No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint

Wei Ji Ma

Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA

Wei Huang

Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA

Human ability to simultaneously track multiple items declines with set size. This effect is commonly attributed to a fixed limit on the number of items that can be attended to, a notion that is formalized in limited-capacity and slot models. Instead, we propose that observers are constrained by stimulus uncertainty that increases with the number of items but use Bayesian inference to achieve optimal performance. We model five data sets from published deviation discrimination experiments that varied set size, number of deviations, and magnitude of deviation. A constrained Bayesian observer better explains each data set than do the traditional limited-capacity model, the recently proposed slots-plus-averaging model, a fixed-uncertainty Bayesian model, a Bayesian model with capacity limit, and a simple averaging model. This indicates that the notion of limited capacity in attentional tracking needs to be revised. Moreover, it supports the idea that Bayesian optimality of human perception extends to high-level perceptual computations.

Keywords: multiple-object tracking, Bayesian observer, attention, change detection

Citation: Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision, 9*(11):3, 1–30, http://journalofvision.org/9/11/3/, doi:10.1167/9.11.3.

Introduction

Multiple-object tracking is a leading paradigm for studying visual attention (Cavanagh & Alvarez, 2005), in part because it resembles real-world tasks—think about a basketball player tracking other players on the court, or a driver tracking other cars. Pylyshyn and Storm (1988) demonstrated that humans can track multiple, independently moving objects continuously over several seconds. This work challenged the older notion that attention always has a single focus and raised the question of how many items can be tracked. Tracking performance is often alleged to be limited by the number of items that can be attended to, also called the capacity (Cavanagh & Alvarez, 2005; Hulleman, 2005; Oksama & Hyona, 2008; Pylyshyn & Storm, 1988). This view, in which items are encoded in all-or-none fashion, is very similar to limited-capacity theories of working memory (Cowan, 2001; Luck & Vogel, 1997; Pashler, 1988; Rensink, 2000). In contrast, flexible-resource theories (Alvarez & Franconeri, 2007; Palmer, 1990) claim that a continuous resource gets distributed over all items, with less resource per item translating into greater uncertainty. Errors arise as a consequence of uncertainty, much like in basic psychophysics. Consistent with this view, object speed (Alvarez & Franconeri, 2007) and perceptual learning (Green & Bavelier, 2006) affect tracking capacity. Similar theories have been put forward for working memory (Bays & Husain, 2008; Wilken & Ma, 2004).

A shortcoming of flexible-resource theories has been the lack of a mathematical formulation of how uncertainty affects performance as set size varies. This shortcoming is particularly glaring since limited-capacity models are mathematically clear-cut. Of course, probabilistic approaches such as signal detection theory (Green & Swets, 1966) and Bayesian inference (Knill & Pouget, 2004; Knill & Richards, 1996) have been used extensively to describe the effects of uncertainty on human perception. In these models, an observer probabilistically infers the state of the world from noisy sensory evidence. These approaches have the advantages of being mathematically precise, general, and not needing ad hoc assumptions. Moreover, neural circuits can plausibly implement Bayes-optimal computations (Beck et al., 2008; Ma, Beck, Latham, & Pouget, 2006). However, most probabilistic models are created for relative simple judgments, in which only a single feature of a single stimulus is task relevant (such as in cue integration). Many perceptual decisions, including the tracking task we will study here, instead require the extraction of a global, abstract variable from a constellation of multiple stimuli that are in and of themselves not of interest. Several studies have used Bayesian approaches to understand human perception in such tasks, including causal inference in cue combination (Kording et al., 2007; Sato, Toyoizumi, & Aihara, 2007),

doi: 10.1167/9.11.3 Received June 18, 2009; published October 5, 2009 ISSN 1534-7362

oddity detection (Hospedales & Vijayakumar, 2009), object recognition (Kersten, Mamassian, & Yuille, 2004), and visual search (Nolte & Jaarsma, 1966; Palmer, Verghese, & Pavel, 2000; Rosenholtz, 2001; Vincent, Baddeley, Troscianko, & Gilchrist, 2009).

With this in mind, we develop here a model for optimal performance under uncertainty in an attentional tracking task, where uncertainty increases with set size due to a flexible resource constraint. We will refer to this shorthand as a constrained Bayesian observer, even though the inference process itself is not constrained. We will pit it against five alternative models: (a) an observer with limited capacity K, who perfectly encodes K attended items and does not encode other items, if any, at all (Hulleman, 2005); (b) the slots-plus-averaging model, originally proposed for working memory (Zhang & Luck, 2008), in which K items are attended (and others not at all) but their encoding is noisy; (c) an unconstrained (fixed-uncertainty) Bayesian observer; (d) a Bayesian observer with capacity limit K instead of a resource constraint; and (e) an observer who extracts a global motion signal by averaging over all trajectories. In this paper, we mean by "Bayesian" that the observer takes into account probabilities over variables in making a decision rather than only best estimates, even though the prior distribution might be flat (for a concrete example, see the paragraph below Equation 18).

Theory and methods

Deviation discrimination data

We model published data from a variation of multiple-object tracking that has fewer degrees of freedom and is therefore easier to analyze (Tripathy & Barrett, 2004; Tripathy, Narasimhan, & Barrett, 2007). On each trial, N dots moved from left to right in linear trajectories (Figure 1). Of these, D dots changed direction while on the vertical midline ($1 \leq D \leq N$). In conditions where multiple dots deviate, they do so by the same angle and in the same direction. Subjects reported whether the deviation was counterclockwise ($C = 1$) or clockwise ($C = -1$).

We examine the results from five independent experiments that used either near-threshold or suprathreshold deviations and differed by the number of deviating trajectories. Our goal is to qualitatively understand these results using a single model. In Experiment 1, all trajectories deviated in the same direction and by the same angle ($D = N$). The percentage of "counterclockwise" responses was measured as a function of the magnitude of deviation, when all trajectories deviate. The 84.1% correct ($d' = 1$) threshold was computed as a function of N and found to be virtually independent of N (Tripathy & Barrett, 2004). In Experiment 2, one trajectory deviated ($D = 1$). Threshold was found to increase steeply with N, suggesting a very low attentional capacity (Tripathy & Barrett, 2004). In Experiment 3, $D = 1$ but fixed, suprathreshold deviations of magnitudes $\Delta = 19°, 38°, 76°$ were used. As N was varied, performance decreased, but the curves for the different angles of deviation were clearly separated (Tripathy et al., 2007). In Experiment 4, the same suprathreshold deviations were used, but N was fixed at 6 or 8, and D was varied between blocks. Effective number tracked was defined as the capacity of a hypothetical limited-capacity observer achieving the measured percentage correct. To the surprise of the investigators, this number was found to depend strongly on deviation magnitude, but only weakly on N and D (Tripathy et al., 2007). In Experiment 5, different values of Δ and D were interleaved within a block, making it impossible for subjects to know the difficulty of a trial beforehand. The values $N = 10$, $D = 1, 2$, and $\Delta = 19°, 38°, 57°$ were used (Tripathy et al., 2007). Effective number tracked was again strongly dependent on Δ, and only weakly on D.

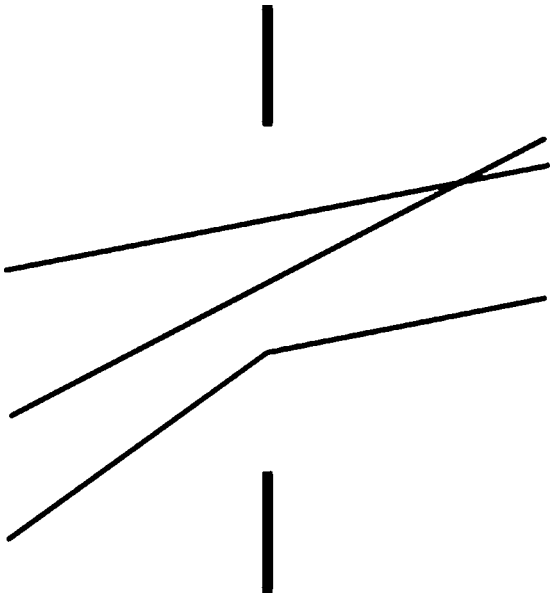

Figure 1. Trajectories on an example $C = -1$ trial (Tripathy & Barrett, 2004). Thin lines represent dots moving from left to right. The thick lines mark the midline and are present in each display.

Bayesian observer

To describe the constrained Bayesian observer, we first explain the neural constraint that causes a particular increase of uncertainty with set size.

Neural constraint on uncertainty

To quantify the notion of limited but flexible resources, we assume that each item is encoded by a similar neural population with Poisson-like variability (Ma et al., 2006). This is a physiologically plausible family of neural

variability distributions that allows for within-population noise correlations and non-unity Fano factors (the Fano factor is the ratio of variance to mean of the spike count of a single neuron; it is 1 for Poisson neurons, but values different from 1 are found in recordings). The constraint is that the total amount of action potentials expended to track N items is roughly independent of N. For a single item, the gain g, which is the mean amplitude of the population pattern of activity, is then roughly proportional to $1/N$. (An exactly equal division of spikes over locations is as unlikely as it is unnecessary. The allocation proportions are flexible and will be influenced by spatial attention.) This could be implemented through divisive normalization, already a key operation in many models of attention (Reynolds & Heeger, 2009). Under Poisson-like variability, gain is proportional to Fisher information, $I(s)$, which in turn is inversely proportional to the stimulus uncertainty σ squared, $g \propto I(s) \propto 1/\sigma^2$ (Seung & Sompolinsky, 1993) (Figure 2). It follows that uncertainty increases as $\sigma \propto 1/\sqrt{g} \propto \sqrt{N}$.

The same constraint was proposed recently for working memory (Bays & Husain, 2008), with two minor differences. In that proposal, items are allocated subsets of neurons from a fixed, common set of neurons that always fire at the same gain. Such an implementation might be harder to implement if the populations encoding different items are spatially separated. Secondly, it was suggested that deviations from an exact $\sigma \propto \sqrt{N}$ relationship (Bays & Husain, 2008; Wilken & Ma, 2004) might be due to correlations. However, the argument above allows for correlations and still predicts $\sigma \propto \sqrt{N}$. Instead, deviations that are still of power-law form, i.e., $\sigma \propto N^{\alpha}$ with $\alpha \neq ½$, could originate from deviations from Poisson-like variability or from a dependence of the total number of available spikes on N. In Appendix C, we comment on the consequences of $\alpha \neq ½$.

The relationship $\sigma \propto \sqrt{N}$ has been proposed earlier based on a sampling argument (Palmer, 1990; Shaw, 1980), but without neural justification. If a fixed total number of S samples is available, then on average, S/N samples will be available per item. By averaging these observations, a better estimate of a single item is obtained. Since the standard deviation of an average of a number of observations generated by the same random process is inversely proportional to the number of terms in the average, we find $\sigma \propto 1/\sqrt{S/N} \propto \sqrt{N}$. While intuitive, this argument does not specify the nature of these samples. Moreover, it is independent of the form of neural variability, while we claim that a different form of variability would produce a different increase of uncertainty with N. For example, if neural variability were additive and Gaussian, then $I(s) \propto g^2$ and $\sigma \propto N$, a completely different relationship.

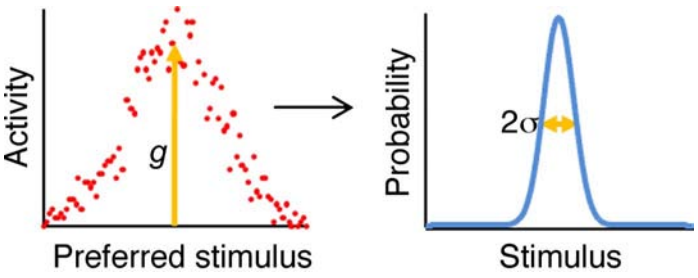

Figure 2. If neuronal variability is Poisson-like, then the gain of a neural population (left; neurons are ordered by their preferred stimulus and the population pattern of activity on a single trial is shown) is inversely proportional to the square of the uncertainty about the encoded stimulus (right): $g \propto 1/\sigma^2$. Under a spike constraint, this implies $\sigma^2 \propto 1/N$.

Generative model

A Bayesian observer (Kersten et al., 2004; Knill & Pouget, 2004; Knill & Richards, 1996) uses the statistical structure of the task to infer a probability distribution over the variable of interest. Unlike limited-capacity models, Bayesian models assume that all observations are noisy, but that little or no random guessing occurs in arriving at a response (Figure 3a). In our case, the probability that $C = 1$ (or $C = -1$) is inferred based on N noisy observations of pre- and post-midline motion directions, which we denote by $\mathbf{x} = (x_1,\ldots, x_N)$ and $\mathbf{y} = (y_1,\ldots, y_N)$. Since C is a binary variable, this probability is uniquely specified by the log odds,

$$d = \log \frac{p(C = 1|\mathbf{x}, \mathbf{y})}{p(C = -1|\mathbf{x}, \mathbf{y})}. \quad (1)$$

The optimal decision on a given trial is to respond "counterclockwise" if $d > 0$ on that trial. This is a straightforward generalization of the likelihood formulation of signal detection theory (Green & Swets, 1966; Wickens, 2002).

For any experimental condition, we can use Bayes' rule and other probability calculus to derive an expression for d in terms of the observations $\mathbf{x}$ and $\mathbf{y}$. To do this, we need to specify the generative model of the task, i.e., a description of the stochastic processes through which the observations are generated by the task-relevant variables. The generative model is depicted graphically in Figure 3b. Besides C, D, $\mathbf{x}$, and $\mathbf{y}$, this diagram contains the following variables: Δ, the angle of deviation; $\mathbf{I}$, the indices of deviating trajectories (a subset of $1,\ldots, N$); $\vec{\Delta}$, the vector of direction changes; and $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, the vectors of pre- and post-midline directions, respectively (known to the experimenter, but not to the observer). The vectors $\mathbf{x}$ and $\mathbf{y}$ consist of sensory observations generated by $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, respectively; they are known to the observer but not to the experimenter.

