Optimal attentional allocation in the presence of capacity constraints in uncued and cued visual search

Christopher J. Bates

Department of Psychology, Harvard University, Cambridge, MA, USA

Robert A. Jacobs

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA

The vision sciences literature contains a large diversity of experimental and theoretical approaches to the study of visual attention. We argue that this diversity arises, at least in part, from the field's inability to unify differing theoretical perspectives. In particular, the field has been hindered by a lack of a principled formal framework for simultaneously thinking about both optimal attentional processing and capacity-limited attentional processing, where capacity is limited in a general, task-independent manner. Here, we supply such a framework based on rate-distortion theory (RDT) and optimal lossy compression. Our approach defines Bayes-optimal performance when an upper limit on information processing rate is imposed. In this article, we compare Bayesian and RDT accounts in both uncued and cued visual search tasks. We start by highlighting a typical shortcoming of unlimited-capacity Bayesian models that is not shared by RDT models, namely, that they often overestimate task performance when information-processing demands are increased. Next, we reexamine data from two cued-search experiments that have previously been modeled as the result of unlimited-capacity Bayesian inference and demonstrate that they can just as easily be explained as the result of optimal lossy compression. To model cued visual search, we introduce the concept of a "conditional communication channel." This simple extension generalizes the lossy-compression framework such that it can, in principle, predict optimal attentional-shift behavior in any kind of perceptual task, even when inputs to the model are raw sensory data such as image pixels. To demonstrate this idea's viability, we compare our idealized model of cued search, which operates on a simplified abstraction of the stimulus, to a deep neural network version that performs approximately optimal lossy compression on the real (pixel-level) experimental stimuli.

Introduction

The scientific literature on visual attention contains a wide variety of experimental approaches and theoretical perspectives. For example, many publications 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 (Desimone & Duncan, 1995; Reynolds et al., 2000; Spitzer et al., 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). These findings have been interpreted by some (but not all) as evidence for resource reallocation.

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 (a so-called data-limited process) 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 data-limited processing (Eckstein, 1998, 2011, 2017; Eckstein et al., 2006, 2002, 2009; Palmer et al., 1993, 2000; Palmer, 1994; Shimozaki et al., 2003, 2012) and other experiments supporting limited capacity processing (Davis et al., 2003; Lu & Dosher, 1998; Palmer et al., 2011). In additional experiments on visual search, data are found to be well explained by unlimited-capacity Bayesian or signal-detection models, but these models are not always compared to capacity-limited models (Eckstein et al., 2000; Ma et al., 2011; Schoonveld et al., 2007). In still other experiments, data support a role for capacity limits and compression of information (e.g., Rosenholtz et al., 2012, though note that this work concerns differences between peripheral and foveal representations).

Data from cued visual search tasks are often explained in terms of capacity limits (Carrasco & Yeshurun, 1998; Posner, 1980). In these tasks, subjects are given a cue before the stimulus appears as to which

Citation: Bates, C. J., & Jacobs, R. A. (2021). Optimal attentional allocation in the presence of capacity constraints in uncued and cued visual search. *Journal of Vision*, *21*(5):3, 1–23, https://doi.org/10.1167/jov.21.5.3.

https://doi.org/10.1167/jov.21.5.3 Received May 26, 2020; published May 4, 2021 ISSN 1534-7362

of N locations is likely to contain a target object. Then they report target presence or absence. 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, such as 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 influences a Bayesian prior distribution over target locations. A Bayesian observer can treat a cue as indicating a high prior probability of a target appearing at the cued location and multiply this probability by the likelihood of the target given the sensory measurement (Shimozaki et al., 2003, 2012).

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 a display (Alvarez & Franconeri, 2007; Alvarez & Oliva, 2008, 2009; Tombu & Seiffert, 2008).

In this article, we argue that the wide range of experimental findings and theoretical claims in the vision sciences literature arises, at least in part, from the field's inability to unify differing theoretical perspectives. In particular, the field has been hindered by a lack of a principled conceptual and formal framework for simultaneously thinking about both optimal (Bayesian) attentional processing and capacity-limited attentional processing. A goal of this article is to supply such a framework, which is based on rate-distortion theory (RDT) and optimal lossy compression. We argue that this framework can reconcile the simultaneous successes of Bayesian ideal observers and of capacity limits in explaining performance in perceptual tasks.

Here, we apply RDT models to both uncued and cued visual search tasks. When examining uncued search, we present data suggesting previously unexplored limitations to data-limited models of visual search and providing additional evidence that the search is more capacity limited than data limited. Next, we extend the framework to the domain of cued visual search, where we model results from two cueing experiments that have previously been taken as evidence for an absence of a capacity limit in search (Shimozaki et al., 2003, 2012). We show that these results can just as easily be explained within our optimal, capacity-limited framework.

The article is organized as follows. The following two sections explain RDT at intuitive and formal (mathematical) levels. The subsequent two sections compare our RDT framework to existing theories of attentional capacity limits in visual search and to unlimited-capacity signal detection and Bayesian theories, respectively.

The following section reports our experimental and modeling work on uncued visual search. 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 (roughly, the uncertainty in the prior distribution over target locations) 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 viewed displays with vertically oriented distractors and nonvertically oriented targets and reported the direction of tilt away from vertical on all targets in the display. We designed our stimuli such that RDT and unlimited-capacity Bayesian accounts made widely divergent predictions in the one- versus two-target conditions so we could easily distinguish between them. The two models make different predictions because the RDT model is more sensitive to stimulus entropy than the Bayesian model due to its capacity limitations.[1]

The next section considers cued search. We revisit important experimental studies by Shimozaki and colleagues (Shimozaki et al., 2003, 2012). We offer an explanation for how subjects respond to cues that is different from the authors' strictly Bayesian explanation and addresses its known limitations. Most important, the explanation by Shimozaki et al. does not generalize beyond the case of signal-detection per se. For instance, it is well known that people can modulate their attention to different aspects of a stimulus (e.g., different features in an image) even when the prior probabilities over target locations have not changed, meaning that attentional modulation is not equivalent to prior probability modulation. By contrast, our RDT-based explanation generalizes to any kind of attentional shift. In the context of cued visual search, it posits that subjects respond to cues by allocating more representational resources (i.e., bits) to locations where the target is more likely to appear.

Finally, we demonstrate that our approach based on efficient lossy compression can feasibly scale up, such that we can predict behavior at the level of raw sensory signals. Specifically, we implement an image-computable model of one of the cued search experiments that takes real stimulus images (pixel values) as input. We compare its predictions to the idealized model, which operates on simplified abstractions of the stimuli, as is typical in psychological modeling.

RDT and optimal lossy compression: An intuitive account

This section provides an intuitive overview of RDT and optimal 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 et al. (2012).

Our framework regards a visual observer as a type of "communication channel" that, speaking loosely, needs to communicate information from visual portions of its brain to portions that control decision making and behavior. Let's consider the problem of communicating a signal, such as a visual image, denoted x. 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 encoding, 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 in biological observers, always have limited capacity. It is therefore imperative to find an "efficient" code satisfying the following two properties. The first property is that a code must be compressed, meaning that, on average, it represents signals using fewer representational resources (i.e., bits) than the channel's maximum capacity. For example, if a channel has a small capacity, then a compressed code for an image might convey the coarse structure of the image but not its fine details. This property is a hard constraint that cannot be violated.

The second property is that a channel's code should be maximally informative about signals in a task-dependent manner, meaning that it represents as much task-relevant information in the signals as possible. This criterion is a soft constraint in the sense that the representation of task-relevant information should be maximized to the greatest extent possible. Importantly, these two properties interact with each other. A low-capacity channel will need to use highly compressed codes that might convey only some of the task-relevant information. In contrast, a high-capacity channel can use codes that are only mildly compressed (or perhaps not compressed at all) and therefore can convey more task-relevant information.

In our RDT framework, the notion of task-dependent lossy compression is critically important to the idea of capacity-limited visual attention. In a communication channel—or a visual observer—a lossy code might convey the detailed structure of one portion of an image (a portion within the observer's focus of attention) but convey only the coarse structure of other portions (portions outside the focus of attention).

How does RDT find a task-dependent capacity-limited lossy code? It does so based on three inputs provided by the user. One input is the capacity of the communication channel. It is through this input that a user quantifies the "bottleneck" or capacity limit on performance. As indicated in the previous paragraph, our RDT framework conjectures that attention helps alleviate problems associated with low capacity. In a visual search task, for example, attention is a way of representing some portions of an image with high fidelity (namely, those portions likely to contain a target) while simultaneously representing other portions (regions unlikely to contain a target) with low fidelity.

Capacity limits are typically formalized using the information-theoretic unit of "bits." This is an appealing feature of RDT because this unit is "assumption free," meaning that it does not depend on the nature of a task or the nature of an observer's internal representations and operations. This helps explain the broad generality of RDT.

A second input is a loss function that quantifies the penalties for mismatches between signals and their task-dependent reconstructions based on codes. It is through the loss function that a user specifies properties of a task. For example, consider a TSA officer visually searching bags at an airport. If the officer's code for an image fails to represent a weapon, then that failure would carry a large penalty, whereas if the code fails to represent a toothbrush, then that failure would carry a small penalty.

The loss function is another appealing aspect of RDT. In cases in which subjects' understanding of a task differs from an experimenter's understanding, the experimenter may want to try multiple loss functions to see which one provides the best fit to subjects' data. An experimenter can also attempt to infer subjects' loss function from their data (Sims, 2015).

The final input is a prior distribution over visual stimuli. This distribution carries information about the statistical regularities in a visual environment. Similar to Bayesian approaches, this information is used by RDT to find optimal codes (lossy codes that minimize task errors as quantified by the loss function). This prior should match the knowledge of the subject being modeled. In some circumstances, the experimenter may choose to assume a prior distribution that they believe approximately matches the subject's, while in other circumstances, the experimenter may fit a parametric prior to subjects' responses (see, e.g., Bates et al., 2019).

