounted for by the full RDT model, but not the full Bayesian model (right panel).

Tables 1 and 2 report the results of our maximum likelihood fits for the full and one-parameter models, respectively. We find that when comparing the full models, the log-likelihood values favor the RDT model over the Bayesian model when considering all experimental trials, and also when considering only the one-target or two-target trials.

Comparing parameter values fit to one-target trials versus two-target trials versus both sets of trials provides an opportunity for important sanity checks. Ideally, a single set of parameter values should be able to explain data in both conditions, as it is unrealistic to presume that, for instance, channel capacity or sensory noise magnitude change across conditions. For the full RDT model, we found that C and α had very similar values regardless of whether these values were fit to one-target trials or all trials (recall that τ does not play a role in one-target trials). We found somewhat unexpected values when fitting the full RDT model to the two-target condition alone, as they should ideally be close to the values found when fitting both conditions together and when fitting the one-target condition alone. We believe this can be explained in part by the finding that there was higher inter-subject variance in the two-target condition and performance was non-normally distributed with a long tail toward poorer performances. As a result, the optimizer required many more optimization steps to converge and the gradients were very small.

For the Bayesian models, we found a higher value for noise parameter σ in the two-target condition relative to the one-target condition when fitting a model to each condition separately. When fitting to both conditions, the most likely value was found to be between those values.

A blank entry in a table indicates that a parameter did not impact model predictions for the given model (e.g., τ in the one-target condition), or was not applicable. In addition, in some cases, tables specify a range of values for the Bayesian model parameters. This is due to the 'min' operator in those models, which results in ranges of parameter space that give identical predictions. We used a grid search over starting values of parameters used by the optimizer to compensate for the fact that gradients are flat in those areas.

Discussion

We motivated the present experiment by noting an apparent conflict in the attention literature between capacity-unlimited Bayesian models and capacity-limited models. We argued that RDT provides a natural reconciliation of the conflicting viewpoints, because RDT models often make similar predictions to Bayesian models while still being capacity-limited. Our experiment provides further evidence for capacity limits in visual search.

Fits of Bayesian and RDT models to the experimental data show that Bayesian models over-estimated task performance, particularly when information-processing demands were high (e.g., the two-target experimental condition), whereas RDT models provided highly accurate accounts of subjects' responses. We conclude that capacity constraints played a significant role in limiting subject performance when information processing demands were high. We also conclude that our RDT framework provides a useful computational formalism for characterizing subjects' allocation of attention in the presence of capacity constraints.

We found that while the full RDT model strongly benefited from two added parameters, the Bayesian model was not as sensitive to these parameters. Future work could search for other possible parameterizations that benefit the Bayesian model, although this comes with the drawback that it becomes more difficult to compare the RDT and Bayesian approaches as their assumptions diverge.

Here, we presented just one experiment that was designed to distinguish between capacity-limited versus capacity-unlimited accounts. On the whole, we found stronger evidence for capacity limits. Future experiments will seek to further probe our primary thesis that attentional allocation in visual search is capacity-limited. We have already conducted experiments to extend these ideas to cued visual search to complement the study just presented, as part of a longer manuscript.

Acknowledgments

The first author was supported by an NSF NRT graduate training grant (NRT-1449828) and an NSF Graduate Research Fellowship (DGE-1419118). This work was also supported by an NSF research grant (DRL-1561335).

References

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of vision*, *7*(13), 14–14.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological science*, *19*(4), 392–398.