Each arrow indicates a direct probabilistic dependency. The absence of an arrow indicates the absence of a direct

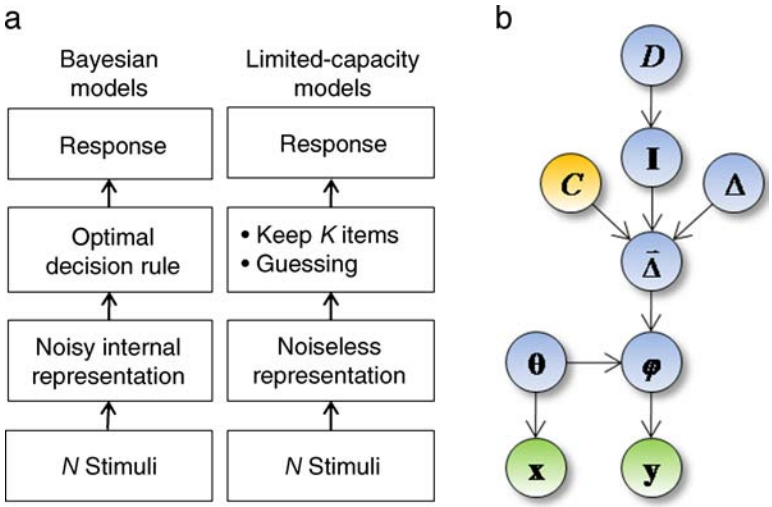

Figure 3. (a) Schematic of decision processes in Bayesian and limited-capacity models. (b) Generative model of the task. Arrows indicate conditional probabilistic dependencies.

probabilistic dependency. For example, we have assumed that the angle of deviation is chosen independently of its sign. Each probabilistic dependency can be formalized as a conditional probability distribution. Most of the distributions in this generative model are common to all experiments and we will first specify those, starting at the bottom of the diagram. We assume that the motion directions of the dots are corrupted by independent sources of sensory variability which obey normal distributions (we ignore the fact that sometimes the dots are so close to each other that the independence assumption may be violated). Before the midline, the vector of observations $\mathbf{x}$ given the actual stimuli $\boldsymbol{\theta}$ is then drawn from the following product of Gaussians:

$$p(\mathbf{x}|\boldsymbol{\theta}) = p(x_1, \ldots, x_N|\theta_1, \ldots, \theta_N) = \prod_{i=1}^{N} p(x_i|\theta_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{\text{pre}}^2}} e^{-\frac{(x_i-\theta_i)^2}{2\sigma_{\text{pre}}^2}}, \quad (2)$$

and similarly for $p(\mathbf{y}\,|\,\boldsymbol{\varphi})$ but with variance σ_{post}^2.

The vector of post-midline motion directions is simply the sum of the vector of pre-midline motion directions and the vector of deviations: $\boldsymbol{\varphi} = \boldsymbol{\theta} + \vec{\Delta}$, or in other words, $p(\boldsymbol{\varphi}\,|\,\boldsymbol{\theta}, \vec{\Delta}) = \delta(\boldsymbol{\varphi} - \boldsymbol{\theta} - \vec{\Delta})$, where δ is the Dirac delta distribution. In turn, the vector of deviations $\vec{\Delta}$ is uniquely determined by the deviation angle Δ, the deviation sign C, and the indices of the deviating trajectories, $\mathbf{I}$: $p(\vec{\Delta}\,|\,\Delta, \mathbf{I}) = \delta(\vec{\Delta} - C\Delta\mathbf{1}_{\mathbf{I}})$, where $\mathbf{1}_{\mathbf{I}}$ is a vector of length N with 1's at the indices in $\mathbf{I}$ and 0's everywhere else.

Finally, the set of indices of deviating trajectories, $\mathbf{I}$, is a randomly chosen subset of size D of the set $1,\ldots, N$. Since there are $\binom{N}{D}$ subsets of size D, the probability of $\mathbf{I}$ given D is equal to $1/\binom{N}{D}$ if $|\mathbf{I}| = D$ and 0 otherwise. Throughout, N is assumed fixed and known to the observer.

Now we have expressed the stochastic dependencies of all variables in the diagram in terms of the top-level variables, C, D, Δ, and $\boldsymbol{\theta}$. What remains is to specify the probability distributions over these variables. Since they are top-level, they do not depend on other variables. Therefore, their distributions are prior distributions, reflecting assumed or learned knowledge about the statistics of the stimuli. Throughout, we will assume flat prior distributions for C and $\boldsymbol{\theta}$, that is, $p(C = 1) = p(C = -1) = ½$ and $p(\theta_i) =$ constant for all i.

The prior distributions over D and Δ depend on the experiment. The distribution over D is a delta distribution (when there is a fixed number of deviating trajectories on each trial, as in Experiments 1–4) or a sum of delta distributions (when the number of deviating trajectories takes one of multiple possible values, as in Experiment 5). The distribution over Δ is uniform (in the threshold Experiments, 1 and 2), a delta distribution (in the suprathreshold Experiments 3, 4, and 6), or a sum of delta distributions (when trials with multiple different deviation angles are interleaved, as in Experiment 5).

Inference

A Bayesian (optimal) observer uses the structure of the generative model to decide which value of C to report on each trial. Specifically, the observer computes the posterior probabilities $p(C = 1\,|\,\mathbf{x}, \mathbf{y})$ and $p(C = -1\,|\,\mathbf{x}, \mathbf{y})$ and reports "$C = 1$" if $p(C = 1\,|\,\mathbf{x}, \mathbf{y}) > p(C = -1\,|\,\mathbf{x}, \mathbf{y})$. We will derive this decision rule in terms of $\mathbf{x}$ and $\mathbf{y}$ for each of the five experiments. Once the decision rule is known, average behavior over a large number of trials can be simulated (or sometimes computed analytically), so that the Bayesian observer can be compared with behavioral data.

Bayes' rule expresses the posterior probability $p(C|\mathbf{x}, \mathbf{y})$ in terms of the likelihood $p(\mathbf{x}, \mathbf{y}|C)$ and the prior $p(C)$:

$$p(C|\mathbf{x},\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y}|C)p(C)}{p(\mathbf{x},\mathbf{y})}. \quad (3)$$

For binary variables such as C, it is convenient to consider the log posterior ratio (log odds), Equation 1, which can, using Equation 3, be rewritten as the sum of a log likelihood ratio and a log prior ratio:

$$d = \log\frac{p(\mathbf{x},\mathbf{y}|C=1)}{p(\mathbf{x},\mathbf{y}|C=-1)} + \log\frac{p(C=1)}{p(C=-1)}. \quad (4)$$

Because of our prior on C, the log odds reduce to the log likelihood ratio. The difficulty in computing the likelihoods lies in the fact that although C is the only task-relevant variable, the probability of $\mathbf{x}$ and $\mathbf{y}$ is also influenced by unknowns which are themselves not of interest, such as $\boldsymbol{\theta}$, Δ (in the threshold experiments), and D (when the number of deviating trajectories is unknown). A Bayesian observer solves this marginalization problem by averaging over these random variables, as we will now examine case by case.

Experiment 1: Near-threshold, N of N

In Experiment 1, all trajectories deviate, so $D = N$ and $\mathbf{I} = \{1,\ldots, N\}$. A threshold paradigm is used, so Δ is varied over a wide range. We evaluate the likelihood $p(\mathbf{x}, \mathbf{y}|C)$ by writing it as an average over the scalar Δ:

$$p(\mathbf{x},\mathbf{y}|C) = \int_0^\infty p(\mathbf{x},\mathbf{y}|C,\Delta)p(\Delta)d\Delta. \quad (5)$$

This is also called marginalizing out or integrating out Δ. The probability of $\mathbf{x}$ and $\mathbf{y}$ conditioned on both C and the scalar Δ is computed by integrating out the vector $\vec{\Delta}$:

$$\begin{aligned} p(\mathbf{x},\mathbf{y}|C,\Delta) &= \int p\left(\mathbf{x},\mathbf{y}|\vec{\Delta},C,\Delta\right)p\left(\vec{\Delta}|C,\Delta\right)d\vec{\Delta} \\ &= \int p\left(\mathbf{x},\mathbf{y}|\vec{\Delta}\right)p\left(\vec{\Delta}|C,\Delta\right)d\vec{\Delta} \\ &= \int p\left(\mathbf{x},\mathbf{y}|\vec{\Delta}\right)\delta\left(\vec{\Delta}-C\Delta\mathbf{1}\right)d\vec{\Delta} \\ &= p\left(\mathbf{x},\mathbf{y}|\vec{\Delta}=C\Delta\mathbf{1}\right), \end{aligned} \quad (6)$$

where we used the facts that $\mathbf{x}$ and $\mathbf{y}$ depend on C and Δ only through $\vec{\Delta}$, and that $\vec{\Delta} = C\Delta\mathbf{1}$, where $\mathbf{1}$ is a vector of length N consisting of only 1's. Next, we integrate out $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$:

$$\begin{aligned} p(\mathbf{x},\mathbf{y}|C,\Delta) &= p\left(\mathbf{x},\mathbf{y}|\vec{\Delta}=C\Delta\mathbf{1}\right) = \\ &= \iint p\left(\mathbf{x},\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\varphi},\vec{\Delta}=C\Delta\mathbf{1}\right)p\left(\boldsymbol{\varphi}|\boldsymbol{\theta},\vec{\Delta}=C\Delta\mathbf{1}\right)p(\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{\varphi} \\ &= \int p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\varphi}=\boldsymbol{\theta}+C\Delta\mathbf{1})p(\boldsymbol{\theta})d\boldsymbol{\theta}, \end{aligned} \quad (7)$$

where again we used the structure of the generative model to simplify the conditional probabilities. We can now explicitly evaluate this integral. We assume a uniform distribution over $\boldsymbol{\theta}$ ($p(\boldsymbol{\theta})$ is then a constant, α) and integrate over the real line (strictly speaking, motion direction is a periodic variable and lives on the circle, but there is little difference if the variability distributions are relatively narrow, as they are here). Then we have

$$\begin{aligned} p(\mathbf{x},\mathbf{y}|C,\Delta) &= \alpha\int p(\mathbf{x}|\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\varphi}=\boldsymbol{\theta}+C\Delta\mathbf{1})d\boldsymbol{\theta} \\ &= \alpha\prod_{i=1}^{N}\int \frac{1}{\sqrt{2\pi\sigma_{\text{pre}}^2}}e^{-\frac{(x_i-\theta_i)^2}{2\sigma_{\text{pre}}^2}}\frac{1}{\sqrt{2\pi\sigma_{\text{post}}^2}}e^{-\frac{(y_i-\theta_i-C\Delta)^2}{2\sigma_{\text{post}}^2}}d\theta_i \\ &= \alpha\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-x_i-C\Delta)^2}{2\sigma^2}} \\ &= \alpha(2\pi\sigma^2)^{-\frac{N}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-x_i-C\Delta)^2}, \end{aligned} \quad (8)$$

where we used a standard formula for the integral over the product of two Gaussians and introduced the notation $\sigma^2 = \sigma_{\text{pre}}^2 + \sigma_{\text{post}}^2$. Inserting this result back into Equation 5, we find

$$\begin{aligned} p(\mathbf{x},\mathbf{y}|C) &= \alpha(2\pi\sigma^2)^{-\frac{N}{2}}\int_0^\infty e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-x_i-C\Delta)^2}p(\Delta)d\Delta \\ &= \alpha\left(2\pi\sigma^2\right)^{-\frac{N}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-x_i)^2}\int_0^\infty e^{-\frac{\Delta^2}{2\sigma^2}+\frac{C\Delta}{\sigma^2}\sum_{i=1}^{N}(y_i-x_i)}p(\Delta)d\Delta, \end{aligned} \quad (9)$$

where we used $C^2 = 1$. Finally, we obtain the log odds from Equation 4:

$$d = \log\frac{\int_0^\infty e^{-\frac{\Delta^2}{2\sigma^2}+\frac{\Delta}{\sigma^2}\sum_{i=1}^{N}(y_i-x_i)}p(\Delta)d\Delta}{\int_0^\infty e^{-\frac{\Delta^2}{2\sigma^2}-\frac{\Delta}{\sigma^2}\sum_{i=1}^{N}(y_i-x_i)}p(\Delta)d\Delta}. \quad (10)$$

If $\Sigma_i(x_i - y_i) > 0$, then the integrand in the numerator is larger than the integrand in the denominator for any Δ

(since $\Delta/\sigma^2 > 0$). Moreover, both integrands are non-negative functions on the entire domain of Δ. It follows that $d > 0$. Similarly, if $\Sigma_i(x_i - y_i) < 0$, then $d < 0$. From this, we conclude that $d > 0$ is equivalent to

$$\sum_{i=1}^{N} \frac{y_i - x_i}{\sigma^2} > 0, \tag{11}$$

regardless of the form of $p(\Delta)$. Note that we kept the constant factor $1/\sigma^2$, in anticipation of situations where uncertainty might differ between dots ($1/\sigma_i^2$), a case of which we will discuss in the Predictions section. The decision rule is to report "$C = 1$" when the average post-midline motion direction is larger than the average pre-midline one. That this is the optimal strategy is intuitive and could have been guessed without doing any calculations: since each trajectory deviates by the same amount, maximum information about the deviation is obtained by averaging all N observations. However, the same calculation method will be used in more complex conditions.