The RDT framework studied here should be regarded as a "computational theory" of vision in the sense of Marr (1982). That is, it provides an analysis of the visual task faced by an observer, identifying the problems that need to be solved during task performance along with their optimal solutions. Although Bayesian approaches also provide computational theories, the RDT framework differs from Bayesian approaches because RDT considers capacity limits—formalized in a general, task-independent manner—when it computes optimal solutions to problems. Like Bayesian approaches, RDT does not provide an analysis at the level of "representation and algorithm," which seeks to identify

an observer's mental representations and operations, or at the level of "hardware implementation," which seeks to understand an observer's underlying hardware.

Historically, RDT has been developed primarily within the engineering literature to characterize the trade-off between rate (or capacity) and distortion (or loss). We believe that the success of RDT in the engineering literature over many decades—and in a wide variety of task contexts—bodes well for its success in the vision sciences. In general, other information-theoretic approaches to attention and other aspects of cognition have been tried in the past (e.g., Kahneman, 1973). These approaches fell out of favor in much of psychology as they failed to explain many important phenomena (Luce, 2003). A crucial shortcoming of earlier attempts has been their use of lossless compression. Because a lossless code represents all information in signals—both task relevant and task irrelevant—it has limited applicability to biological perception and cognition in real-world settings. In contrast, RDT emphasizes capacity constraints that lead to task-dependent lossy compression. Capacity constraints mean that the end use (or task goal) of a code (or representation) is critical for its design. Due to capacity constraints, an efficient communication channel (or visual observer) must choose which information in the signal to prioritize. Hence, task goals (loss function) and the statistical nature of a channel's environment (prior or stimulus distribution) play essential roles in shaping a channel's performance. By taking both of these factors into account, RDT is capable of accounting for structured psychological representations and in fact provides an especially strong account of psychological similarity and Shepard's law of generalization (Sims, 2018).

Finally, it is important to emphasize that while traditional information theory applications address the case of "signal reconstruction," the theory applies equally when the coded message that is transmitted is not ultimately used for reconstruction, per se. Therefore, it is applicable to the brain, which does not generally try to reconstruct the signal. For example, the engineering application of communicating over video chat concerns signal reconstruction: The pixels sent from your computer are first mapped to a compressed code, this code is sent to a receiving computer, and finally the receiving computer tries to reconstruct the original image from the code. This same process occurs in the brain, except that the end use for the code is not reconstruction of the original signal but rather some downstream task. For example, there is no brain region devoted to reconstructing retinal activations (this would be highly inefficient!). Rather, the retinal signal is decoded to extract behavior-relevant information about scene and object categories, ensemble statistics, material and geometric properties, and so on. The neural code from which these kinds of information are extracted should be carefully designed to contain as much task-relevant and as little task-irrelevant information as possible.

RDT and optimal lossy compression: A formal account

With this intuitive account of RDT as a foundation, we next provide a formal (mathematical) account. 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 the code $\hat{x}$ conveys about signal x or vice versa) cannot exceed the capacity of the communication channel. This constrained optimization problem is stated as follows:

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

where $\mathcal{C}$ denotes the channel's capacity (in bits) and $\mathcal{L}(x,\hat{x})$ denotes the loss function. Mutual information is given by

$$\mathrm{MI}(x;\hat{x}) = \sum_x \sum_{\hat{x}} p(x,\hat{x}) \log \frac{p(x,\hat{x})}{p(x)p(\hat{x})}. \tag{2}$$

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 signals, 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).

In general, the constrained optimization problem is computationally tractable only when signals are low-dimensional. In the context of visual perception, that means that exact RDT cannot use images when these images are represented at the pixel level (at least, not with current conventional computers). Instead, a user needs to make assumptions about what types of high-order visual features might be important to observers (e.g., orientation, color, size) and then develop an abstract, low-dimensional representation for images based on these features. This is the strategy that we used in many of the simulations reported below.

Fortunately, however, researchers studying deep neural networks (DNNs) have developed DNNs that approximately solve the RDT constrained optimization

problem, often with excellent results, even when images are represented at the high-dimensional pixel level. Consequently, these DNNs are referred to as "image-computable" models. An advantage of these models is that the user does not need to make any assumptions about high-order visual features that observers might be using. Below we present simulation results using an image-computable DNN.

Relation to previous theories of attentional capacity limits in visual search

Previous theories of visual attention in the context of visual search often conjecture that visual processing is capacity limited. Critically, however, these theories have failed to identify a benchmark for optimal capacity-limited performance, which makes it difficult to know when performance in a task is efficient (or at least "good"). This matters because claims about efficiency are central to theories of capacity limits in visual search (Wolfe & Pashler, 1998).

All current behavioral approaches to assessing capacity limits in visual search involve measuring performance as a function of task demands. Consider, for instance, the "workload" approach to visual attention. In a single-target visual search task, the number of distractor items (or workload) can be varied. It is found that an increase in the number of distractors often leads to worse performance (higher search times, lower accuracy), but not always (Wolfe & Pashler, 1998; Eckstein, 2011). When performance is constant as a function of load for some stimulus set, it is often said that capacity limits did not play a role. This interpretation of capacity limits fits naturally with notions of parallel versus serial computation—when computation is parallel, it is efficient and therefore "unlimited capacity." Formal mathematical models have been developed along this line of thinking to predict response-time distributions as a function of processing architecture (Townsend & Nozawa, 1995).

The workload approach to assessing capacity limits has had limited success. Researchers have found interactions between set size and performance (response times or accuracy) to vary greatly depending on the particular stimuli involved, and results do not map neatly onto a strict parallel versus serial processing dichotomy in which "primitive" features are encoded in parallel across the visual field and composite features or objects are encoded in a more laborious, serial manner (Treisman & Gelade, 1980; Hochstein & Ahissar, 2002). As a result, it has proven difficult to quantitatively predict search efficiency on novel stimuli. However, it is worth noting that some progress has been made within the isolated domain of simple stimuli on making quantitative predictions of performance as a function of workload. This line of work assumes that perceptual processing of each feature or stimulus is limited by a fixed number of noisy samples (or recruitable, statistically independent neurons), which can be distributed to different locations in the display as needed (Palmer et al., 1993; Eckstein et al., 2009). These models can be related to our approach, but they fundamentally differ in that they rely on a priori assumptions about perceptual noise distributions that are generally suboptimal from the standpoint of RDT.

Attentional bottlenecks have also been investigated within the "dual-task" paradigm (related to our two-target condition in Experiment 1; Han et al., 2003; Liu et al., 2009; Menneer et al., 2009; Moore & Osman, 1993; Palmer et al., 2020; Pastukhov et al., 2009; Stroud et al., 2012; Sperling & Melchner, 1978; VanRullen et al., 2004). In these tasks, experimenters ask subjects to report on two values (i.e., "tasks") concurrently and try to predict which task pairs will interfere with each other. If performance on one task is degraded by adding another concurrent task, then it is assumed that both tasks share a common, limited resource. For example, it is generally found that reporting feature values from two separate objects or regions within a display is more difficult than reporting from just one but that attention can be split to varying degrees between the two (e.g., Sperling & Melchner, 1978).

Researchers can diagnose the degree of interference for two simultaneous reports by plotting what is known as the attention operating characteristic (AOC) (Gottlob et al., 1999; Pastukhov et al., 2009; Sperling & Melchner, 1978). The AOC plots the performance on one task against performance on the other. When two tasks interfere (and thus presumably compete for the same limited resource pool), greater accuracy on one will come at the expense of lower accuracy on the other. These bottlenecks may occur at the perceptual stage or some later processing stage. However, while plotting psychometric curves such as the AOC can diagnose when two dimensions or tasks draw on the same resource pool, it has proven difficult to generalize results using previous theories. That is, simply knowing how two tasks interfere does not allow us to make quantitative predictions about interference in other pairs of tasks.

The RDT framework suggests several culprits for the difficulties just reviewed. For one, workload variables (e.g., number of distractor objects, similarity between objects) are task specific and experimenter defined, and their correspondences with relevant quantities in the visual system are unclear (Kantowitz, 1987). As already mentioned, major theories of attentional capacity limits claim that "easy" tasks are capacity unlimited while "hard" tasks are capacity limited. But they lack an objective baseline of comparison to know what is easy or hard for a capacity-limited agent. In

RDT analyses, by contrast, the unit of information (i.e., bits) provides a task-agnostic unit of measure that allows cross-task comparisons. Specifically, RDT models consider the number of bits that must be transmitted in order to achieve the observed level of performance in a task. According to RDT, this measure, in turn, critically depends on what stimulus distribution a subject is optimized for and how well their behavioral goals align with the experimental task. For instance, if goals (see Figure 3) or stimulus statistics (see discussion of parameter τ, below) are poorly aligned, the experimenter will find inefficient search performance. Once a subject's goals and environments are specified (or well approximated), it is possible (in principle) to predict performance in any perceptual task with no changes to the RDT model (see below). In this sense, the RDT approach is completely task independent.

A related difficulty with previous theories is that they make assumptions about which high-order visual features (e.g., orientation, spatial frequency, color, size, shape) exist in people's representations of (natural) images (Duncan, 1984, 1993; Duncan & Nimmo-Smith, 1996; Hochstein & Ahissar, 2002; Liu et al., 2009; Scholl, 2001; Treisman & Gelade, 1980). Formed primarily on intuition, these assumptions seem problematic in the context of natural images and ecological tasks (Orhan & Jacobs, 2014). Alternatively, image-computable DNNs trained to analyze (pixel-level) images have become increasingly popular as models of biological processing (Jacobs & Bates, 2019; Kriegeskorte, 2015; Yamins & DiCarlo, 2016), at least in part because they do not make these assumptions. Instead, they discover complex visual features, often difficult to interpret and not necessarily intuitive, during their learning process. Because the RDT approach can be implemented using image-computable DNNs, this approach does not need to make representational assumptions. In principle, DNNs implementing RDT can make predictions about which visual features an optimal capacity-limited agent should develop.