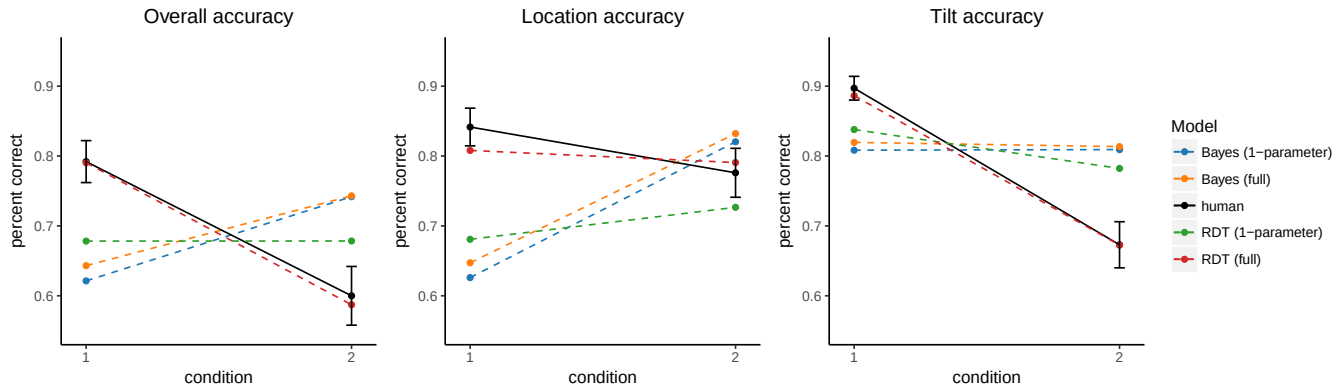

Figure 2: Model predictions and experimental data (overall accuracy, location accuracy alone, and tilt accuracy alone). The percentage of correct responses is plotted in each condition and compared to two versions of model predictions. The simpler version of the model has only one fitted parameter (capacity or sensory noise magnitude). The full model includes three fitted parameters.

Table 1: Inferred model parameter values for full models.

	$\mathcal{C}$	σ	α	τ	log likelihood (all trials)	log likelihood (one condition)
RDT (both)	3.1	-	0.36	0.81	-22435	-
RDT (one-target)	3.1	-	0.38	-	-	-8577
RDT (two-target)	2.4	-	0.24	0.00001	-	-13776
Bayesian (both)	-	0.145	≥ 0.7	< 0.5	-23545	-
Bayesian (one-target)	-	0.126	≥ 0.0	> 0.0	-	-8890
Bayesian (two-target)	-	0.172	$\geq 0.35, \leq 0.65$	> 0.5	-	-13934

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, *106*(18), 7345–7350.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*.

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), 11–11.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision research*, *51*(13), 1484–1525.

Craig, A. (1976). Signal recognition and the probability-matching decision rule. *Perception & Psychophysics*, *20*(3), 157–162.

Da Silva, C. F., Victorino, C. G., Caticha, N., & Baldo, M. V. C. (2017). Exploration and recency as the main proximate causes of probability matching: a reinforcement learning analysis. *Scientific reports*, *7*(1), 1–23.

Davis, E. T., Shikano, T., Peterson, S. A., & Michel, R. K. (2003). Divided attention and visual search for simple versus complex features. *Vision Research*, *43*(21), 2213–2232.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193–222.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, *9*(2), 111–118.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of vision*, *11*(5), 14–14.

Eckstein, M. P. (2017). Probabilistic computations for attention, eye movements, and search. *Annual review of vision science*, *3*, 319–342.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and bayesian priors. *Psychological science*, *17*(11), 973–980.

Eckstein, M. P., Shimozaki, S. S., & Abbey, C. K. (2002). The footprints of visual attention in the posner cueing paradigm revealed by classification images. *Journal of Vision*, *2*(1), 3–3.

Eckstein, M. P., Thomas, J. P., Palmer, J., & Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & psychophysics*, *62*(3), 425–451.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human*

Table 2: Inferred model parameter values for 1-parameter models.

	$\mathcal{C}$	σ	log likelihood (all trials)	log likelihood (one condition)
RDT (both)	2.6		-23862	-
RDT (one-target)	3.0		-	-8939
RDT (two-target)	2.3		-	-14583
Bayesian (both)	-	0.148	-24032	-
Bayesian (one-target)	-	0.126	-	-8890
Bayesian (two-target)	-	0.181	-	-14179

perception and performance, *18*(4), 1030.

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

Ma, W. J., Navalpakkam, V., Beck, J. M., Van Den Berg, R., & Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. *Nature neuroscience*, *14*(6), 783.

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

Palmer, E. M., Fencsik, D. E., Flusberg, S. J., Horowitz, T. S., & Wolfe, J. M. (2011). Signal detection evidence for limited capacity in visual search. *Attention, Perception, & Psychophysics*, *73*(8), 2413–2424.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of v4 neurons. *Neuron*, *26*(3), 703–714.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, *12*(4), 14–14.

Schoonveld, W., Shimozaki, S. S., & Eckstein, M. P. (2007). Optimal observer model of single-fixation oddity search predicts a shallow set-size function. *Journal of Vision*, *7*(10), 1–1.

Shimozaki, S. S., Schoonveld, W. A., & Eckstein, M. P. (2012). A unified bayesian observer analysis for set size and cueing effects on perceptual decisions and saccades. *Journal of Vision*, *12*(6), 27–27.

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

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*, 652–656.

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

Spitzer, H., Desimone, R., & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science*, *240*(4850), 338–340.

Tombu, M., & Seiffert, A. E. (2008). Attentional costs in multiple-object tracking. *Cognition*, *108*(1), 1–25.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS computational biology*, *6*(8).