In Experiments 1 and 2, performance is measured as a deviation threshold, i.e., as the value of Δ for which the observer is 84.1% correct (this number is obtained as ½ + ½ erf($1/\sqrt{2}$), where erf(x) is the error function). To relate the Bayesian model to human performance, we have to apply the Bayesian decision rule to a large number of trials. Usually, this requires simulation, but in this particular case we can do it analytically. In the Bayesian model, probability correct for a given value of Δ is the probability that $d > 0$ when $C = 1$ (or that $d < 0$ when $C = -1$, which is the same). When $C = 1$, each random variable $y_i - x_i$ follows a Gaussian distribution with mean Δ and standard deviation σ, and their average follows a Gaussian distribution with mean Δ and standard deviation $\sigma/\sqrt{N}$. Therefore, the probability that their average is positive is equal to ½ + ½ erf($\Delta/\sigma \cdot \sqrt{N/2}$). Comparing this with the above expression for 84.1% correct yields that the threshold deviation is equal to $\Delta_{\text{thr}} = \sigma/\sqrt{N}$.

In the constrained Bayesian model, the spike constraint causes the standard deviation to scale with the square root of N: $\sigma = \sigma_1\sqrt{N}$, where σ_1 is the combined pre- and post-midline uncertainty when only 1 trajectory is present. It follows that $\Delta_{\text{thr}} = \sigma_1$, indicating that the threshold deviation is independent of N. In other words, the benefit gained from averaging N observations is exactly undone by the increase in uncertainty due to the spread of attention over N items. This also means that, in this task, it does not make a difference how many items are tracked since tracking additional items does not improve performance.

Experiment 2: Near-threshold, 1 of N

Experiment 2 differs from Experiment 1 in that only one, randomly chosen trajectory is deviating. It means that $D = 1$ and $\mathbf{I}$ is a single index j randomly chosen from $\{1,\ldots, N\}$. As a consequence, in Equation 6, $p(\vec{\Delta}|C, \Delta)$ needs to be computed as an average over $\mathbf{I}$:

$$\begin{aligned} p\left(\vec{\Delta}|C,\Delta\right) &= \frac{1}{N}\sum_{j=1}^{N} p\left(\vec{\Delta}|C,\Delta,\mathbf{I}=\{j\}\right) \\ &= \frac{1}{N}\sum_{j=1}^{N} \delta\left(\vec{\Delta} - C\Delta\mathbf{1}_j\right). \end{aligned} \tag{12}$$

Therefore, Equation 6 gets replaced by

$$p(\mathbf{x},\mathbf{y}|C,\Delta) = \frac{1}{N}\sum_{j=1}^{N} p\left(\mathbf{x},\mathbf{y}|\vec{\Delta} = C\Delta\mathbf{1}_j\right). \tag{13}$$

The conditional probability inside the sum indicates that the jth trajectory is deviating, while all others are not. As in Experiment 1, we write this as an average over θ,

$$\begin{aligned} p(\mathbf{x},\mathbf{y}|C,\Delta) &= \\ &= \frac{1}{N}\sum_{j=1}^{N} \int p(\mathbf{x},\mathbf{y}|\theta,\varphi = \theta + C\Delta\mathbf{1}_j)p(\theta)d\theta. \end{aligned} \tag{14}$$

To compute this integral, we write it as a product over all individual trajectories, keeping in mind that the jth factor is different from all others and therefore needs a separate factor:

$$\begin{aligned} &p(\mathbf{x},\mathbf{y}|C,\Delta) = \\ &= \frac{\alpha}{N}\sum_{j=1}^{N} \int \frac{1}{\sqrt{2\pi\sigma_{\text{pre}}^2}} e^{-\frac{(x_j-\theta_j)^2}{2\sigma_{\text{pre}}^2}} \frac{1}{\sqrt{2\pi\sigma_{\text{post}}^2}} e^{-\frac{(y_i-\theta_j-C\Delta)^2}{2\sigma_{\text{post}}^2}} d\theta_j \\ &\cdot \prod_{i\neq j} \int \frac{1}{\sqrt{2\pi\sigma_{\text{pre}}^2}} e^{-\frac{(x_i-\theta_i)^2}{2\sigma_{\text{pre}}^2}} \frac{1}{\sqrt{2\pi\sigma_{\text{post}}^2}} e^{-\frac{(y_i-\theta_i)^2}{2\sigma_{\text{post}}^2}} d\theta_i \\ &= \frac{\alpha}{N}\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-x_i)^2}{2\sigma^2}}\right) \sum_{j=1}^{N} e^{\frac{(y_j-x_j)^2}{2\sigma^2}} e^{-\frac{(y_j-x_j-C\Delta)^2}{2\sigma^2}}. \end{aligned} \tag{15}$$

This form allows us to compute the average over Δ, as in Equation 5. We assume a uniform prior distribution over Δ (this is a threshold paradigm, so this assumption is

reasonable), which takes a small value β on a large interval. Then

$$p(\mathbf{x},\mathbf{y}|C) = \\ = \frac{\alpha\beta}{N}\left(\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-x_i)^2}{2\sigma^2}}\right)\sigma\sqrt{\frac{\pi}{2}}\sum_{j=1}^{N}e^{\frac{(y_j-x_j)^2}{2\sigma^2}}\left(1+\mathrm{erf}\frac{y_j-x_j}{\sigma C\sqrt{2}}\right). \quad (16)$$

Now, we are ready to compute the log odds:

$$d = \log\frac{\sum_{j=1}^{N}e^{\frac{(y_j-x_j)^2}{2\sigma^2}}\left(1+\mathrm{erf}\frac{y_j-x_j}{\sigma\sqrt{2}}\right)}{\sum_{j=1}^{N}e^{\frac{(y_j-x_j)^2}{2\sigma^2}}\left(1-\mathrm{erf}\frac{y_j-x_j}{\sigma\sqrt{2}}\right)}. \quad (17)$$

The decision rule, $d > 0$, thus becomes

$$\sum_{j=1}^{N}e^{\frac{(y_j-x_j)^2}{2\sigma^2}}\mathrm{erf}\frac{y_j-x_j}{\sigma\sqrt{2}} > 0. \quad (18)$$

Contrary to the averaging rule in Experiment 1, this decision rule would have been impossible to guess. Another difference with Experiment 1 is that the decision rule now contains the uncertainty parameter, σ, in an essential way (in Experiment 1, it was irrelevant if uncertainty was equal across items). Therefore, it requires that a neural population encoding motion direction also encodes, on a single trial, the uncertainty about a stimulus, and that this information is used in downstream computations. This utilization of implicit knowledge of one's uncertainty is what we mean by Bayesian inference (even though the prior distribution is flat in this case). Probabilistic population codes (Ma et al., 2006) provide a concrete neural implementation of a Bayes-optimal computation (cue combination).

Positional uncertainty

In Experiment 2, only 1 trajectory is deviating, and therefore we also need to take into account positional uncertainty. By this we mean the uncertainty in the endpoints of the first halves of all trajectories and the starting points of their second halves, on the vertical midline. This uncertainty will lead to mispairings, whereby the first half of one trajectory is mistakenly associated with the second half of a different trajectory. This "correspondence problem" will make correct change discrimination more difficult. Figure 4a shows an illustration.

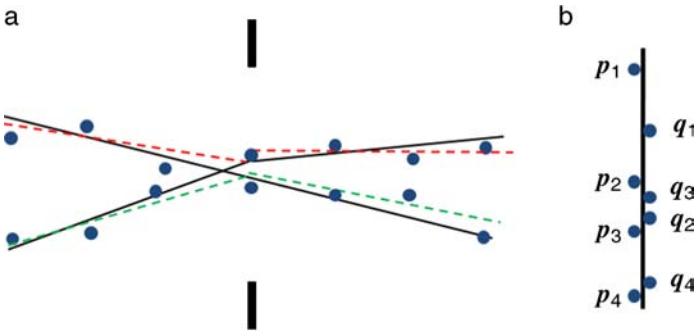

Figure 4. Illustration of positional uncertainty. (a) Two example trajectories are shown in black. Blue dots represent noisy observations of dot positions during the trial. Dashed lines indicate the inferred trajectories. Because of the relative positions of the intersections of the inferred trajectories with the vertical midline, the observer erroneously believes that the red trajectory halves belong together, and the green ones. In truth, the left-hand red segment and the right-hand green segment belong together. (b) In the model, we conceptualize this problem using endpoints of first halves, p_i, and starting points of second halves, q_i (example shown has four trajectories). Correspondence is established by matching the points in p and q after ordering them separately. In this example, this would lead to the pairings (p_1, q_1), (p_2, q_3), (p_3, q_2), and (p_4, q_4).

In Experiment 1, this did not play a role, since the optimal decision rule involved averages of the pre- and post-midline observed motion directions, and averages over all trajectories are not affected by mispairings. However, the numerator and denominator in the log odds of Experiment 2 are not invariant under permutations of the indices in $\mathbf{x}$ or $\mathbf{y}$.

Mispairings are particularly prone to occur whenever two trajectories intersect, which, by experimental design, does not occur exactly on the vertical midline (a minimum distance between the dots is respected at their time of deviation). Therefore, we are considerably simplifying the problem by only taking into account positional uncertainty on the vertical midline. We are essentially summarizing all potential mispairings into one single moment. This simplification is meant to capture the essence of positional uncertainty without the need to model time courses of the positions of all dots, which would make the model much more complex.

Each trajectory consists of two halves. We model positional uncertainty by drawing endpoints of first halves, $\mathbf{p}$, and starting points of second halves, $\mathbf{q}$, from normal distributions centered at common actual positions $\mathbf{L}$. Those positions are specified by the experiment, i.e., they are placed equidistantly on the real line and then corrupted by uniform jitter (Tripathy & Barrett, 2004; Tripathy et al., 2007). The standard deviations of the normal distributions are free parameters and are assumed to be equal. Like the observations of direction, the observations of position are subject to the spike constraint and therefore

these standard deviations increase with $\sqrt{N}$. We will denote the positional uncertainty at set size 1 by $\sigma_{\text{pos},1}$; it is $\sqrt{2}$ times the positional uncertainty in $\mathbf{p}$ or $\mathbf{q}$ separately.

The correspondence between first-half endpoints $\mathbf{p}$ and second-half starting points $\mathbf{q}$ is established by picking the most likely pairing. This is the pairing in which the sorted version of $\mathbf{p}$ corresponds, entry by entry, to the sorted version of $\mathbf{q}$: the smallest entry in $\mathbf{p}$ corresponds to the smallest entry in $\mathbf{q}$, etc. (Figure 4b). Specifically, we denote by $S_{\mathbf{p}}(\mathbf{p})$ the permutation that sorts $\mathbf{p}$. For example, if $\mathbf{L}$ = (8, 23, 27, 45), then $\mathbf{p}$ could be (10.5, 25.1, 24.6, 42). $S_{\mathbf{p}}(\mathbf{p})$ is then the permutation (1, 2, 3, 4) → (1, 3, 2, 4). Similarly, $S_{\mathbf{q}}(\mathbf{q})$ is the permutation that sorts $\mathbf{q}$. We define permuted sets of motion direction observations by applying the same permutations, $S_{\mathbf{p}}$ and $S_{\mathbf{q}}$, to $\mathbf{x}$ and $\mathbf{y}$, respectively: $\mathbf{x}_{\text{new}} = S_{\text{p}}(\mathbf{x})$ and $\mathbf{y}_{\text{new}} = S_{\mathbf{q}}(\mathbf{y})$.

Subsequently, $\mathbf{x}_{\text{new}}$ and $\mathbf{y}_{\text{new}}$ are entered into the decision rule, instead of $\mathbf{x}$ and $\mathbf{y}$. For example, in Experiment 2, the final decision rule becomes

$$\sum_{j=1}^{N} e^{\frac{(y_{\text{new},j}-x_{\text{new},j})^2}{2\sigma^2}} \operatorname{erf}\frac{y_{\text{new},j}-x_{\text{new},j}}{\sigma\sqrt{2}} > 0. \tag{19}$$

We use this permutation procedure in every application of the Bayesian models in this paper, except in Experiment 1, where it is not needed (as discussed above), and in the inset of Figure 14a, where the basic effect is demonstrated without positional uncertainty.

This implementation of positional uncertainty is the only non-Bayesian element of our model. A purely Bayesian observer would average over all possible permutations, weighted by their probabilities, but we believe that this is unlikely, because their number grows as N factorial.

Experiment 3: Suprathreshold, 1 of N

In this experiment, 1 of N trajectories deviates, where the angle of deviation is relatively large. It might seem that the generative model is the same as that of Experiment 2. There however, Δ takes on values over a wide range, whereas here, Δ is fixed within a block. A Bayesian observer incorporates this knowledge through $p(\Delta)$. This means that $p(\mathbf{x}, \mathbf{y}|C) = p(\mathbf{x}, \mathbf{y}|C, \Delta)$. Combined with Equations 4 and 8, this leads to the decision rule

$$\sum_{i=1}^{N} e^{\frac{\Delta(y_i-x_i)}{\sigma^2}} > \sum_{i=1}^{N} e^{-\frac{\Delta(y_i-x_i)}{\sigma^2}}, \tag{20}$$

where Δ is now a specific value rather than a variable. It is interesting to note that in the limit that $y_i - x_i$ is much larger for one value of i than for all others, and similarly for $x_i - y_i$, then the sums on both sides are dominated by the largest terms, and this decision rule simplifies to the so-called signed-max rule (Baldassi & Verghese, 2002):

$$\max_i \frac{y_i - x_i}{\sigma^2} > \max_i \frac{x_i - y_i}{\sigma^2}. \tag{21}$$

(This is only a statement about the decision rule; keep in mind that we make the additional assumption that uncertainty increases as $\sqrt{N}$.) However, this rule is not optimal outside of this limit or in other conditions. The optimal rule, Equation 20, can be regarded as a soft version of the signed max rule, as the exponential function preferentially amplifies larger observed deviations $y_i - x_i$ (or $x_i - y_i$). It is intuitive that this is optimal, as larger observed deviations provide more conclusive evidence about the true direction of deviation than smaller ones.