Relation to previous optimal models

Like other models of optimality, such as Bayesian models, RDT models can be regarded as "ideal observers." In general, ideal observers are useful for at least two reasons. First, they provide a benchmark by defining optimal task performance. By comparing human performance with this optimal benchmark, we can reason about whether human performance is good, poor, or something in between. Second, if we make the assumption that human observers are (approximately) efficient—that is, they understand the nature of their task and they know and use the task-relevant statistical regularities of their visual environment—ideal observers can make predictions that researchers can use as working hypotheses about people's visual representations and operations.

Despite these commonalities, Bayesian and RDT models rely on analogies to different kinds of human-engineered systems. For instance, consider a particular class of Bayesian models, namely, Bayesian-consistent signal detection models. These models are analogous to sensing devices, which are unlimited capacity but whose measurements are corrupted by some degree of internal noise. For example, a large array of identical, noisy detectors could be applied to a search display in order to detect the presence of a target. As the number of locations where the target could appear is increased, the amount of information transmitted by the sensor array (as measured in bits) increases unboundedly. Consequently, task performance may still remain high (though performance could worsen due to the risk of false alarms; see Huang & Pashler, 2005).

The RDT framework, by contrast, is based on an analogy to a noisy communication channel, which is a capacity-limited device. All information-transmitting devices (such as neural circuits) have a maximum bit rate associated with them. This maximum bit rate (or capacity) can be approached using carefully designed codes (including neural codes). When RDT models assume that human observers are efficient, they assume that people achieve information-transmission rates close to their capacity.[2] With this assumption, experimenters can use RDT models to predict what information in visual stimuli is being prioritized by experimental subjects.

Because communication channels are limited capacity, their accuracy declines as informational load is increased. To continue the example above, as the number of possible target locations in a display is increased, the information-processing requirements (in bits) will generally increase, and therefore detection accuracy will decline monotonically as soon as the channel capacity is exceeded.

Importantly, the signal detection and communication channel analogies embody different assumptions about the nature of neural noise. Bayesian models assume there is some fixed amount of sensory noise inherent to the system. This noise hinders task performance but imposes no capacity limit. In contrast, RDT models assume that sensory noise is largely the *result* of capacity limits. In other words, as the amount of processing resources (i.e., bits) devoted to a task increases, the noise in the sensory response decreases. Moreover, RDT assumes we can *design* the noise distributions to be optimally suited to our needs (through evolution, development, or learning). For example, for optimal capacity-limited systems, some kinds of correlated noise may actually be advantageous, and in many cases

the optimal sensory response should be statistically biased (Bates et al., 2019). These properties contrast with i.i.d. (independent and identically distributed) assumptions typical in Bayesian models.

To this point, we have compared conventional Bayesian models, which lack limitations on information processing, with RDT models, which include processing constraints. Admittedly, however, there are Bayesian models in the vision sciences literature that include information-processing limitations. They may include, for example, a sensory stage, which is designed to have limitations resembling those of humans, followed by a Bayesian decision-making stage, which computes an optimal behavioral response based on the output of the sensory stage.

To us, there are at least two drawbacks to these types of models. First, the portions of these models with information-processing limitations tend to be task or domain specific. That is, they make representational or processing assumptions (for example, assumptions about the response profiles of visual neurons) that depend on the specific task to which the model is applied. Therefore, models for different tasks or domains often have little in common. Second, these models separate processing limitations and optimization. Due to this separation, these models do not address the optimal allocation of limited processing resources.

In contrast, RDT successfully addresses both of these points. In regard to the former, RDT constitutes a completely task-independent baseline for achievable performance, and this result follows from the theorems of information theory. Its task independence is what allows a RDT-based modeling framework to be applied to a broad range of perceptual and cognitive domains. In regard to the latter, because RDT comprises a constrained optimization problem with capacity as its constraint, it computes the optimal allocation of limited processing resources.

Which approach—Bayesian or RDT—is best for understanding human visual perception? When using simple stimuli and tasks, such as those commonly found in the vision sciences literature, Bayesian and RDT models often make similar predictions. However, an increasing number of researchers are studying more realistic scenarios. Given the complexity of these scenarios, and given experimental evidence that perception can flexibly allocate its limited resources in a sensible (perhaps optimal) way, we argue that RDT models provide a particularly promising direction for future research.

Uncued visual search

This section reports the results of an experiment in which subjects were asked to perform an uncued visual search task. It also describes classes of RDT and unlimited-capacity Bayesian models and reports the fits of these models to our experimental data.

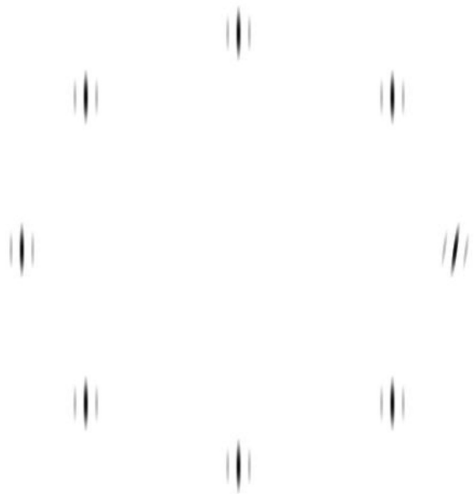

Figure 1. Example stimulus from the one-target condition. In the two-target condition, the two targets were always 180° apart, and the direction of the tilt for each target was chosen randomly.

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,[3] whereas targets were tilted a small fixed amount in the clockwise or counterclockwise direction relative to vertical. (We chose 7.5 degrees based on pilot data to achieve approximately 70–80% correct on single-target search.)

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 $8 \times 2 = 16$ equiprobable unique stimuli in the one-target case, and in the two-target case, there were also $4 \times 4 = 16$ equiprobable unique stimuli. Figure 1 shows a sample stimulus from the one-target condition. Subject responses in the two-target condition were constrained such that it was impossible to choose two locations that were not 180° apart. Subjects were also informed of this constraint in the instructions.

In the two-target case, the placement constraint for the second target was introduced to disincentivize

anticipatory saccades away from the stimulus center and toward 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). The response screen contained initially blank circles surrounding all the previous Gabor locations. When the subject clicked on a blank location, a right-tilted line immediately appeared that extended the diameter of the circle. If the subject clicked on the same location a second time, the line was immediately replaced with a left-tilting one. If the subject clicked on a location for a third time, the location was immediately made blank again and the cycle reset. The degree of tilt of the lines matched that of the Gabors. In the two-target condition, selection of the first location did not automatically make a line appear in the obligatory second location. However, clicks on invalid secondary locations (i.e., ones that were not 180 degrees away from the initially selected location) were ignored. In both conditions, subjects were paid $6.00 to complete 500 trials. Most subjects took approximately 20–30 min to complete the task. Below we analyze only the last 200 trials.

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

$$p(y_\theta, y_{loc}|\hat{x}) = \sum_x p(y_\theta, y_{loc}, x, \hat{x})/p(\hat{x}) \quad (3)$$

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.
- $\hat{x}$ is a model's (noisy) sensory response, code, or representation of x. For tractability, our models assumed restrictions on the space of possible vectors. In particular, our models placed zero probability on sensory responses that would correspond to the wrong number of targets (i.e., greater than 1 in the one-target condition and not equal to 2 in the two-target condition) or tilt values other than the three encountered in the experiment (vertical, left, right). For simplicity, we parameterized $\hat{x}$ in the same way as x. That is, $\hat{x}$ was also a vector of eight dimensions, where each element was a tilt value.

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

$$p(x_s|x) = \frac{\prod_i^N e^{\frac{1}{\sigma}\cos(x_s^{(i)} - x^{(i)})}}{\sum_{x'} \prod_i^N e^{\frac{1}{\sigma}\cos(x'^{(i)} - x^{(i)})}} \quad (4)$$

where i indexes over items in the display, and

$$\hat{x} = \arg\min_{x'} \mathbb{E}_{p(x|x_s)}\mathcal{L}(x, x'). \quad (5)$$

That is, in the Bayesian models, sensory measurement x_s has a discretized von Mises distribution (again, for tractability and to be consistent with the RDT models) where noise is i.i.d. between items i, and sensory response $\hat{x}$ 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 4). The loss function for both was given by

$$\mathcal{L}(x, \hat{x}) = \|\hat{x} - x\|^2. \quad (6)$$

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 3; 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 and 5 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:

$$\mathcal{L}_{SE}(x, \hat{x}) = \frac{\|\hat{x} - x\|^2}{\max_{x'} \|x' - x\|^2} \quad (7)$$

$$\mathcal{L}_{loc}(x, \hat{x}) = \frac{\sum_{n=1}^{N} \mathbb{1}(x_n, \hat{x}_n)}{N_{targets}} \quad (8)$$

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 1 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

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{SE} + \alpha\mathcal{L}_{loc} \quad (9)$$

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

$$L(\phi) = \prod_t p_{y_\theta, y_{loc}|x}\left(x_{resp}^{(t)} | x^{(t)}\right) \quad (10)$$

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 et al., 2017; Wozny et al., 2010; Craig, 1976).

Visualizing the optimal noisy channel: Figures 2 and 3 qualitatively visualize basic predictions of the optimal lossy channel. Figure 2 shows how channel output probabilities $p(\hat{x}|x)$ vary as a function of capacity. In particular, more probability mass is concentrated on $x = \hat{x}$ as channel capacity is increased. Figure 3 shows how the same probabilities vary as a function of what information is emphasized in the loss function for a fixed capacity. The top row corresponds to $\alpha = 0$, which penalizes the squared distance between the stimulus and response in angle space and thus assumes subjects care about both tilt and location. The bottom row corresponds to $\alpha = 1$, which only penalizes location errors and thus assumes subjects only care about location. Intermediate values of α would interpolate between the top and bottom rows.