Experiment 4: Suprathreshold, D of N, blocked

This situation is not fundamentally different from that in Experiment 3; the only difference is that there are now more possible subsets of deviating trajectories, namely $\binom{N}{D}$ of them. The decision rule is in Appendix A.

Experiment 5: Suprathreshold, D of N, interleaved

In Experiment 4, Δ and D were fixed in a given block and only varied between blocks, whereas in Experiment 5, trials with different values of Δ and D were interleaved within a single block. In a Bayesian model, these designs correspond to different assumed distributions over Δ and D. In Experiment 5, the observer is not sure of the values of Δ and D, and therefore marginalizes over these variables. The decision rule is in Appendix A.

Monte Carlo simulations

In order to compare Bayesian model predictions with human data and other models, we apply the Bayesian decision rule, for every experiment separately, to a large number of synthetic observations (Monte Carlo simulation). These synthetic data are generated from the generative model pertaining to that experiment. To simulate a set of observations, we pick a deviation angle from the range used in the experiment. For each trial, we decide beforehand whether the deviation is clockwise or counterclockwise (with equal probability). Then, we randomly choose initial motion directions as well as which trajectory is deviating. Finally, based on the motion directions generated in this way, we synthesize noisy observations $\mathbf{x}$ and $\mathbf{y}$ by drawing independently from Gaussian distributions centered at these motion directions. After generating the observations, we apply the Bayesian decision rule to the observations $\mathbf{x}$ and $\mathbf{y}$ on each trial. Then, we determine the proportion of trials on which the

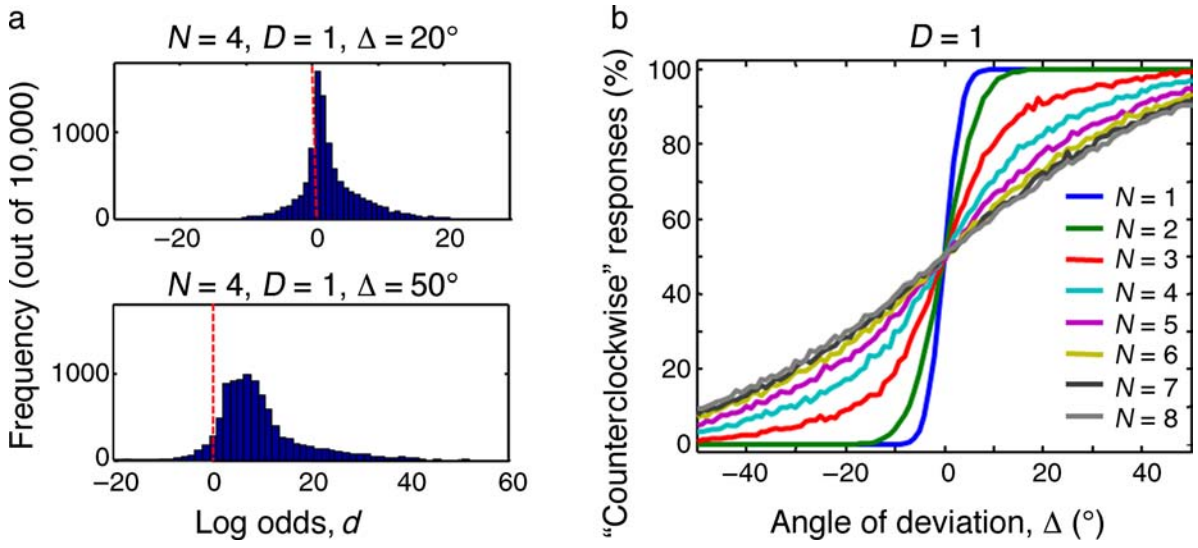

Figure 5. Properties of the constrained Bayesian model in the threshold paradigm (1 of *N* deviating). (a) Histogram of the Bayesian decision variable, the log odds of counterclockwise versus clockwise deviation, when $C = 1$. The log odds follow a non-Gaussian distribution. The dashed red line indicates the optimal criterion, $d = 0$. These examples use $N = 4$, $D = 1$, $\Delta = 20°$ (top), and $\Delta = 50°$ (bottom). (b) Percentage "counterclockwise" responses as a function of the deviation angle, Δ, for different numbers of trajectories, *N*. Saturation performance is 100% for any *N*. In both panels a and b, $\sigma_1 = 3°$ and $\sigma_{pos,1} = 15'$.

model response is correct. In terms of the log odds, *d*, this is the same as calculating a histogram of *d* for (**x**, **y**) pairs drawn under $C = 1$, and counting what portion of the histogram satisfies $d > 0$ (equivalently, $C = -1$ and $d < 0$). The log odds do, in general, not follow a normal distribution (see Figure 5a).

This procedure is then repeated for a large number of deviation angles, so that we can plot the psychometric curve (see Figure 5b). We then fit a cumulative normal function to the psychometric curve, thus obtaining a best estimate for Δ_{thr}. There is no reason why the psychometric curve should have a cumulative normal shape, and in many cases it does not. However, it is a reasonable approximation, and our analysis of model data parallels that of the behavioral data.

In the suprathreshold experiments, effective number tracked was computed in the same way as in the behavioral experiments (Tripathy et al., 2007), namely by asking in a given condition what would be the capacity of a hypothetical limited-capacity observer achieving the same percentage correct as the Bayesian observer. The limited-capacity observer for general *D* is described in the next section. Since capacity is an integer, we interpolate percentage correct as a function of capacity using an exponential fit, again following (Tripathy et al., 2007).

Each experiment has a number of parameters that are set by the experimenter, which we tried to replicate as closely as possible from Tripathy and Barrett (2004) and Tripathy et al. (2007); these are not free. In addition, each Bayesian model has one or two free parameters: the directional uncertainty at set size 1, σ_1, and in Experiments 2–5, the positional uncertainty at set size 1, $\sigma_{pos,1}$. The values of these free parameters were taken to be consistent across experiments. In Experiment 1, we expect the constant deviation threshold to be approximately equal to σ_1. Moreover, threshold for discriminating the deviation sign in a single trajectory decreases was measured separately as a function of dot speed (Tripathy & Barrett, 2004); we used approximately those values for σ_1, keeping in mind that different experiments used different dot speeds. $\sigma_{pos,1}$ was fitted by hand. The experiment-specific parameters are listed in Appendix B. In all simulations, we used at least 10,000 trials per condition.

Unconstrained Bayesian model

The unconstrained Bayesian model is identical to the constrained Bayesian model except that the uncertainty per item does not increase with *N*, i.e., $\sigma = \sigma_1$ and $\sigma_{pos} = \sigma_{pos,1}$.

Limited-capacity model

In the traditional limited-capacity model (Cowan, 2001; Hulleman, 2005; Luck & Vogel, 1997; Oksama & Hyona, 2008; Pashler, 1988; Pylyshyn & Storm, 1988), a capacity limit *K* (a positive integer) is assumed, meaning that on each trial, *K* trajectories are randomly selected; if $K \geq N$, all are selected. If the deviating trajectory is among these, the observer will report the correct sign of the deviation. If none of the *K* selected trajectories deviates, then the observer guesses about the sign of the deviation, picking "clockwise" or "counterclockwise" each with probability ½.

A characteristic feature of the limited-capacity model (as well as its variations, which will be discussed in the next sections) is that when one trajectory deviates, the deviation threshold is infinite for $N > 1.46K$. We will prove this below. Another characteristic is that performance is independent of deviation angle, which is relevant in Experiments 3–5.

Experiment 1: N of N

The traditional limited-capacity model predicts 100% performance and a threshold of 0° regardless of K or N, since the selected trajectory or trajectories will always be deviating and their sign of deviation will be known with absolute certainty.

Experiments 2 and 3: 1 of N

In Experiments 2 and 3, 1 of N trajectories is deviating. The proportion of trials on which the observer responds correctly is then 1 if $N \leq K$, and

$$\mathrm{PC}(N,1) = \frac{K}{N} + \frac{1}{2}\left(1 - \frac{K}{N}\right) = \frac{1}{2}\left(1 + \frac{K}{N}\right), \qquad (22)$$

if $N \geq K$. This is shown as a function of N, for different values of K, in Figure 6a. The limited-capacity model does not take into account the angle of deviation, Δ. For any set size, N, and any value of the capacity, K, percentage correct is independent of deviation angle (see Figure 6b). Since deviation threshold is defined as the smallest value of the deviation angle for which proportion correct exceeds 0.841, the limited-capacity model predicts that only two possible threshold values are possible: zero and infinity, depending on whether PC in Equation 22 randomly is larger or smaller than 0.841, respectively. We can compute for which values of N the deviation threshold is infinite by solving the equation $\mathrm{PC}(N, 1) < ½ + ½\ \mathrm{erf}(1/\sqrt{2})$. The solution is

$$N > K/\mathrm{erf}(1/\sqrt{2}) \approx 1.46K. \qquad (23)$$

This means that if $K = 3$, the threshold deviation according to the limited-capacity model will be infinite whenever $N \geq 5$. This leads to threshold-versus-set size curves that look like those in Figure 6c.

In the limited-capacity model, there is no difference between near-threshold and suprathreshold paradigms, so Equation 22 is valid for both Experiments 2 and 3.

Experiments 4 and 5: D of N

When D of N trajectories deviate, the limited-capacity model predicts that $\mathrm{PC}(N, D) = 1$ when $N < K + D$ (since then at least one deviating trajectory is attended to). When $N \geq K + D$, the model predicts, through a basic combinatorial argument, the following proportion correct:

$$\mathrm{PC}(N,D) = 1 - \frac{1}{2}\binom{N-D}{K} \Big/ \binom{N}{K}. \qquad (24)$$

This is shown as a function of D and for different values of K in Figures 6d and 6e (with $N = 6$ and $N = 8$, respectively). Again, percentage correct is independent of deviation angle (Figure 6f). Equation 24 is valid regardless of whether different values of D are blocked or interleaved.

The limited-capacity observer is used to define the notion of effective capacity or effective number tracked (Tripathy et al., 2007). When percentage correct is measured for given D and N, the effective capacity of the human observer is defined as the capacity of a hypothetical limited-capacity observer with the same percentage correct, where PC is interpolated between integer values of K using an exponential fit. We applied the same method to the simulated responses of different models, to obtain model effective capacity.

Slots-plus-averaging model

The slots-plus-averaging model (Zhang & Luck, 2008) is a variation of the limited-capacity model that takes uncertainty into account to some extent but still assumes that no more than K items can receive resource. The model was developed for short-term memory but can also be applied to attentional tracking. The shortcomings of the limited-capacity model for visual-short term memory were pointed out in a set of two-interval suprathreshold feature change detection experiments (Wilken & Ma, 2004). This work suggested instead that short-term memory limitations originate from the variability in the sensory encoding of items combined with a finite but continuous resource. A direct estimation task was introduced, in which subjects estimated in the second interval the feature value of one item, which was among multiple items present in the first interval. This confirmed that precision with which items are maintained in memory smoothly decreases with set size. In response, Zhang and Luck proposed the slots-plus-averaging model, which attempted to address the decline of precision with set size by postulating that resources come in a small number of discrete chunks, K (slots). When there are fewer slots than items ($N \geq K$), the number of slots is equivalent to the capacity in the traditional limited-capacity model. However, when there are more slots than items ($N < K$), multiple slots will be allocated to the same item, thereby increasing the quality of its encoding, in a way similar to the sampling argument

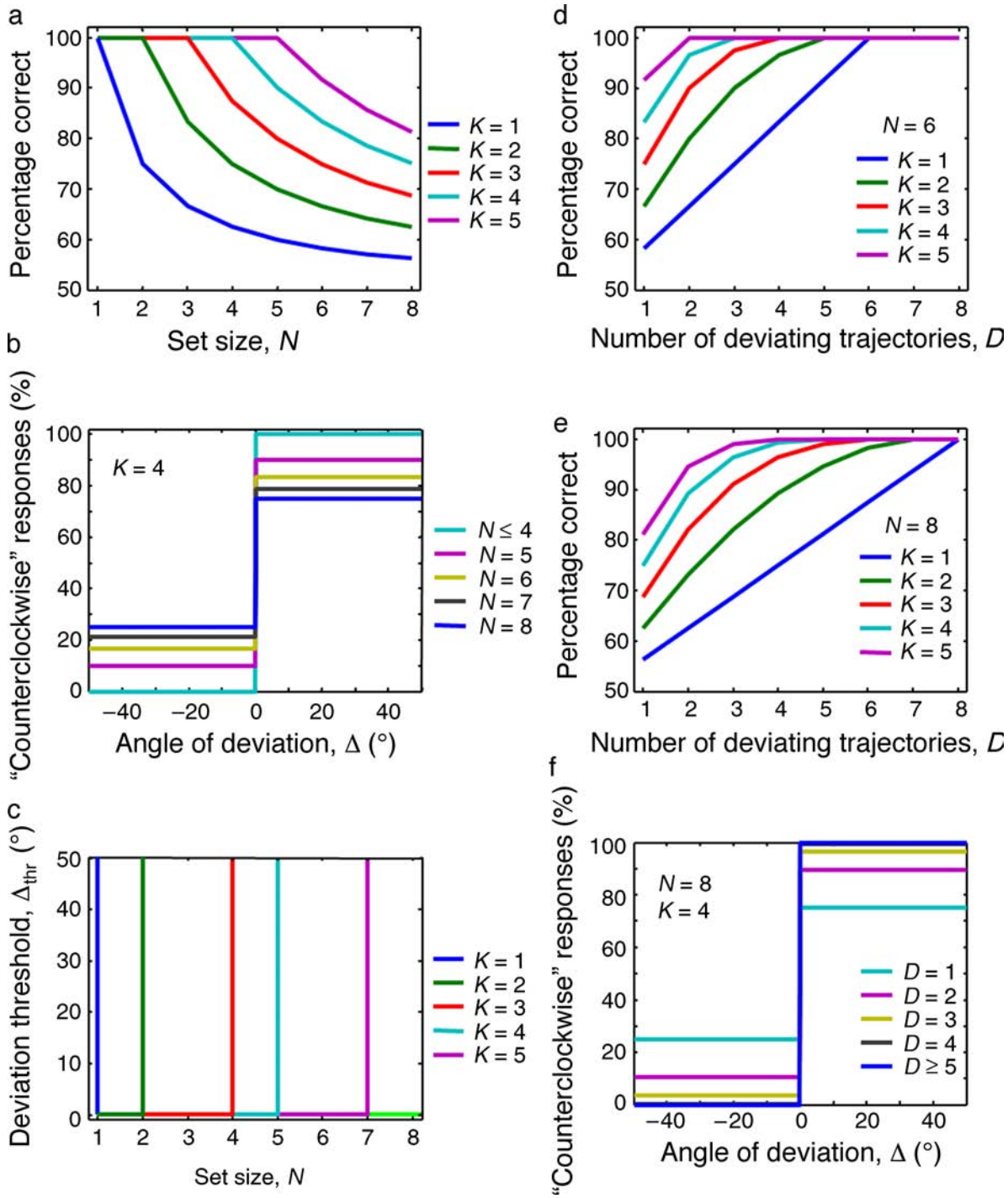

Figure 6. Predictions of the traditional limited-capacity model when 1 (a–c) or *D* (d–f) of *N* trajectories deviate. (a) Percentage correct as a function of set size, *N,* for different values of the capacity, *K*. (b) Psychometric functions for different set sizes. In this example, $K = 4$. (c) Deviation threshold (at 84.1% correct) as a function of set size for different values of *K*. The cyan curve ($K = 4$) is derived from the psychometric functions in panel b. The bright green line (flat at 0°) represents the deviation threshold when all *N* trajectories deviate (Experiment 1). (d) Percentage correct as a function of *D,* when $N = 6$, for different values of *K*. (e) Same when $N = 8$. (f) Psychometric functions for $N = 8$, $K = 4$, and different values of *D* ($D = 4$ and $D \geq 5$ lines nearly overlap). Percentage correct is independent of deviation angle. Its values in this example can be read off from the cyan curve ($K = 4$) in panel e.

mentioned in "Neural constraint on uncertainty." Items that do not receive any slot are not maintained at all. Therefore, this model is a hybrid between the limited-capacity model and a continuous-resource model like the constrained Bayesian observer. Here, we only describe aspects of the slots-plus-averaging model that can be directly applied to the tracking task we study.

As the slots-plus-averaging model has not yet been applied to change detection or attentional tracking, we make our best guess as to how the concepts outlined in (Zhang & Luck, 2008) would translate to our context. As in the traditional limited-capacity model, the slots-plus-averaging model would postulate a categorical distinction between attended and unattended items, with the observer having no information at all about an unattended item. However, unlike the traditional limited-capacity model, the slots-plus-averaging model acknowledges the existence of variability in the representation of attended items. One would expect this variability in internal representation to lead the observer to sometimes mistake a non-deviating trajectory for a deviating one, or the other way round, just like in any near-threshold discrimination task. Bayesian models automatically take these confusions into account and prescribe how noisy observations from different trajectories should be combined into a single, optimal decision rule. However, since probabilistic inference across multiple items is contrary to the thinking behind limited-capacity models, we will instead assume that the observer somehow knows which of the trajectories were deviating.

We use PC(Δ, N, D) to denote predicted proportion correct at deviation angle Δ, when D of N trajectories are deviating. For the deviation threshold when D of N trajectories are deviating, we use the notation $\Delta_{\text{thr}}(N, D)$.

Experiment 1: N of N

We call the number of slots K. When $N \geq K$, each attended item will be encoded with a certain standard deviation σ_K (one slot each). This standard deviation corresponds to the combined pre- and post-midline uncertainty. Averaging K such observations will produce standard deviation $\sigma_K/\sqrt{K}$. When $N \leq K$, each item will be encoded with standard deviation $\sigma_N = \sigma_K \sqrt{N/K}$ (one slot each). Averaging N such observations will produce standard deviation $\sigma_K/\sqrt{K}$. We conclude that the deviation threshold will be $\Delta_{\text{thr}}(N, N) = \sigma_K/\sqrt{K}$, independent of N. That this is the same as in the Bayesian model is not surprising, since the naïve averaging operation happens to be optimal when all trajectories deviate.

Experiments 2 and 3: 1 of N

When $N \geq K$, there is a probability of K/N that the deviating trajectory is allocated a slot. When this happens, its internal representation will have standard deviation σ_K. Then, probability correct equals the probability that an observation drawn from a normal distribution with mean Δ (which is positive) and standard deviation σ_K is itself positive (so that the correct deviation sign will be reported).

This probability is $\frac{1}{2} + \frac{1}{2}\,\text{erf}(\Delta/(\sigma_K\sqrt{2}))$. On the other hand, when the deviating trajectory is not allocated a slot, performance will be at chance. Consequently, the predicted proportion of correct responses is, after simplification,

$$\text{PC}(\Delta, N \geq K, 1) = \frac{1}{2}\left(1 + \frac{K}{N}\,\text{erf}\,\frac{\Delta}{\sigma_K\sqrt{2}}\right). \quad (25)$$

As a check, when Δ is large compared to σ_K, the error function will tend to 1, and PC is the same as in the traditional limited-capacity model, Equation 22. It follows from Equation 25 that proportion correct is bounded by $\text{PC}(\Delta = \infty, N, 1) = \frac{1}{2}(1 + K/N)$. Just as in the traditional limited-capacity model, PC never reaches 0.841 if $N > 1.46K$ (see Equation 23). For these values of N, $\Delta_{\text{thr}}(N, 1) = \infty$. For other values of N, i.e., those which satisfy $K \leq N < 1.46K$, we can compute the threshold deviation from Equation 25 as:

$$\Delta_{\text{thr}}(K \leq N \leq 1.46K, 1) = \sigma_K\sqrt{2}\,\text{erf}^{-1}\left(\frac{N}{K}\,\text{erf}\,\frac{1}{\sqrt{2}}\right). \quad (26)$$

When $N \leq K$, each item will be attended and have at least one slot allocated to it. The average number of slots allocated will be K/N. As a consequence, the standard deviation of its internal representation will be reduced by a factor $\sqrt{K/N}$: $\sigma_N = \sigma_K\sqrt{N/K}$. (This ignores the discrete nature of the slots, which prevent allocation of fractional slots. For example, when there are four slots and three items, one item will get one more slot than the other two. Taking that into account would yield a minor correction and is not essential to the model comparison.)

We assume that the observer somehow knows which of the trajectories was deviating, leaving the task of determining the sign of the deviation. Therefore, proportion correct will be

$$\text{PC}(\Delta, N \leq K, 1) = \frac{1}{2} + \frac{1}{2}\,\text{erf}\,\frac{\Delta}{\sigma_K}\sqrt{\frac{K}{2N}}. \quad (27)$$

When $K = N$, this is equal to Equation 25, and when Δ is large compared to σ_K, it is equal to 1 as in the traditional limited-capacity model. Thus, for $N \leq K$, threshold is $\Delta_{\text{thr}}(N \leq K, 1) = \sigma_K\sqrt{N/K}$. The percentage of "counter-clockwise" responses is plotted as a function of Δ for different values of N in Figure 7a. The psychometric function for $N = 1$ is identical to the one when N of N

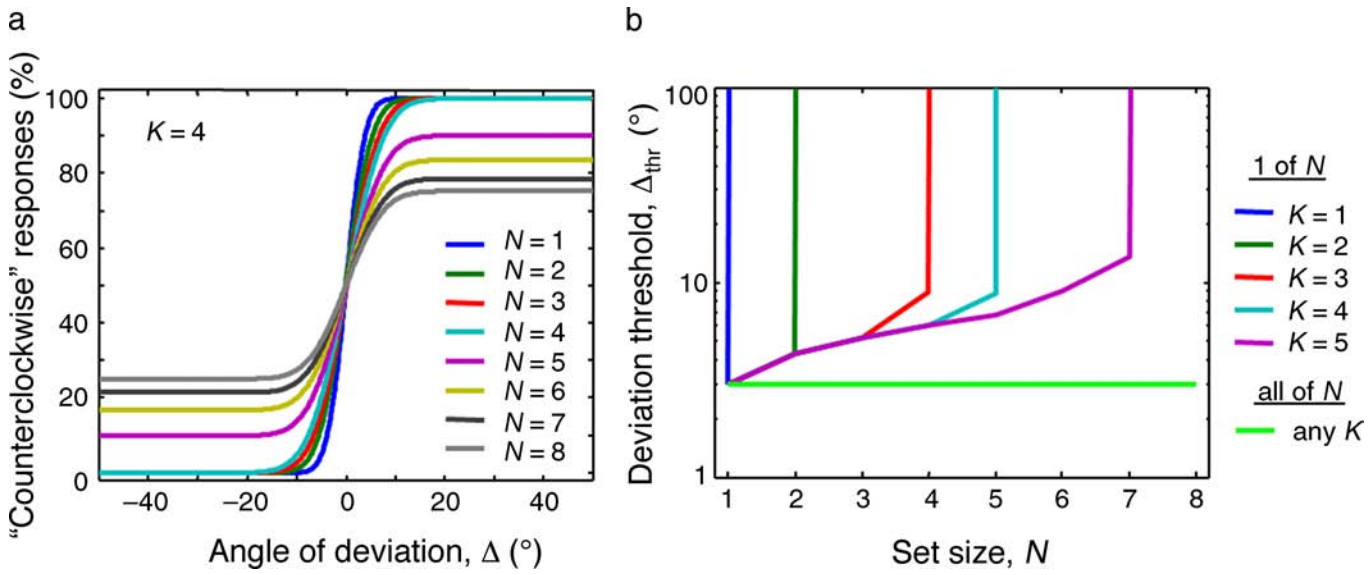

Figure 7. Predictions of the slots-plus-averaging model for the threshold experiments. (a) Percentage "counterclockwise" responses as a function of the deviation angle, Δ, for different numbers of trajectories, N, when 1 trajectory is deviating. In this example, $K = 4$. Note that saturation performance is far below 100% when $N > K$, as is characteristic of all limited-capacity models. (b) Deviation threshold as a function of set size for different numbers of slots, K, when either 1 of N or all N trajectories are deviating. For example, the cyan curve ($K = 4$) is derived from the psychometric functions in panel a. To ensure that the threshold at $N = 1$ is identical for all K (at a value close to the data), we chose $\sigma_K = \sigma_{K=1}\sqrt{K}$ (see text for details). The solid lines show $\sigma_{K=1} = \sigma_{N=1} = 3°$. Changing $\sigma_{K=1}$ only has the effect of shifting the curves vertically. Figure 6a represents asymptotic performance (at very large Δ) in the slots-plus-averaging model.

trajectories are deviating (regardless of N). Performance saturates at a level far below 100% for any sufficiently large N, a distinctive feature that casts serious doubt on any limited-capacity model (Bays & Husain, 2009).

The predictions of the slots-plus-averaging model for threshold versus set size are shown, both for 1-of-N and for N-of-N, in Figure 7b. In order to give each value of K a fair chance to fit the data, we chose $\sigma_K = \sigma_{K=1}\sqrt{K}$, with $\sigma_{K=1} = 3°$. (Note that $\sigma_{K=1} = \sigma_{N=1}$.) With this choice, $\Delta_{\text{thr}}(1, 1) = 3°$ for any K, close to the measured value.

Experiments 4 and 5: D of N

When multiple trajectories deviate, the combinatorics get slightly more complicated, but the logic is the same. We assume that when multiple deviating trajectories receive a slot, the observer averages the corresponding observations (on top of averaging over multiple slots allocated to the same item, when $N < K$). The predictions are in Appendix A.

Bayesian model with capacity limit

In this model, there exists a capacity limit K, but inference is optimal within the selected subset of K items. We consider this model in order to determine whether a capacity limit on a Bayesian ideal observer can describe the data better than a continuous resource constraint. Predictions are derived in Appendix A.

Averaging model

Finally, we consider a simple model in which the observer makes a judgment by comparing the average motion directions across all dots before and after they pass the midline (Tripathy et al., 2007). Since all motion is to the right, a vector average of these directions is well approximated by a linear average. Therefore, this observer responds "$C = 1$" when

$$\sum_{i=1}^{N} \frac{y_i}{\sigma^2} > \sum_{i=1}^{N} \frac{x_i}{\sigma^2}. \quad (28)$$

This is equivalent to the summation model in signal detection theory, in which the signals and noise from local detectors get summed (Baldassi & Verghese, 2002; Graham, Kramer, & Yager, 1987). In fact, Equation 28 is identical to Equation 11, the optimal decision rule when all trajectories deviate (like there, we keep the $1/\sigma^2$ factor for generality). This rule is not optimal in the other experiments. However, since it relies solely on the computation of a single global motion signal, it does not require the tracking of individual dots. Therefore, if human observers would be following this strategy, one could even

question whether the task under study is a tracking paradigm.