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 4). 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 4, 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

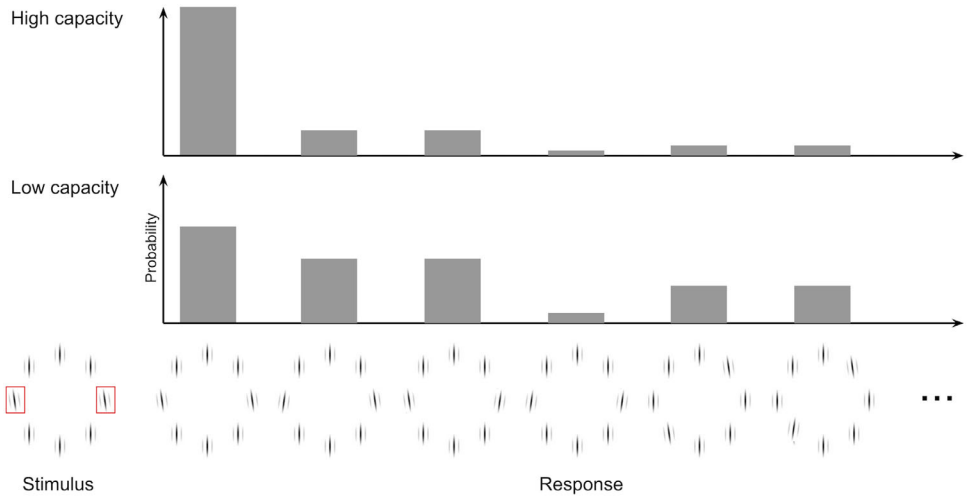

Figure 2. Cartoon visualizing how sensory response probabilities $p(\hat{x}|x)$ vary as a function of capacity. Bar plots depict the probabilities for a subset of possible responses $\hat{x}$ given the stimulus x depicted on the left, where images along the horizontal axis depict the members of that subset. As capacity increases, more mass is concentrated on the correct sensory response ($x = \hat{x}$). Red rectangles indicate target location and were not present in the experiment. Gabor spacing in the figure is more condensed than in the real stimuli.

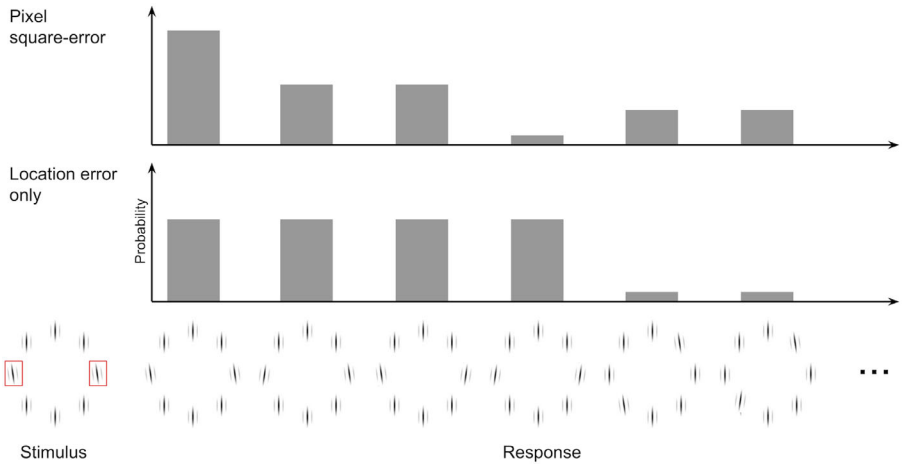

Figure 3. Cartoon visualizing how sensory response probabilities $p(\hat{x}|x)$ vary as a function of what information is emphasized by the loss function. Bar plots depict the probabilities for a subset of possible responses $\hat{x}$ given the stimulus x, where images along horizontal axis depict the members of that subset. Top row corresponds to $\alpha = 0$ (equivalent to pixel square error). Bottom row corresponds to $\alpha = 1$ (no penalty for tilt errors, only location errors). In this case, the first four bars are equal because the target locations are the same as in the stimulus. Red rectangles simply indicate target location and were not present in the experiment. Gabor spacing in the figure is more condensed than in the real stimuli.

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.

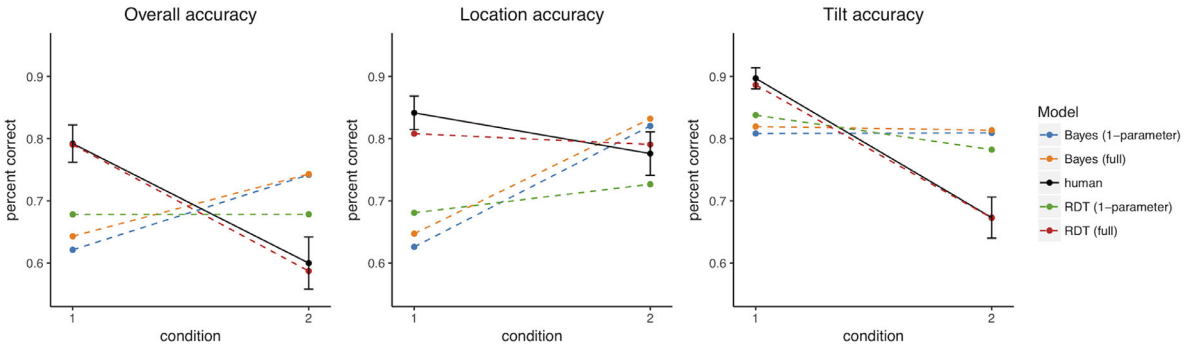

Figure 4. Model predictions and experimental data (overall accuracy, location accuracy alone, and tilt accuracy alone). The percentage of correct responses is plotted in each condition and compared to two versions of model predictions. The simpler version of the model has only one fitted parameter (capacity or sensory noise magnitude). The full model includes three fitted parameters.

	$\mathcal{C}$	σ	α	τ	Log likelihood (all trials)	Log likelihood (one condition)
RDT (both)	3.1	—	0.36	0.81	−22,435	—
RDT (one-target)	3.1	—	0.38	—	—	−85,77
RDT (two-target)	2.4	—	0.24	0.00001	—	−13,776
Bayesian (both)	—	0.145	$\geq$0.7	$<$0.5	−23,545	—
Bayesian (one-target)	—	0.126	$\geq$0.0	$>$0.0	—	−8,890
Bayesian (two-target)	—	0.172	$\geq$0.35, $\leq$ 0.65	$>$0.5	—	−13,934

Table 1. Inferred model parameter values for full models.

	$\mathcal{C}$	σ	Log likelihood (all trials)	Log likelihood (one condition)
RDT (both)	2.6		−23,862	—
RDT (one-target)	3.0		—	−8,939
RDT (two-target)	2.3		—	−14,583
Bayesian (both)	—	0.148	−24,032	—
Bayesian (one-target)	—	0.126	—	−8,890
Bayesian (two-target)	—	0.181	—	−14,179

Table 2. Inferred model parameter values for one-parameter models.

The middle panel of Figure 4 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 all conditions, as it is unrealistic to presume that, for instance, channel capacity or sensory noise magnitude changes across conditions. For the full RDT model, we found that $\mathcal{C}$ 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 intersubject variance in the two-target condition and performance was nonnormally 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

In summary, subjects in our experiment performed an uncued visual search task when stimuli contained either one or two targets. Fits of Bayesian and RDT models to the experimental data show that Bayesian models overestimated 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 in our uncued search experiment 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.

Cued visual search

In this section, we model behavioral data from experiments using cued visual search tasks. These experiments were originally presented in Shimozaki et al. (2003) and Shimozaki et al. (2012). In both experiments we will analyze, subjects were presented with displays containing multiple locations and were asked to detect the brief appearance of a target at one of the locations. The target only appeared on half of the trials and was formed by adding a "Gaussian disk" to the white-noise background, resulting in brighter pixels at the selected location. The variance of the background noise relative to the disk brightness determined task difficulty. Subjects did not need to report the location of the target. For purposes of analysis, trials were categorized as "valid," "invalid," or "target-absent." Valid trials were trials in which the target was present and occurred at the location that was cued. Invalid trials were trials in which the target was present but occurred at a location that was incongruent with the cue. The studies differed primarily in number of locations, and therefore the details of our model change very little across the two studies.

We chose these two studies because they examine behavior under a range of cueing conditions. While the experiment we analyze in Shimozaki et al. (2003) had just two locations in the display and high cue validity, like many classic cueing studies, Shimozaki et al. (2012) was more complex with eight locations and had a lower cue validity. An "effective setsize" (see below) was established using a set of secondary cues, which informed the subject about the set of locations they could safely ignore. Both the primary and secondary cue locations varied randomly from trial to trial.

In both studies, we had access to summary performance statistics but not the raw subject responses. For this reason, our analyses will remain qualitative, and our primary goal will be to demonstrate both the plausibility and feasibility of a modeling approach based on efficient lossy compression, rather than to adjudicate definitively between Bayesian and RDT predictions in the present cueing experiments. We will leave to future work more extensive and quantitative model comparisons.

We note that the Bayesian models originally presented in Shimozaki et al. (2003) and Shimozaki et al. (2012) were able to achieve closer fits to the empirical data than our own models, because they opt to fit parameters at the individual subject level, while we opt to fit only at the aggregate level. For a visual comparison of our results to theirs, we refer the interested reader to the original studies.

To model how subjects take advantage of the precues on each trial, we employ a crucial assumption: Subjects respond to a cue by changing how they weight the importance of each of the locations in the display, allocating more resources to locations where the target is more likely to occur. In modeling terms, this amounts to the assumption that subjects' visual communication channels are conditioned on the cue. Thus, the set of conditional probabilities $p(\hat{x}|x)$ is instead written as

$p(\hat{x}|x, C)$, where C is a vector representing the cue location(s). The channel conditioned on some C is the channel that has been optimized for a vector of weights $\omega = f(C)$ over locations in the loss function (see below). Intuitively, these weights determine how much errors in transmission are penalized at each location in the display. To be optimal, ω should assign higher weight to cued locations and lower weight to uncued locations, as long as the cues are positively correlated with target locations.

We present two kinds of models below. First, we model subject responses using the same framework as before ("idealized model"). That is, we abstract the stimuli into vectors containing the average pixel intensity at each target location (discretized into bins) and operate on these abstract low-dimensional vectors, rather than the true vector of pixel values. As above, we use the Blahut algorithm to find the optimal lossy channel, as this algorithm is tractable in low-dimensional spaces and guarantees a global optimal solution. Next we will demonstrate the scalability of our approach by modeling the experiment in Shimozaki et al. (2003) at the pixel level, using a deep neural network that approximates optimal lossy compression ("image-computable model").

Idealized model

Models for the cued search tasks were highly similar to those used with the uncued search experiment. The principle differences are that (i) the decision variable in the model for the cued visual search studies was a binary variable, denoted z, where the probability that z equals 1 is the probability that the target was present in the display at any location, and (ii) the channel distributions were conditioned on the cue C. Specifically, we used the joint distribution $p(x, \hat{x}, z|C) = p(z)\ p(x|z, C)\ p(\hat{x}|x, C)$, where x and $\hat{x}$ are the stimulus and sensory response, respectively, and z is the binary variable representing target presence, which is independent of the cue. The decision rule is given by

$$p(z|\hat{x}, C) = \sum_{x} p(z, x, \hat{x}|C)/p(\hat{x}|C). \quad (11)$$

For the purposes of the visual communication channel $p(\hat{x}|x, C)$, the space of possible stimuli was discretized. A stimulus x was represented by an N-dimensional vector, where N is the number of potential target locations. The stimulus value corresponding to a location could be one of K evenly spaced values ranging from 0 to ν, where ν represented the strongest possible signal. For nontarget locations, this value was set to 0. It was nonzero for the target location. The sensory response $\hat{x}$ was represented in the same way. The value of K (determining the number of discrete bins; 50 in Shimozaki et al. [2003], 8 in Shimozaki et al. [2012]) was chosen to strike a balance between realism and computational tractability.

As in the uncued experiment, we assumed that at decision time, subjects employed veridical knowledge of both the task and their sensory response distributions but that their sensory responses were not necessarily optimal for the task. Accordingly, computation of $p(z|\hat{x}, C)$ (computed during the decision stage of the model) assumed x and z to follow the true, experimental stimulus prior. In contrast, $p(\hat{x}|x, C)$ (computed during the sensory stage) was the solution to a constrained optimization problem, as defined by RDT. This optimization problem requires the specification of a visual prior distribution and a loss function, defined next.

Visual prior. For the visual prior distribution $p(x)$, we assumed a uniform distribution over all possible target-absent and target-present stimuli.

Loss function. The loss function was a weighted squared error:

$$\mathcal{L}(x, \hat{x}, C) = \sum_{i} \omega_i (x_i - \hat{x}_i)^2 \quad (12)$$

where i indexes over locations, ω_i is the weight given to location i, $\sum_i \omega_i = 1$, and $\omega_i \geq 0 \forall_i$.

The values of the weights $\omega = \{\omega_i\}$ were computed on each trial based on the trial's primary and secondary cue locations (again, see below). Consequently, these weight values can be interpreted as specifying an allocation of attention to different locations. We assumed that people are only able to reallocate some of their cognitive resources to the cued locations on each trial, and we captured the degree of reallocation in the parameter $\epsilon \geq 0$, which controls the amount of smoothing and fully determines ω. Precisely,

$$\hat{\omega}_i = \begin{cases} c_v \epsilon + 1 & \text{if } i \in C_1 \\ \frac{(1-c_v)\epsilon}{|C_2|} + 1 & \text{if } i \in C_2 \\ 1 & \text{otherwise} \end{cases}$$

$$\omega_i = \hat{\omega}_i / \sum_j \hat{\omega}_j \quad (13)$$

where $c_v = 0.5$ is the primary cue validity; i and j index over locations; C_1 and C_2 are the set of primary and secondary cue locations, respectively; and $|C_2|$ is the number of secondary cues (i.e., the effective setsize minus 1). This smoothing scheme was chosen such that weights approach their task-optimal values as ϵ goes to infinity and become uniform over all locations as ϵ goes to zero. These equations ensure that the weights are strictly positive and sum to 1, which are necessary conditions for stability during optimization.

Image-computable model

Modeling optimal lossy compression at the level of image pixels is computationally intractable due to the cost of solving the RDT-constrained optimization problem. Therefore, it is necessary to use approximate methods. Here, we used a deep neural network architecture (see Bates & Jacobs, 2020) to approximate $p(\hat{x}|x, C)$, where x comprised real stimulus images (i.e., pixel values). Our architecture can be considered a variant of β-variational autoencoders (β-VAEs; see Kingma & Welling, 2013; Higgins et al., 2017) and comprises two main components.

First, the encoder network q_ϕ (with parameters ϕ) is stochastic and learns a mapping from pixel values x to a probability distribution over abstract, latent vector representations, denoted y. It is assumed that $p(y|x, C)$ is a normal distribution with diagonal covariance matrix, and the network produces the mean and variance for each element of y. Given these means and variances, samples y are drawn. These samples can be viewed as analogous to neural responses.

Next, a deterministic decoder network f_θ (with parameters θ) is used. Unlike typical decoders, our decoder does not attempt to reconstruct stimulus x. Instead, it attempts to reconstruct a low-dimensional representation of x. Let $\overline{x}$ denote a vector of average pixel intensities in each potential target region for a particular stimulus. Then, the decoder network maps samples y (obtained from the encoder network) to corresponding estimates of $\overline{x}$, denoted $\hat{x}$. Both encoder and decoder are trained using standard gradient descent methods.

Finally, to approximate $p(z|\hat{x}, C)$, we train an additional decision network that takes an input $\hat{x}$ from the first (i.e., pretrained β-VAE) network and outputs the probability of target presence.

It is important to highlight a conceptual difference between this neural network model and the idealized (Blahut algorithm) model presented above. While in the idealized model, we took $\hat{x}$ to be the output of the channel, here we introduce an intermediate variable y. Mathematically, these assumptions are equivalent (by the information-processing inequality since f_θ deterministically maps from y to $\hat{x}$), but the inclusion of the intermediate variable y, which is an abstract latent code, affords a more direct analogy to neural implementation.

In general, directly implementing a rate-distortion objective in neural networks is a difficult problem, since estimating mutual information is notoriously hard. Here, we have chosen β-VAEs as the backbone of our approach because they can be interpreted as approximating a rate-distortion objective (Alemi et al., 2016, 2017; Ballé et al., 2016; Burgess et al., 2018) and are easy to implement. While their solutions are in general not strictly rate-distortion optimal, they achieve state-of-the-art performance in many applications and exhibit a systematic trade-off between rate (or capacity) and distortion (or error), and therefore serve our present purposes.

In β-VAEs, "β" refers to an optimization parameter that controls the trade-off between information content of the network and distortion or accuracy on the task objective. In our case, β controls the information rate of latent code y. A larger value for β forces y to be less informative, and therefore training loss will generally be higher. Conversely, a lower β value puts less constraint on y, and thus the network can generally be more accurate on the task.

Specifically, the objective is given by

$$\mathcal{L}(x, \hat{x}, C) = \sum_i \omega_i(\overline{x}_i - \hat{x}_i)^2 - \beta\,\mathrm{KL}[p(y|x, C)\,||\,p(y)]. \tag{14}$$

In this equation, the first term is analogous to the loss function used by the idealized model (Equation 12), where the weights $\omega = \{\omega_i\}$ are based on the cue C in the same manner as in the idealized model (Equation 13). The second term acts as a capacity regularizer (or constraint on information capacity) by controlling the difference between the posterior distribution $q_\phi(x) = p(y|x, C)$ over latent variable y and its prior distribution $p(y)$. Specifically, this term is calculated as the Kullback-Leibler (KL) divergence between the probability of the compressed neural code y given the input x and cue C (the conditional channel distribution) and the prior (marginal) probability of the code. Intuitively, if the code is constrained to have a posterior distribution near its prior, then it cannot store as much information about the input, and thus its information rate will be limited. The marginal $p(y)$ is assumed to be a spherical Gaussian (i.e., $p(y) = \mathcal{N}(0, I)$), as is common in the VAE literature. This assumption affords an analytical expression for the KL divergence term and its gradients. During training, a β-VAE uses an iterative stochastic gradient descent algorithm, often requiring thousands or millions of iterations, to adjust its weights (ϕ and θ) to minimize Equation 14.

Network details. The encoder network q_ϕ comprised two standard convolutional layers, with 32 and 64 filters in the first and second layers, respectively. The second layer fed into a fully connected layer (2,000 units, rectified-linear activation), which in turn fed into the latent layer representing $p(y|x, C)$ (500 units for the mean values of the latent variables and 500 units for their variances). All convolutional layers had a stride of 2 and kernel size of 3 and used rectified-linear activations. The decoder network f_θ mapping from y

to $\hat{x}$ was a fully connected layer with linear activation function. The decision network mapping from $\hat{x}$ to binary decision z (denoting model response, either target present or target absent) was a two-layer multilayer perceptron where the hidden layer had 100 units and rectified-linear activations, and a one-unit output layer with sigmoid activation function to keep output values, representing $p(z|\hat{x}, C)$, between 0 and 1.