Tripathy and colleagues (2007) conducted an experiment to rule this model out. In this experiment, there were always two targets, both deviating at either 38° or −38°. Set size was 6, 8, or 10. However, the distractors (all 4, 6, or 8 of them) also deviated, half of them clockwise by a fixed angle, and half of them counterclockwise by the same angle. This fixed angle was varied by block and could take the values 0°, 19°, 38°, 57°, and 76°. The task was to determine the direction of deviation of the two targets. If human observers were averaging the motion directions before and after the midline, then performance should not be affected by the angle of deviation of the distractors since the mean and variance of the average are not affected. However, percentage correct was found to decrease monotonically with the fixed deviation angle of the distractors.

In this paper, we add to this evidence by including the averaging model in our model comparisons. Note that "averaging" in this model has a different meaning than in the slots-plus-averaging model; in the former, averaging means pooling over *all* items to extract a global signal, while in the latter, averaging is over the observations provided by multiple discrete slots allocated to the *same* item.

Results

Experiment 1: Near-threshold, *N* of *N*

The data show that deviation threshold is more or less independent of N, with a value of about 3° (Figure 8a, green line). The slots-plus-averaging model (Figure 8c) and the constrained Bayesian model (Figure 8f) predict the same. The latter is because the benefit of averaging over N observations is canceled by the increase of uncertainty with N, as explained in the Theory and methods section. The traditional limited-capacity model predicts that threshold is exactly zero for any N (Figure 8b). The unconstrained Bayesian model and the averaging model predict that threshold decreases as $1/\sqrt{N}$ (Figures 8d and 8g), reflecting only the benefit of averaging without an increase in uncertainty. The Bayesian model with capacity limit predicts the same decrease for $N \leq K$, and a constant threshold for $N > K$ (Figure 8e). The traditional limited-capacity model, the unconstrained Bayesian model, the Bayesian model with capacity limit, and the averaging model can be ruled out based on this experiment.

Experiment 2: Near-threshold, 1 of *N*

The data show that the deviation threshold increases rapidly with N, taking values of more than 30° at $N = 4$ (Figure 8a, red line). The three limited-capacity models (traditional capacity limit, slots-plus-averaging, unconstrained Bayesian) predict that threshold will rise to infinity at set sizes exceeding $1.46K$ (Equation 23; Figures 8b, 8c, and 8e). This occurs because performance, even for very large Δ, is limited by the fact that a subset of size K is chosen and all other items are ignored. Indeed, asymptotic proportion correct is $(1 + K/N) / 2$ (Equations 22 and 25, Figures 6b and 7a). Moreover, for smaller values of N, the limited-capacity model and the slots-plus-averaging model predict no or only a slow increase of threshold with set size, because they ignore uncertainty in stimulus representation (Figure 8b) or uncertainty about which trajectory deviates (Figures 8b and 8c).

The slots-plus-averaging model has threshold grow as $\sqrt{N}$ for $N \leq K$ (Equation 27) and predicts $\Delta_{thr} = 9°$ when $K = 3$ and $N = 4$ (Equation 26), which is far from the observed value. The effects of changing K were explored in Figures 6c and 7b. According to the averaging model, threshold grows as $\sqrt{N}$ throughout (Figure 8g).

Both the unconstrained and the constrained Bayesian model describe these data better (Figures 8d and 8f). The constrained model predicts the fastest increase of all models due to the increase of uncertainty with set size. However, the increase in the data is still not completely explained. This model comparison would be aided by data at $N = 5$. In the Bayesian models, an increased jitter in the initial motion directions leads to an increase in threshold, as was found but not expected in (Tripathy & Barrett, 2004).

Experiment 3: Suprathreshold, 1 of *N*

The data show that percentage correct declines smoothly with set size for any given value of Δ (Figure 9a). The curves are clearly separated for different values of Δ. The limited-capacity model predicts no separation between the curves, regardless of the value of K (Figure 9b).

The slots-plus-averaging model and the Bayesian model with capacity limit behave very similarly in this experiment (Figures 9c and 9e). Both are limited-capacity models for $K \geq N$, and thus the difference between performance and chance drops as $1/N$ (see Equations 25 and A12). They predict an abrupt transition at $N = K$ for $\Delta = 76°$, contrary to the data. Both also predict a much smaller separation between the curves at $\Delta = 38°$ and $\Delta = 76°$ than is indicated by the data. In the unconstrained Bayesian model and the averaging model, performance decline is too slow to fit the data (Figures 9d and 9g); this is because uncertainty is constant with set size. Only the constrained Bayesian model describes the dependence of performance on N and Δ reasonably well (Figure 9f) because of the increase of uncertainty with set size. All models, except for the otherwise implausible limited-capacity model, underpredict performance at $N = 1$ and $\Delta = 19°$.

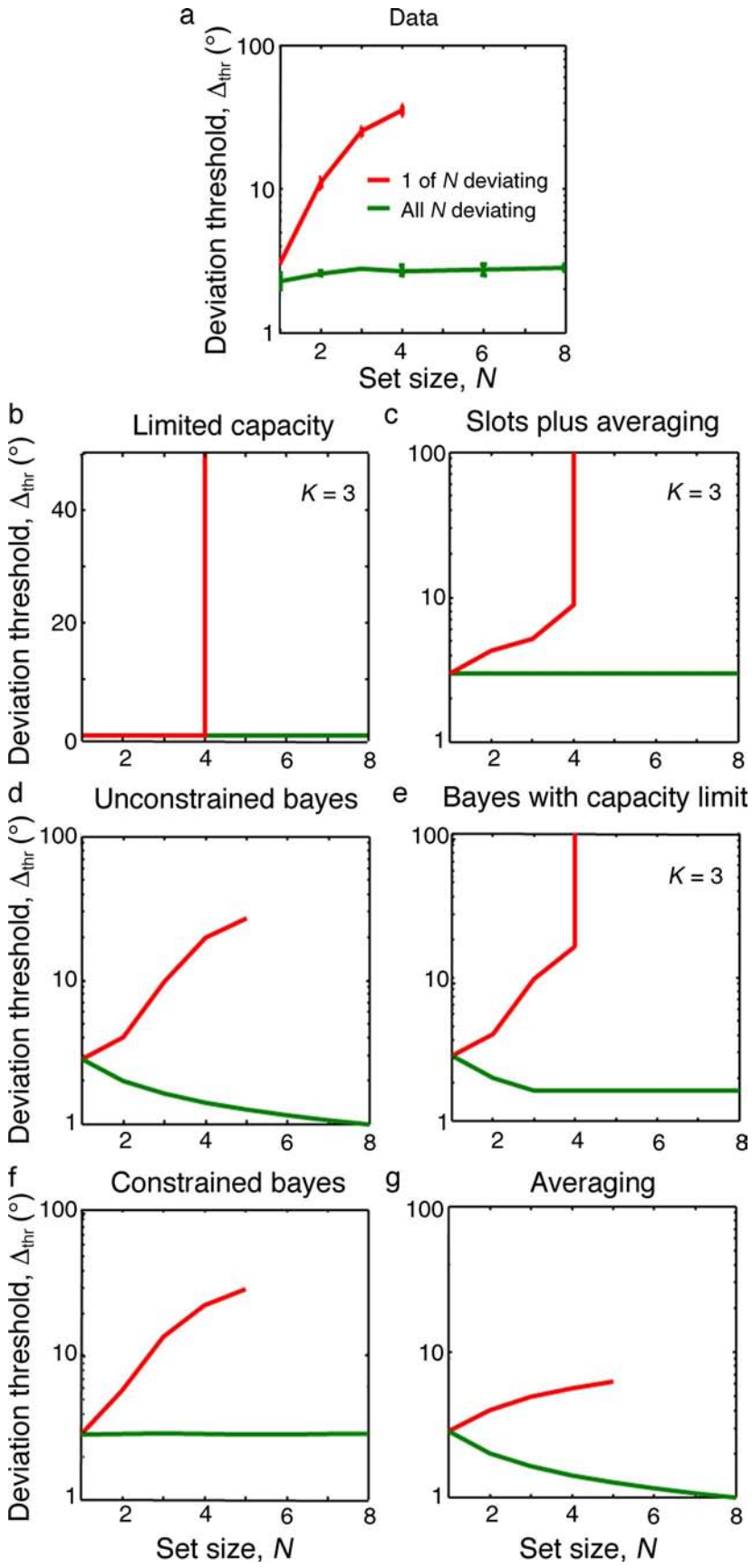

Figure 8. Model comparison for Experiments 1 and 2. (a) Deviation threshold at 84.1% correct versus set size, replotted from Tripathy and Barrett (2004) (error bars are *SEM,* 2 subjects). (b–g) Model predictions.

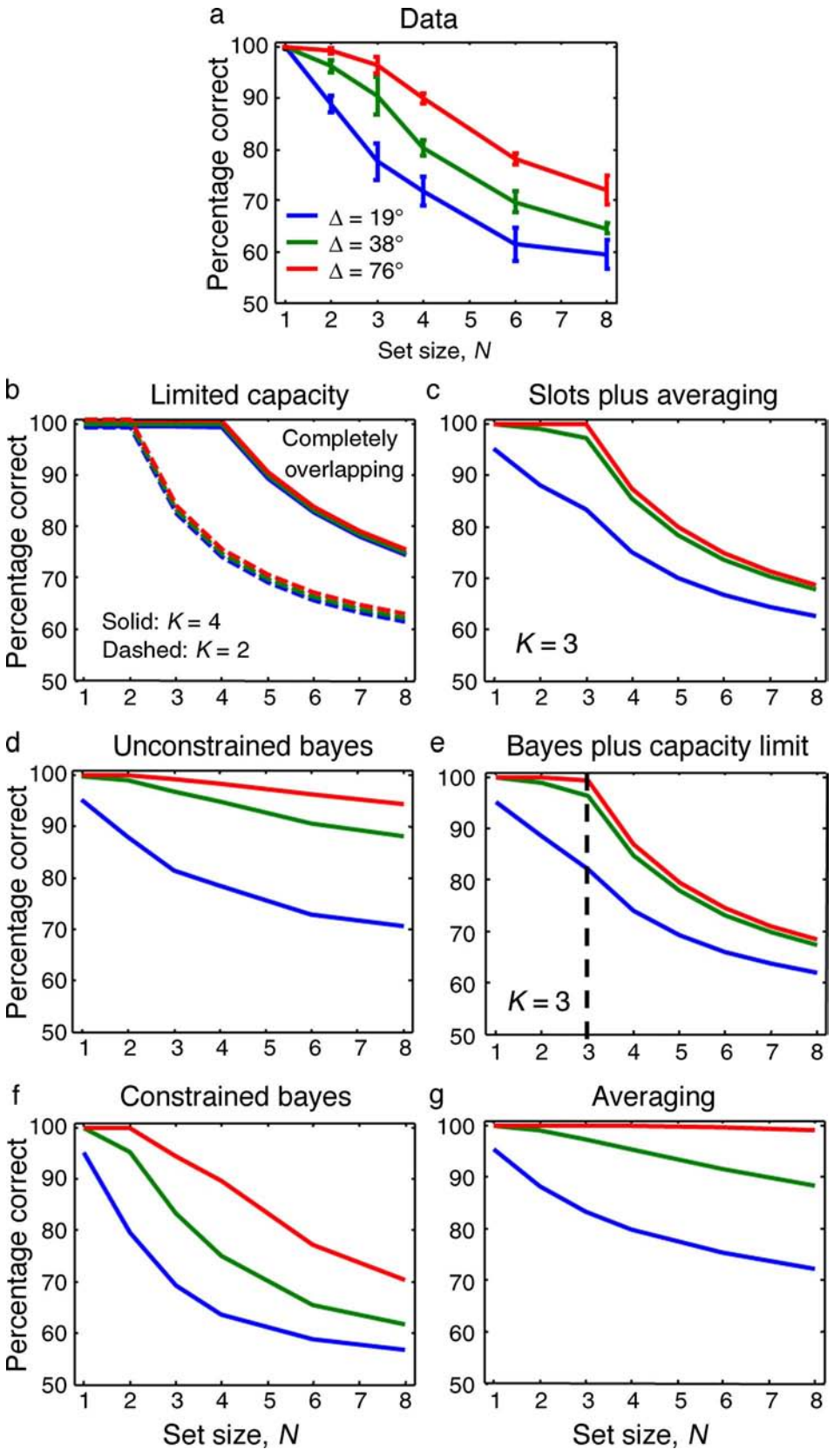

Figure 9. Model comparison for Experiment 3 (suprathreshold, 1 trajectory deviating). Percentage correct as a function of the number of trajectories, for three different angles of deviation, Δ. (a) Data replotted from Tripathy et al. (2007) (three subjects). (b–g) Model predictions. In panel b, for a given *K,* all curves overlap but have been separated slightly for visibility. In panel e, the vertical dashed line (at the capacity limit, $N = K$) marks the transition from unconstrained Bayesian behavior to limited-capacity behavior.