Visual stimuli. Training images were 64×64 pixels and were constructed in a two-step process. First, an image was initialized with white (Gaussian) background noise. That is, each pixel value was drawn from a normal distribution with mean zero and standard deviation σ_{noise}. Next, if a target was present, a "Gaussian disk" was added to the image at one of the potential target locations. A Gaussian disk is produced by calculating the probability density function of a two-dimensional Gaussian whose mean coincides with the center of the target location. The value to be added at each pixel is given by that pixel's vector distance from the mean, plugging that vector into the expression for probability density. Stimuli were randomly generated throughout training. Figure 7 depicts two example stimulus images.

Study 1: Shimozaki et al. (2003)

The experiment we analyze from Shimozaki et al. (2003) used displays with just two locations. Subjects saw a precue on every trial, which correctly indicated the location of the subsequent target 80% of the time. The pixels at the target location were lighter (more white) on average than those in other areas of the image. The experimental manipulation was the signal-to-noise ratio (SNR), or roughly how much lighter the target location was than nontarget locations. As the SNR increased, it became easier for subjects to detect the target.

As we did not have access to the raw subject responses, our analyses remain qualitative. In analyses below, our primary goal is to demonstrate the plausibility of a modeling approach based on RDT or lossy compression, rather than to adjudicate definitively between Bayesian and RDT predictions in the present cueing experiment. We demonstrate that (i) the idealized and image-computable models are qualitatively consistent in their predictions, suggesting that the image-computable model is a reasonable approximation of the idealized model; (ii) both models capture important trends in how averaged human hit- and false-alarm rates varied with SNR; and (iii) our models suggest subjects likely allocated some but not all of their attention to the cued location. We also find, however, that our idealized model does not seem to fit the exact shape of the empirical cueing-effect curves, suggesting further investigations are needed to diagnose the reasons for this mismatch. Our image-computable model results in a somewhat better qualitative match.

Modeling details. In both models, we assumed people's visual channels $p(\hat{x}|x, C)$ were designed to communicate a larger range of target intensities than seen in the experiment. The maximum intensity that a channel can communicate is denoted ν, ν^* denotes a target intensity on a particular trial, and ν^*_{max} denotes the largest target intensity in the experiment.[4]

In analyses using the idealized model (in which representations of stimuli were discretized), we assumed that stimulus symbols to the sensory communication channel represented the mean pixel intensity at the two potential target locations. Parameter values were chosen or inferred as follows. First, we set $\nu = 50$, a large enough value so that the discretization of the stimulus space was a reasonable approximation of the underlying continuous space. We regard $\mathcal{C}$ and ν^*_{max} as the only free parameters of the model, and we set $\nu^*_{max} = 8$ and $\mathcal{C} = 1$ bit.[5] The parameter ν^*_{max} determined the range of intensities that the channel is designed to handle compared to what was seen in the experiment (for our choice of parameter values, this ratio was $8/50 = 0.16$). The values for ν^*_{max} and $\mathcal{C}$ were chosen by visual inspection so that the idealized model's predictions coarsely matched the overall shape of the experimental data. Given these model parameter settings, we then varied ϵ (used to calculate location weights ω in Equation 13) to infer how much subjects reallocated their attention in response to cues.

The image-computable model used continuous rather than discrete variables. Therefore, we set $\nu = 1$ arbitrarily and found reasonable results when setting $\nu^*_{max} = 0.4$ and $\beta = 0.0025$.

In both models, the true decision (stimulus-present) prior used during inference, $p(z|\hat{x}, C)$, assigned a probability of 0.5 to the target-absent stimulus, a probability of 0.4 to the event in which the target occurs at the cued location, and a probability of 0.1 to the event in which the target occurs at the uncued location.

Results

Idealized model. To assess the idealized model, we qualitatively compared its predictions to the summary data reported in Figure 7 of Shimozaki et al. (2003) (replotted here in Figure 5, left panel). These data present hit rates within valid and invalid trials, respectively, and false-alarm rates. Of greatest interest is how cueing effect varies with SNR. The cueing effect is the difference between hit rates for the valid and invalid trials.

The primary goal of our analysis is to assess approximately how much subjects likely reallocated their attention in response to cues on average. In the middle and right panels of Figure 5, we show model predictions for $\epsilon = 0.5$ (corresponding to

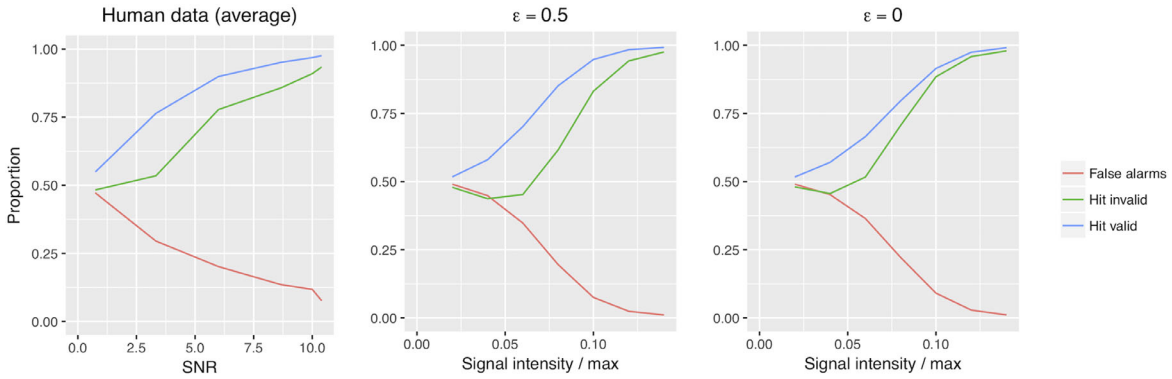

Figure 5. Comparing human performance data (hit rates and false-alarm rates) from Shimozaki et al. (2003) to idealized RDT models for two representative parameter settings. Human data are averaged over three subjects.

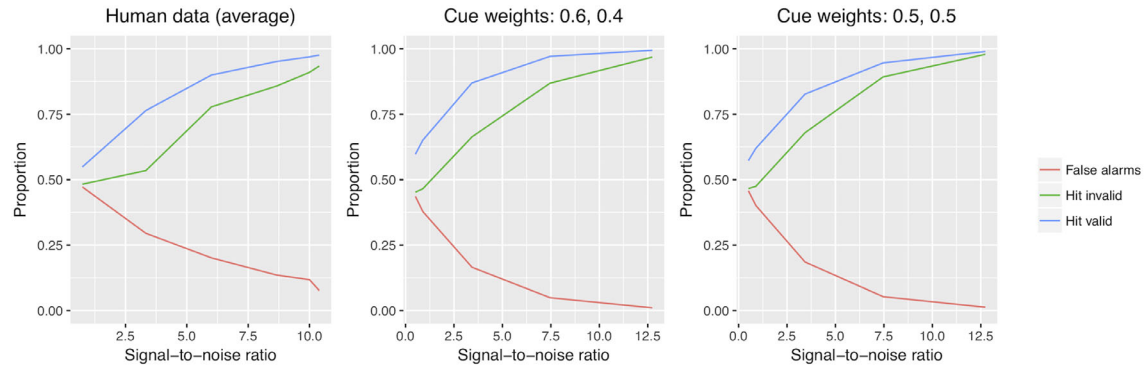

Figure 6. Comparing human performance data (hit rates and false-alarm rates) from Shimozaki et al. (2003) to image computable RDT models for two representative parameter settings. Human data are averaged over three subjects.

$\omega = [0.56, 0.44]$, where the first number is the weight on the cued location) and $\epsilon = 0$ (no reallocation). We found that the model exhibited a cueing effect (hit-rate valid minus hit-rate invalid), even when there was no reallocation in response to the cue ($\epsilon = 0$). This is expected behavior, since the number of invalid-cue trials was less than the number of valid-cue trials, so inference under uncertainty should be biased toward the cued location. By visual inspection, we find that an $\epsilon > 0$ provides a better explanation of the data and that the cueing effect becomes clearly larger than the human data for values greater than about 0.5, across the range of SNRs.

Image-computable model. As shown in Figure 6, the results with the image-computable model are similar to those with the idealized model. The left panel replots the human data, the middle panel corresponds to a 60/40 reallocation given the cue, and the right panel corresponds to no reallocation. By visual inspection, we find that the size of subjects' cueing effect as a function of SNR is best accounted for by the degree of reallocation in the middle panel.

Figure 7 visualizes how the behavior of the 60/40 reallocation model varies with SNR. The leftmost panels are the stimulus images input to the encoder, and the remaining panels are outputs of the decoder (i.e., approximately optimal reconstructions $\hat{x}$ of the stimulus x). Each sample image corresponds to an independent sample of $\hat{x}$ conditioned on x. Reconstructions in the top row show that for relatively low SNRs, the latent code y exhibits less certainty about what stimulus was observed. Reconstructions in the bottom show higher certainty. On average, the pixel square error in the cued location (here, left) is lower than in the uncued location in this model. Since the cued location is left and the target location is right, both rows show an invalid-cue trial.