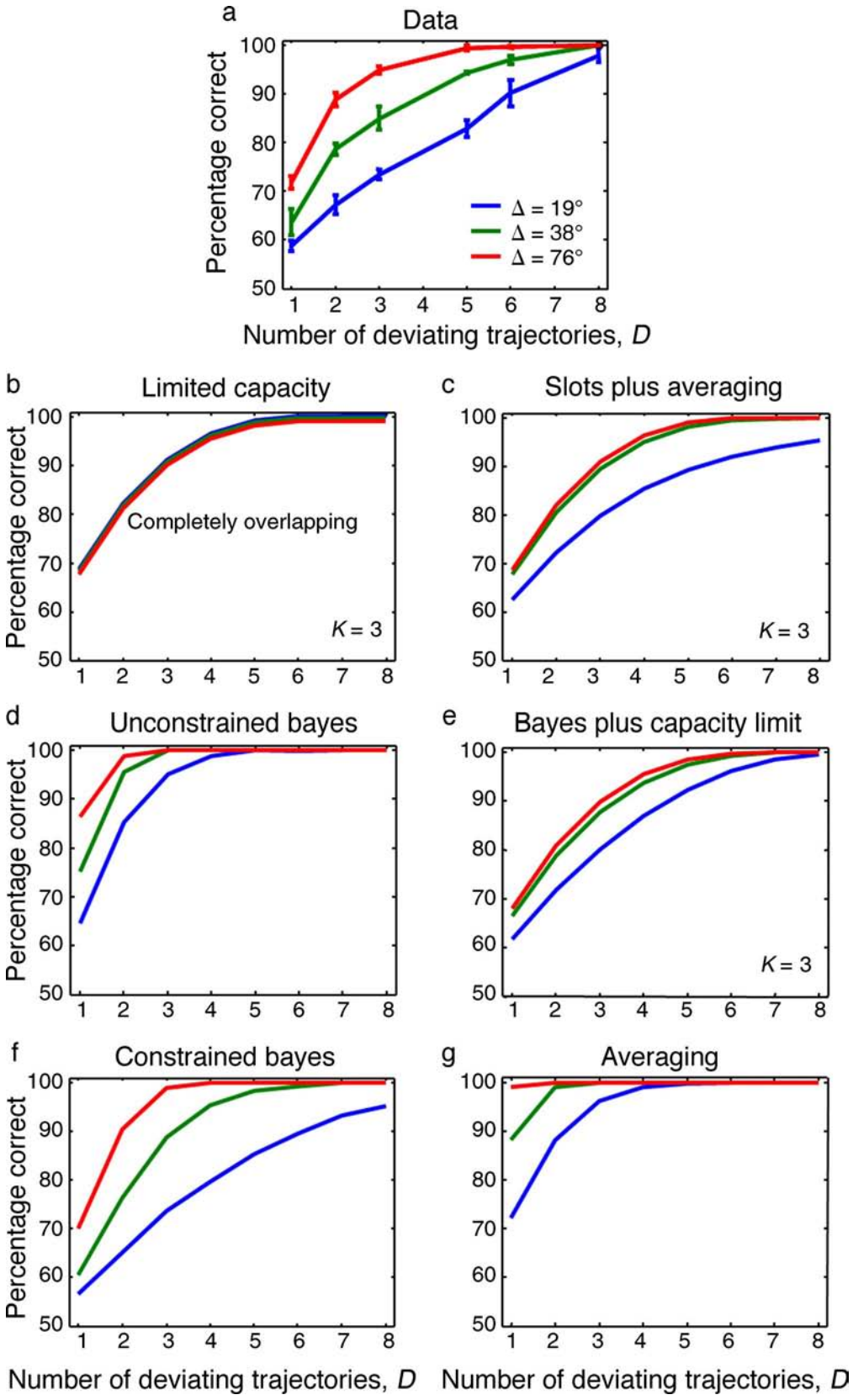

Figure 10. Model comparison for Experiment 4 (suprathreshold, $N = 8$, D trajectories deviating, trials blocked). Percentage correct as a function of the number of deviating trajectories, at $N = 8$, for three different angles of deviation, Δ. (a) Data replotted from Tripathy et al. (2007) (three subjects). (b–g) Model predictions. In panels b, c, and e, $K = 3$. In panels b and c, overlapping curves have been separated slightly for visibility.

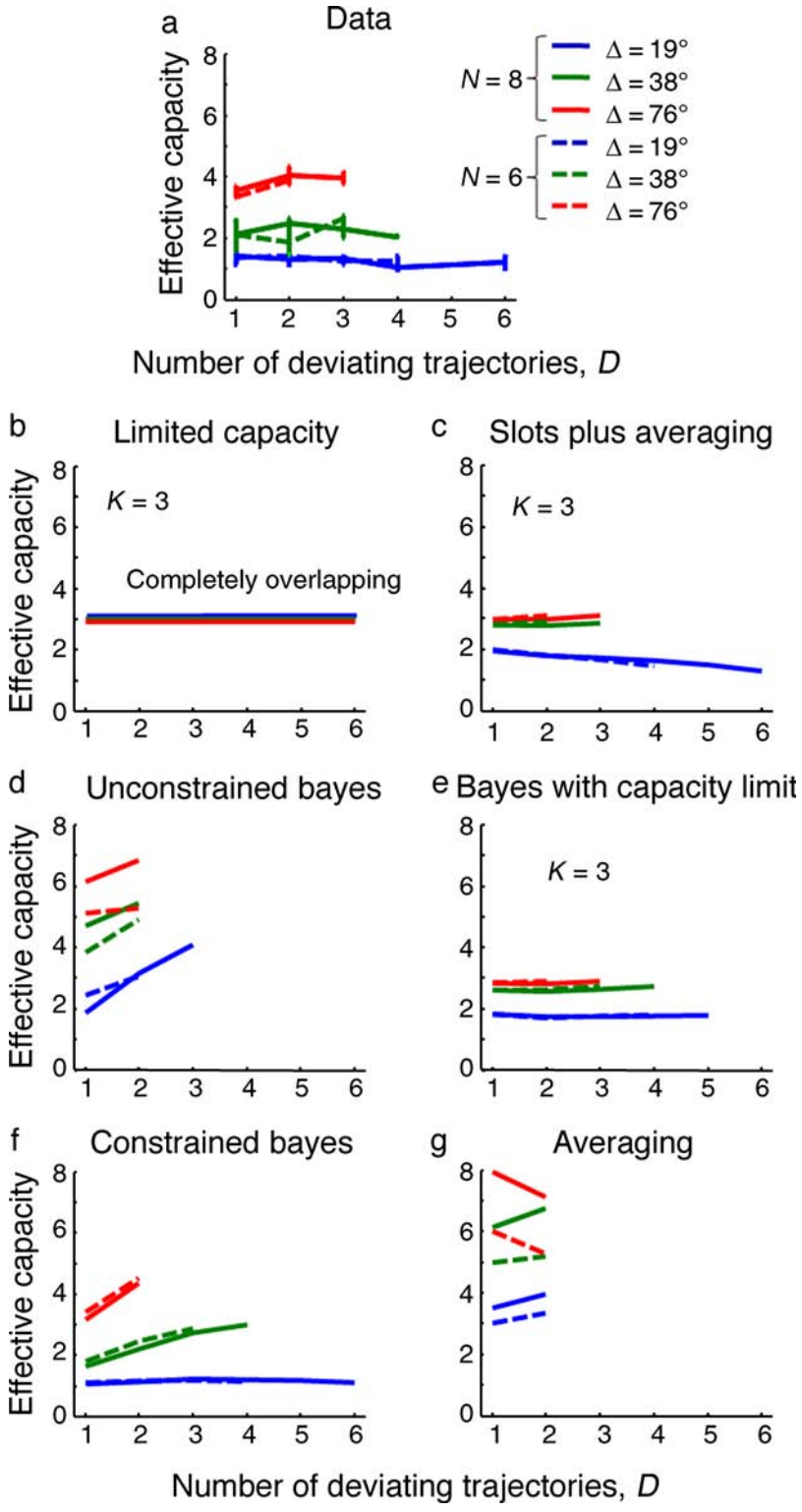

Figure 11. Effective number tracked versus number of deviations, *D,* in a suprathreshold paradigm, for set sizes $N = 6$, 8, and for different angles of deviation. Inset: percentage correct versus D at $N = 8$, used to compute effective number tracked. (a) Data replotted from Tripathy et al. (2007). (b–g) Model predictions. In panels c–f, $D > 2$ points are not shown when model performance exceeds 96%, since effective number tracked cannot be reliably estimated then.

Experiment 4: Suprathreshold, *D* of *N*, blocked

Percentage correct increases smoothly with the number of deviating trajectories (Figure 10a). Again, the data show a clear separation of the curves for different values of Δ.

The limited-capacity model predicts no separation between the curves, regardless of the value of K (Figure 10b). The slots-plus-averaging model (Figure 10c) and the Bayesian model with capacity limit (Figure 10e) are again virtually identical in this experiment, but both predict too little separation between the curves at Δ = 38° and Δ = 76°, as does the unconstrained Bayesian model. Limited-capacity models predict that performance is independent of Δ at least for sufficiently large Δ (Figures 10b, 10c, and 10e), while the unconstrained Bayesian model and the averaging model overestimate performance (Figures 10d and 10g). The constrained Bayesian model reproduces the correct values and dependencies (Figure 10f), notably only with two free parameters. Since N is fixed in this experiment, the differences between models predictions arise from the numerical value of σ_1 rather than on the dependence of σ on N. σ_1 was taken to be 11.3°, in rough accordance with a separate experiment (see Appendix B). A larger value would allow the slots-plus-averaging model to fit better than it does now but be less consistent with that experiment.

Effective capacity, the capacity of an equivalent limited-capacity observer, is an alternative way of expressing these results. The data show that effective capacity depends on the angle of deviation, but not much on set size or number of deviations (Figure 11a). Again, this is reproduced by the constrained Bayesian model (Figure 11f) and not as well by the other models (Figures 11b–11e and 11g). Effective capacity is sensitive to small changes in proportion correct, especially near ceiling. Following the experiment (Tripathy et al., 2007), points were left out when performance was near ceiling. The absence of a point at D = 3 and Δ = 76° in Figure 11f indicates that the model overestimates performance there.

Experiment 5: Suprathreshold, *D* of *N*, interleaved

Trials with different values of Δ (19°, 38°, 57°) and different values of D (1, 2) were interleaved, with N = 10.

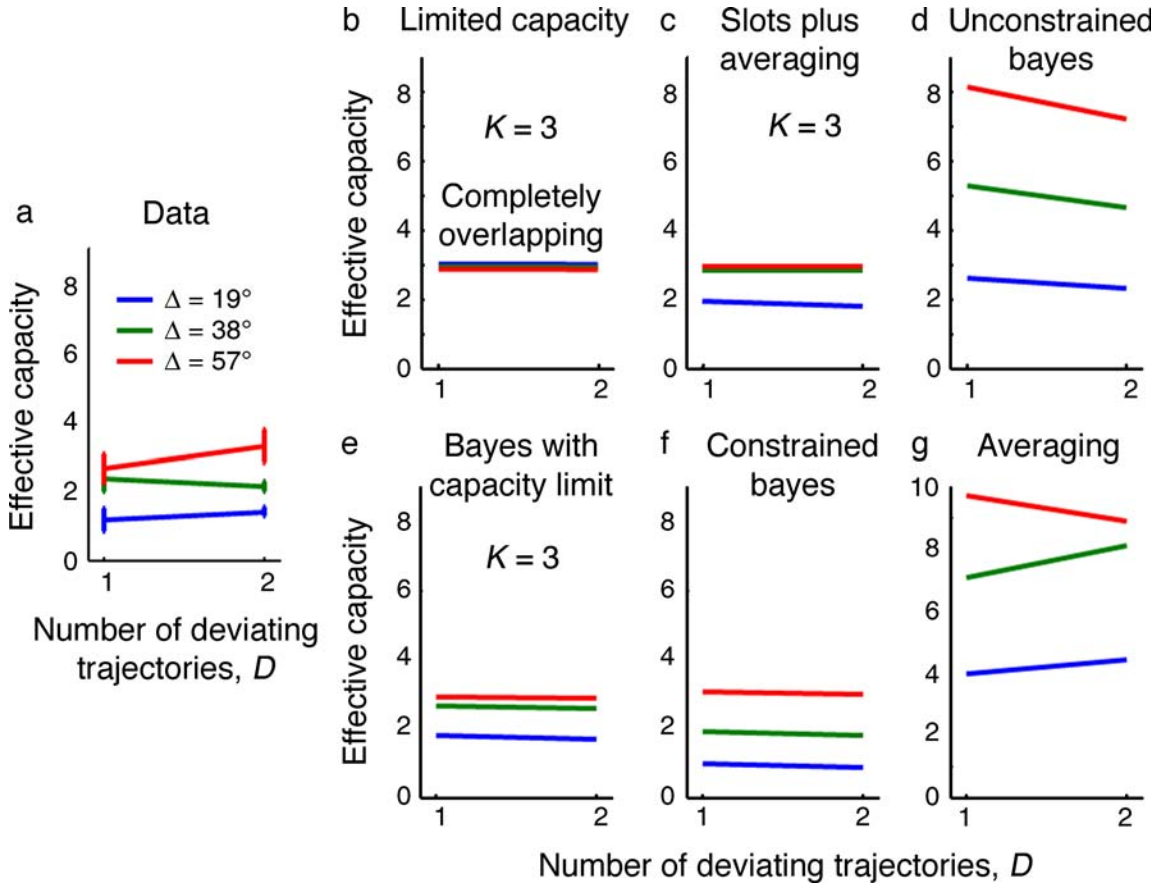

Figure 12. Model comparison for Experiment 5 (suprathreshold, N = 10, D trajectories deviating, trials interleaved). Effective number tracked as a function of the number of deviating trajectories, for N = 10 and three different angles of deviation, Δ. Different values of Δ, and different values of D, were interleaved in this experiment. (a) Data replotted from Tripathy et al. (2007) (three subjects). (b–g) Model predictions. In panels b, c, and e, K = 3. In panels b and c, overlapping curves have been separated slightly for visibility.