Study 2: Shimozaki et al. (2012)

Here, we model the single-fixation condition from Shimozaki et al. (2012). This experiment used eight locations and two kinds of cues—a primary cue that cued a single location and a set of secondary cues

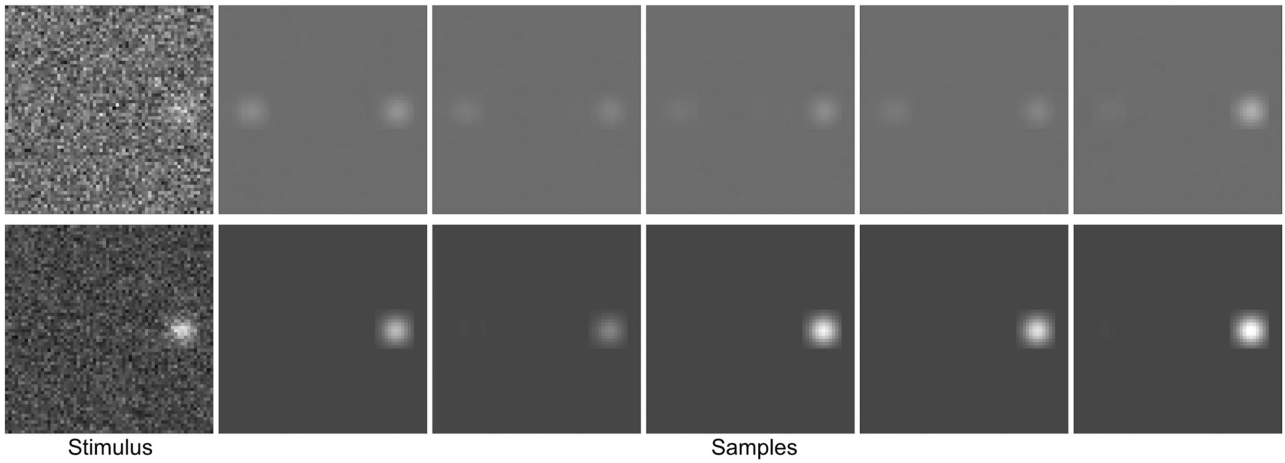

Figure 7. Visualization of neural network model behavior in a two-location cueing task. On the trial illustrated here, the left location was cued, but the target appeared at the right location (i.e., this is an invalid-cue trial). Top and bottom rows correspond to low and high signal-to-noise ratios, respectively. Leftmost panels are the stimulus image and remaining panels are random samples from a decoder network (trained separately for visualization purposes) that produces maximum likelihood images $\hat{x}$ given noisy latent representation y. Pixel values were normalized separately for each row to span the entire range of values (from 0 to 255).

that established the "effective setsize" by indicating the subset of locations where it was possible for the target to appear. The primary cue was 50% valid and the secondary cues (numbering 1, 2, 4, or 7) were collectively 50% valid. That is, the primary cue indicated the target location half the time, and the set of secondary cues indicated the set of alternative locations where the cue could appear the other 50% of the time. For example, if the number of secondary cues was two, then there were a total of three locations where the target could appear, with probability 0.5 that it appeared at the primary cued location and probability 0.25 that it appeared at either of the remaining locations. The locations of the primary and secondary cues were chosen randomly on each trial and thus could have any nonoverlapping configuration.

Calculating the true prior. This calculation was slightly different from the other cueing experiment because there were secondary cues in addition to the primary cue. For each possible arrangement of cues (i.e., which locations were cued by primary and secondary cues), we assigned 0.5 probability to the stimulus value for which ν^* occurred at the primary cue location and $0.5/|C_2|$ probability to each stimulus value for which ν^* occurred at one of the secondary cue locations. All other stimuli were assigned zero probability. We set ν^*/ν to be as close as possible to the value in the previous experiment. (The ratio could not be exactly equal between studies because the denominator ν differed.)

Fitting model to data. As we did not have access to the raw subject response data, we fit the models so as to make similar hit and false-alarm rate predictions to the aggregate data. We minimized the sum of absolute errors with respect to the valid-trial hit rates, invalid-trial hit rates, and false-alarm rates (averaged over subjects) with equal weighting on each data point. As above, we compared results for $\epsilon = 0.5$ and $\epsilon = 0$. For both ϵ values, we searched for the optimal $\mathcal{C}$ using R's `optim` function to conduct model fits. Note that we should not expect the capacity in this study to match the previous study. This is because in the two-location study, we only measure how much information people stored about two locations in the display, instead of eight. Thus, what we find in the first study should be a more severe underestimate of capacity.

Results

Figure 8 shows a comparison of averaged human data (four subjects; replotted from Shimozaki et al., 2012) and our idealized model's predictions for $\epsilon = 0.5$ and $\epsilon = 0$. As was the case with the previous study, we find a cueing effect even with no reallocation. Importantly, we again find that the experimental data are better accounted for by an amount of reallocation given by $\epsilon = 0.5$. Prediction error was lower when $\epsilon = 0.5$ (0.35) compared to when $\epsilon = 0$ (0.41). We find $\epsilon = 0.5$ to yield quite good predictions, despite the fact that it was chosen based on results from the previous study. In fact, we refit our model treating ϵ as a free parameter along with capacity but found a very similar value (0.51). Our estimate of capacity was 4.83 bits, which was larger than our estimate of 1 bit from the previous study, as expected. Overall, we find that the idealized model does an adequate job of capturing changes in hit and false-alarm rates and cueing effect

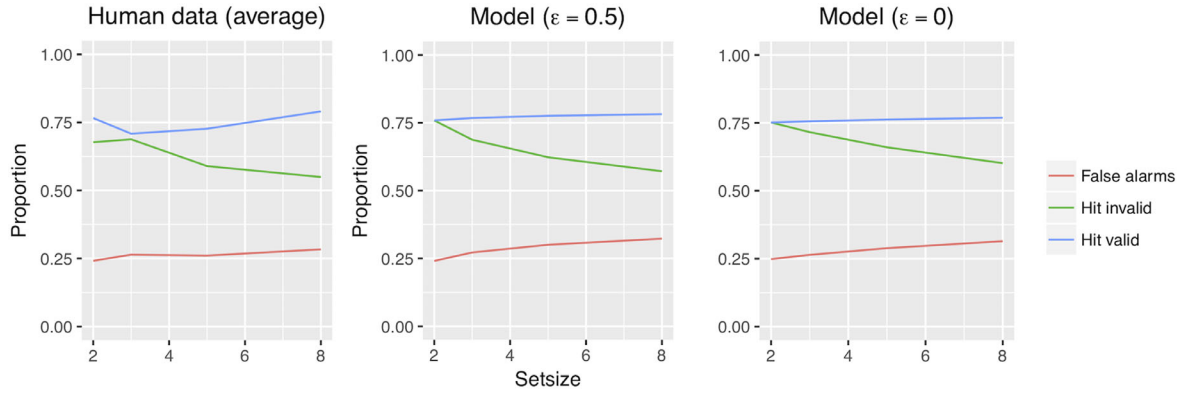

Figure 8. Comparing performance (hit rates and false-alarm rates) between humans and RDT models from the single-fixation condition in Shimozaki et al. (2012). Human data are averaged over four subjects.

size as a function of setsize, despite only having a single free parameter (capacity $\mathcal{C}$).

Discussion

In this section, we presented an idealized RDT model and used it to account for subjects' responses in two experiments using cued visual search tasks. We also presented an image-computable RDT model and used it to account for data from the first experiment. Our results indicate that, qualitatively, these models provide adequate accounts, indicating that capacity constraints may have played significant roles in limiting subject performance. In addition, our results with the image-computable model suggest that deep neural networks may provide useful models of capacity-constrained visual performance in real-world settings (i.e., using pixel values instead of simplified features handcrafted by researchers).

While the authors of the studies we just examined fit Bayesian models to individual subject data, here we have opted to analyze subject averages. Shimozaki et al. (2012) explained the subject-level variability in their eight-location experiment by allowing the decision criterion to vary by both subject and setsize. Similarly, in Shimozaki et al. (2003), the authors fit their models within-subject and allowed d' sensitivity to vary as a linear function of SNR in addition to fitting a parameter for how much the prior was weighted. Here, we have opted not to try to model decision-making idiosyncrasies at the levels of subject and condition.

In comparing the idealized and image-computable models on Shimozaki et al. (2003), we found that the image-computable model qualitatively captured some aspects of the data better. For example, the valid-trial hit rates appeared strictly concave in shape in the image-computable model, whereas in the idealized model, there was an inflection point. One possibility is that this difference reflects suboptimalities in the neural network model. If so, then future work could investigate whether these suboptimalities are a plausible explanation of people's behavior. More generally, as compared to the idealized model, our image-computable modeling framework allows for more easily exploring sources of suboptimality in perceptual tasks, as they can be viewed as instantiating a "process" model, which seeks to identify specific biological mechanisms driving behavior (McClelland et al., 2010).

General discussion

We motivated our experimental and modeling work by noting the large diversity of experimental and theoretical approaches to the study of visual attention in the vision sciences literature and arguing that this diversity arises, at least in part, from the field's inability to unify differing theoretical perspectives. In particular, the field has been hindered by a lack of a principled theoretical framework for simultaneously thinking about both optimal (Bayesian) attentional processing and capacity-limited attentional processing. Here, we have presented such a framework based on rate-distortion theory and optimal lossy compression.

Taken as a whole, our results suggest that visual attention may be *both* capacity limited *and* approximately optimal. That is, performance in any task is limited by capacity, but vision's limited computational resources are efficiently allocated, leading to behavioral signatures that can often look similar to a noisy Bayesian observer. In the uncued search experiment, our RDT models provided a better overall explanation of the data than the Bayesian models, because their capacity limits result in more sensitivity to stimulus entropy. In the cued search experiments, we found our models explained the hit and false-alarm rates well with only one or two free parameters. Our results

suggest that subjects likely reallocated some of their computational resources in response to cues but did not completely adapt to the task.

In our studies here, we made relatively few assumptions specific to human vision. But future work could seek to incorporate known constraints of the visual system into RDT models. For example, we initially explored a version of the cued-search model that included a smoothness constraint on spatial reallocation of attention. This constraint controlled how "jagged" or discontinuous the spotlight of attention could be. Higher smoothing meant that adjacent locations needed to be more similar to each other, resulting in a more smooth and continuous spotlight. Another constraint we explored was a capacity limit on top-down executive control of allocation. However, we found these additional model components had little effect on the present results and therefore omitted them.

Finally, an important conceptual contribution of our work is the idea of a "conditional" communication channel. We employed this idea to model attentional shifts in response to cues, but the idea can be applied to any kind of attentional shift. For example, someone may wish to reallocate their attention to color features within a scene at one moment but later attend to shape features or particular locations (Ehinger et al., 2009). A key strength of our idea is that it can be easily incorporated into image-computable models, like those we presented here and previously in Bates and Jacobs (2020) to model pop-out effects in search. Image-computable models of attention have the potential to be extremely fruitful because such models are capable of implementing nonlinear image filters, mathematical constructs borrowed from electrical engineering that have been highly productive both in understanding computations in the visual system and designing computer vision systems (Olshausen & Field, 1996; Carandini et al., 2005; Simoncelli & Olshausen, 2001; Torralba et al., 2010). Importantly, we anticipate that conceptualizing attention in terms of task-optimized (nonlinear) filters will allow for more concrete, computational hypotheses about the mechansims distinguishing different kinds of attentional allocation (e.g., spatial versus feature-based attention; see Galashan & Siemann, 2017).

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

Acknowledgments

The authors thank Miguel Eckstein and an anonymous reviewer for their helpful comments. The first author was supported by an NSF NRT graduate training grant (NRT-1449828) and an NSF Graduate Research Fellowship (DGE-1419118).

Supported by an NSF research grant (DRL-1561335). A preliminary version of a portion of this work appeared in a paper in the *Proceedings of the 42nd (2020) Annual Conference of the Cognitive Science Society*.

Commercial relationships: none.
Corresponding author: Robert A. Jacobs.
Email: rjacobs@ur.rochester.edu.
Address: Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA.

Footnotes

[1]Some researchers would consider our two-target condition to fall into a separate category of task, known as "dual tasks" (e.g., Han et al., 2003; Liu et al., 2009; Sperling & Melchner, 1978), in which people's performance is sometimes found to be capacity limited. However, we argue that our modeling paradigm provides a general framework that easily incorporates a wide range of tasks, without the need for significant modifications for seemingly special classes of tasks.

[2]Note that RDT itself does not specify the computations needed to approach the rate-distortion bound. Furthermore, exactly achieving the bound requires infinite code-block lengths, something that is infeasible for real systems, which cannot afford to wait for infinite time to receive a message. Nonetheless, the decades since the theorems of information theory were first presented have seen remarkable progress in finding practical coding schemes that operate surprisingly close to the bound (Cover & Thomas, 2006). However, more efficient coding schemes generally come at the cost of requiring greater compute when mapping to and from code space. Thus, an important but challenging target of research going forward should be to determine how close to the bound neural codes can feasibly operate.

[3]Gabors patches were generated using a standard two-dimensional Gabor filter that was rectified so that values below zero were set to zero, where white was assigned to zero and black was assigned to the maximum value.

[4]If x was a vector of pixel values, SNR was defined as $\left(\sum_i x_i^2/\sigma^2\right)^{1/2}$, where noise pixel values were sampled independently from a zero-mean normal distribution with variance σ^2. Because the idealized model does not use pixel values, SNR for this model was computed using a proxy measure, namely, v^*/v.

[5]Note that the choice of v impacts the choice of capacity. For instance, if v is large, then stimulus entropy will be higher and therefore capacity would need to increase to maintain a fixed accuracy.

References

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck. *arXiv preprint. arXiv:1612.00410*.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2017). Fixing a broken elbo. *arXiv preprint. arXiv:1711.00464*.

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, doi:10.1167/7.13.14.

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

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.

Ballé, J., Laparra, V., & Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint. arXiv:1611.01704*.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review, 127*, 891–917.

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, doi:10.1167/19.2.11.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., . . . Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv preprint. arXiv:1804.03599*.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., . . . Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience, 25*(46), 10577–10597.

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

Carrasco, M., & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 24*(2), 673.

Cover, T. M., & Thomas, J. A. (2006). Elements of information theory (Wiley series in telecommunications and signal processing). USA: Wiley-Interscience.

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

da Silva, F. C., 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.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*(4), 501.

Duncan, J. (1993). Similarity between concurrent visual discriminations: Dimensions and objects. *Perception & Psychophysics, 54*(4), 425–430.

Duncan, J., & Nimmo-Smith, I. (1996). Objects and attributes in divided attention: Surface and boundary systems. *Perception & Psychophysics, 58*(7), 1076–1084.

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, doi:10.1167/11.5.14.

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., Peterson, M. F., Pham, B. T., & Droll, J. A. (2009). Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision Research, 49*(10), 1097–1128.

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, doi:10.1167/2.1.3.

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.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling visual search in a thousand scenes: The roles of saliency, target features, and scene context. *Journal of Vision, 9*(8), 1199, doi:10.1167/9.8.1199.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance, 18*(4), 1030.

Galashan, D., & Siemann, J. (2017). Differences and similarities for spatial and feature-based selective attentional orienting. *Frontiers in Neuroscience, 11*, 283.

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

Gottlob, L. R., Cheal, M., & Lyon, D. R. (1999). Attention operating characteristics in a location-cuing task. *Journal of General Psychology, 126*(3), 271–287.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science, 7*(2), 217–229.

Han, S., Dosher, B. A., & Lu, Z.-L. (2003). Object attention revisited: Identifying mechanisms and boundary conditions. *Psychological Science, 14*(6), 598–604.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., . . . Lerchner, A. (2017). β-VAE: Learning basic visual concepts with a constrained variational framework [paper presentation]. *Proceedings of the 2017 International Conference on Learning Representations, Palais des Congrès Neptune, Toulon, France*, https://openreview.net/references/pdf?id=Sy2fzU9gl.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*(5), 791–804.

Huang, L., & Pashler, H. (2005). Attention capacity and task difficulty in visual search. *Cognition, 94*(3), B101–B111.

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

Jacobs, R. A., & Bates, C. J. (2019). Comparing the visual representations and performance of humans and deep neural networks. *Current Directions in Psychological Science, 28*(1), 34–39.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063, pp. 218–226). Englewood Cliffs, NJ: Prentice-Hall.

Kantowitz, B. H. (1987). 3. Mental workload. In: P. A. Hancock (Ed.), *Human Factors Psychology: Advances in Psychology, Vol. 47* (pp. 81–121). North-Holland, https://doi.org/10.1016/S0166-4115(08)62307-9, https://www.sciencedirect.com/science/article/pii/S0166411508623079.

Kinchla, R. A., Chen, Z., & Evert, D. (1995). Precue effects in visual search: Data or resource limited? *Perception & Psychophysics, 57*(4), 441–450.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint. arXiv:1312.6114*.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446.

Liu, S.-H., Dosher, B. A., & Lu, Z.-L. (2009). The role of judgment frames and task precision in object attention: Reduced template sharpness limits dual-object performance. *Vision Research, 49*(10), 1336–1351.

Lu, Z.-L., & Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research, 38*(9), 1183–1198.

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of General Psychology, 7*(2), 183–188.

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.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Henry Holt.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., . . . Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14*(8), 348–356.

Menneer, T., Cave, K. R., & Donnelly, N. (2009). The cost of search for multiple targets: effects of practice and target similarity. *Journal of Experimental Psychology: Applied, 15*(2), 125.

Moore, C. M., & Osman, A. M. (1993). Looking for two targets at the same time: One search or two? *Perception & Psychophysics, 53*(4), 381–390.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

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 short-term memory research. *Attention, Perception, & Psychophysics, 76*(7), 2158–2170.

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.

Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision Research, 34*(13), 1703–1721.

Palmer, J., Ames, C. T., & Lindsey, D. T. (1993). Measuring the effect of attention on simple visual

search. *Journal of Experimental Psychology: Human Perception and Performance, 19*(1), 108.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40*(10–12), 1227–1268.

Palmer, J., White, A. L., Moore, C. M., & Boynton, G. M. (2020). Divided attention in perception: A unified analysis of dual-task deficits and congruency effects. *bioRxiv*, doi:10.1101/2020.01.23.917492.

Pastukhov, A., Fischer, L., & Braun, J. (2009). Visual attention is a single, integrated resource. *Vision Research, 49*(10), 1166–1173.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3–25.

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, doi:10.1167/12.4.14.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*(1–2), 1–46.

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, doi:10.1167/7.10.1.

Shimozaki, S. S., Eckstein, M. P., & Abbey, C. K. (2003). Comparison of two weighted integration models for the cueing task: Linear and likelihood. *Journal of Vision, 3*(3), 3, doi:10.1167/3.3.3.

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, doi:10.1167/12.6.27.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*(1), 1193–1216.

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

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.

Sperling, G., & Melchner, M. J. (1978). The attention operating characteristic: Examples from visual search. *Science, 202*(4365), 315–318.

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

Srivastava, N., & Vul, E. (2016). Attention modulates spatial precision in multiple-object tracking. *Topics in Cognitive Science, 8*(1), 335–348.

Stroud, M. J., Menneer, T., Cave, K. R., & Donnelly, N. (2012). Using the dual-target cost to explore the nature of search target representations. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 113.

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

Torralba, A., Murphy, K. P., & Freeman, W. T. (2010). Using the forest to see the trees: exploiting context for visual object detection and localization. *Communications of the ACM, 53*(3), 107–114.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 39*(4), 321–359.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, 95*(1), 15.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

VanRullen, R., Reddy, L., & Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience, 16*(1), 4–14.

Vul, E., Alvarez, G., Tenenbaum, J. B., & Black, M. J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In: *Advances in neural information processing systems 22*, pp. 1955–1963.

Wolfe, J., & Pashler, H. (1998). *Attention, chapter visual search*. London, UK: University College London Press.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 419.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Computational Biology, 6*(8), e1000871.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365.

Yildirim, I. (2020). Modeling temporal attention in dynamic scenes: Hypothesis-driven resource allocation using adaptive computation explains both objective tracking performance and subjective effort judgments [Conference presentation abstract]. *42nd Annual Virtual Meeting of the Cognitive Science Society*, https://cognitivesciencesociety.org/cogsci20/papers/0185/0185.pdf.