Model	Number of free parameters	Free parameters
Traditional limited capacity	1	K
Slots plus averaging	2	K, σ_K
Unconstrained Bayes	2	σ_1, $\sigma_{\text{pos},1}$
Bayes with capacity limit	3	K, σ_1, $\sigma_{\text{pos},1}$
Constrained Bayes	2	σ_1, $\sigma_{\text{pos},1}$
Averaging	1	σ_1

Table 1. Free parameters of the models considered.

The limited-capacity and slots-plus-averaging models behave the same as in Experiment 4. The Bayesian models do not, since Bayesian observers make use of the statistical structure of the task. In Experiment 5, the observer does no longer know Δ or D on a trial-by-trial basis. The key point of this experiment is again that effective capacity depends on the magnitude of the change, contrary to the prediction of the traditional limited-capacity model. The values and the separation of the data are best matched by the constrained Bayesian model (Figure 12).

Free parameters

Better model fits are sometimes caused by a larger number of free parameters. This is not the case here. The free parameters of the different models are as in Table 1. Thus, the constrained Bayesian model has a number of free parameters that are smaller than or equal to that of three of the five alternative models. The models with fewer parameters, the traditional limited-capacity model and the averaging model, can be ruled out on the basis of their poor fits to the data.

The number of parameters is low in all models because the models do not incorporate the details of the task, such as the time course of the trajectories and their intersections, or the effect of eccentricity. However, since all trends in the data are well reproduced by the constrained Bayesian model, it is likely that it captures the key factors that determine performance.

Predictions

Although we found that the constrained Bayesian model fits the available experimental data well, a critical test of its strength is whether it can make new predictions different from those of other models. Here, we make two predictions for experiments that have not yet been done, offering opportunities to falsify the constrained Bayesian model.

Prediction 1: Near-threshold, *D* of *N*

We would like to predict threshold as a function of the number of deviating trajectories, for a fixed total number of trajectories. This requires a different Bayesian decision rule than Experiment 4 because the latter was suprathreshold with Δ fixed and known.

The decision rule is derived in Appendix A and used to generate the prediction in Figure 13a. The dependence of threshold deviation on D at fixed N will exhibit a rapid decrease of threshold between $D = 1$ and $D = 2$. The slots-plus-averaging model predicts a very different pattern than the constrained Bayesian model, with the largest differences occurring at $D = 1$ (Figure 13b).

Prediction 2: Unequal reliabilities

The Bayesian models so far have assumed that the total uncertainty in the direction of each trajectory, σ, is the same for all trajectories. However, the theory is by no means restricted to this. Since Bayesian theory assumes that a probability distribution over each stimulus is encoded *on a single trial*, a Bayesian observer should be able to take into account uncertainty when making a perceptual decision (Knill & Pouget, 2004; Knill & Richards, 1996). In other words, if different trajectories within a single display come with different amounts of uncertainty, the more uncertain ones should be weighted less. A powerful test of Bayesian optimality is therefore to vary the reliability of different objects in the same display and predict performance. This is routinely done in cue combination studies.

To make this specific, we consider the suprathreshold paradigm with 1 of N trajectories deviating (Experiment 3). We assume that each trajectory comes with its own uncertainty, σ_j for the jth trajectory. The Bayesian decision rule is now a modified version of Equation 20, namely

$$\sum_{i=1}^{N} e^{-\frac{\Delta^2}{2\sigma_i^2}} \left(e^{\frac{\Delta(y_i - x_i)}{\sigma_i^2}} - e^{\frac{-\Delta(y_i - x_i)}{\sigma_i^2}} \right) > 0. \quad (29)$$

(obtained from the same derivation, but with a different σ_i for each trajectory). From this expression, it is clear that the extent to which a trajectory contributes to a decision is inversely proportional to its variance. This is interesting when different values σ_i are present within the same display.

Simpson's paradox

For clarity of explanation, we will first ignore positional uncertainty (imagine an experiment in which trajectories never intersect and are well separated). We assume two

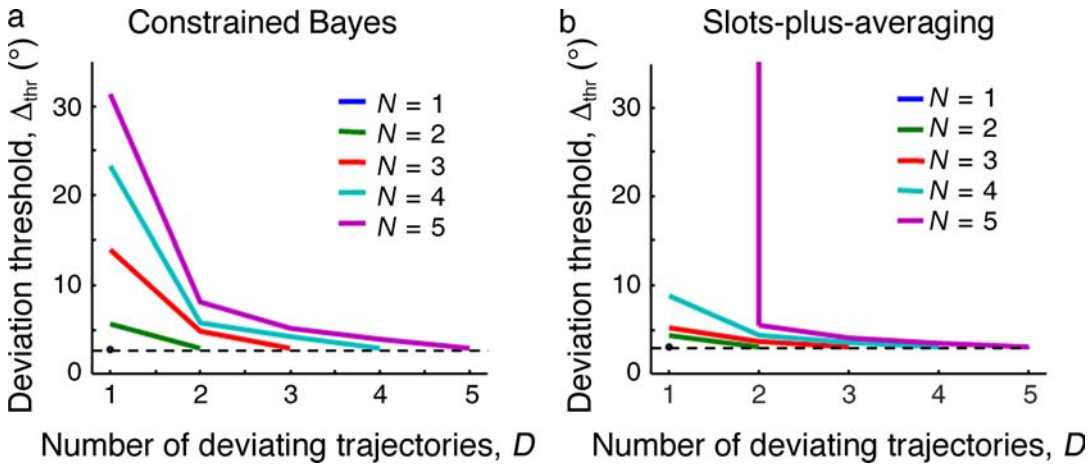

Figure 13. Predicted deviation threshold as a function of the number of deviating trajectories, D, for various set sizes. (a) Constrained Bayesian model. Since $D \leq N$, the blue curve ($N = 1$) consists of a single point. The dashed line indicates the common deviation threshold when N of N trajectories deviate. (b) Slots-plus-averaging model. For $N = 5$ and $D = 1$, threshold is infinite (see also Figure 8e).

levels of directional uncertainty: high and low—these could correspond to low and high contrast, respectively. We consider a suprathreshold discrimination experiment with 1 of N trajectories deviating (N fixed), where on each trial, we randomly assign high uncertainty (low contrast) to H items and low uncertainty (high contrast) to the remaining $N - H$ ones. This leads to displays in which the number of high-uncertainty items can be anywhere from 0 to N. We take $N = 6$ and $\Delta = 38°$ and call the single deviating trajectory the target. Then the blue line in the inset in Figure 14a indicates percentage correct as a function of H, according to the constrained Bayesian model (parameters were taken identical to those in Experiment 3). It is not surprising that overall performance declines monotonically with H, as high-uncertainty items contain less information. However, if we divide the trials into two classes by the uncertainty of the target, then class-conditioned performance *increases* with H. The reason this happens is that because in each class, since target uncertainty is fixed, all that changes when H increases is the number of high-uncertainty *distractors*. The Bayesian decision rule, Equation 29, suppresses evidence from high-uncertainty items, so if those items are distractors, this increases the relative contribution of the target and therefore the performance on the discrimination task.

How is it possible that percentage correct increases in each of two classes but decreases when the two classes are combined into a single data set? This is a well-known phenomenon known as Simpson's paradox (Yule, 1903). Overall percentage correct is not a fixed weighted average of the class-conditional percentages. At each value of H, the weights in the averaging are determined by the relative numbers of observations in each class at that value of H, but these proportions depend on H. The higher H, the higher the probability that the target has high uncertainty. This probability is H/N, giving increasing weight to the percentage for the high-uncertainty class as H increases. In an equation:

$$\mathrm{PC}(\Delta, N \geq K, 1, H) = \\ = \frac{H}{N}\mathrm{PC}_{\mathrm{high}}(\Delta, N \geq K, 1, H) + \left(1 - \frac{H}{N}\right)\mathrm{PC}_{\mathrm{low}}(\Delta, N \geq K, 1, H). \quad (30)$$

In the presence of positional uncertainty, an additional effect is that high-uncertainty items can corrupt the information from low-uncertainty ones. The combination of both effects leads to performance curves like those in Figure 14a (main figure). We still see that between $H = 3$ and $H = 5$, conditioned performance increases while combined performance decreases or stays constant. Moreover, percentage correct is distinctly non-flat as a function of H.

Finally, we compute the prediction of the slots-plus-average model for this hypothetical experiment. We assume that K is given, $K < N$, and that the internal representation of a particular item has standard deviation $\sigma_{K,\mathrm{low}}$ when uncertainty is low and $\sigma_{K,\mathrm{high}}$ when uncertainty is high. Conditioned performance $\mathrm{PC}_{\mathrm{low}}$ and $\mathrm{PC}_{\mathrm{high}}$ are obtained by using $\sigma_{K,\mathrm{low}}$ and $\sigma_{K,\mathrm{high}}$ in Equation 25, respectively. Neither depends on H, since this model assumes that the observer has perfect knowledge about which trajectory, if any, changed among the attended ones. Unconditioned performance is again a weighted average of the conditioned performances. This produces the prediction of Figure 14b, in which conditioned performance is independent of H. Conducting this experiment could

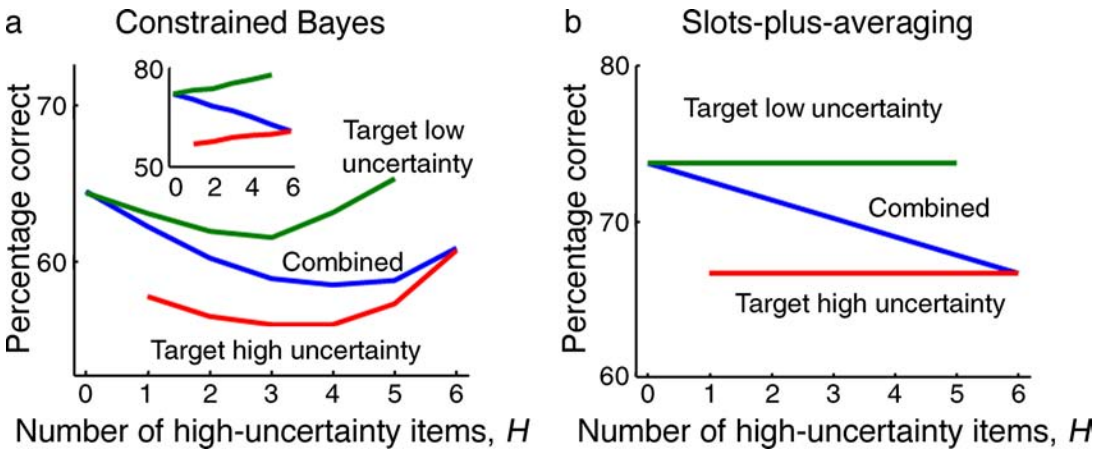

Figure 14. Predicted percentage correct predicted as a function of the number of high-uncertainty items, *H,* and conditioned on the target having low (green) or high (red) uncertainty, or not conditioned (blue). In this example, $N = 6$, $D = 1$, and $\Delta = 38°$. (a) Constrained Bayesian model. The inset shows the basic effect (Simpson's paradox) in the absence of positional uncertainty (positional uncertainty is normally present, though it might be reduced by avoiding trajectory intersections). (b) Slots-plus-averaging model. Conditioned performance is predicted to be independent of the number of high-uncertainty items.

provide further evidence to distinguish the constrained Bayesian model from the slots-plus-averaging model and possibly from other models.

Discussion

No item limit in tracking

We compared six models of human performance in tracking multiple dots in a deviation discrimination task, using data from five published experiments, both near-threshold and suprathreshold. The models featured various combinations of a noise model and a decision model. Three of them contained a fixed limit on the number of items that can be tracked, while the others did not. Our main finding was that a continuous increase of uncertainty with set size, derived from a limited number of available spikes, could explain the data better than a fixed item limit. Signatures of item-limit models include an infinite deviation threshold for sufficiently large set sizes and too small performance improvements as the deviation angle increases in suprathreshold conditions. We were able to rule out the recently proposed slots-plus-averaging model and a simple averaging model in which observers do not track individual items. We also eliminated the possibility that the observer was Bayesian without any constraints, or Bayesian with an item limit. The most successful model combines a spike constraint with otherwise optimal inference, in which a decision is based on the statistical structure of the task and the noisy sensory evidence available on a given trial. This model fits in a long line of probabilistic models of perception but adds to it by introducing the spike constraint.

Clearly, one could come up with hybrids between the models tested here. For example, an observer might select a subset of items and average their corresponding observations before and after the midline (instead of averaging just the target observations, as in the slots-plus-averaging model). Alternatively, an observer might be Bayesian but subject to both an item limit and a continuous increase in uncertainty, or average all observations subject to such an increase. Importantly, we cannot rule out that the data can be explained by a suboptimal decision rule coupled with an increase in uncertainty that scales differently from $\sqrt{N}$. However, any alternative model can now be tested against the benchmark set by the constrained Bayesian model. (Results not reported here suggest that the above alternatives are also inadequate.)

The model makes testable predictions for both behavior and neural activity. We predict the pattern of deviation thresholds as D of N trajectories deviate (Figure 13). We predict the occurrence of Simpson's paradox when dots are allowed to vary in contrast (Figure 14). These experiments are expected to yield additional evidence against several alternative models.

At the neural level, we expect that trajectories are encoded in populations whose gain scales roughly as $1/N$. A different population should encode the Bayesian decision variable, d, and therefore the posterior distribution over the binary variable C.

Broader scope

Most cognitive psychologists and psychology textbooks have long fancied limited-capacity models, both for attention and working memory, presumably because of their mathematical simplicity and their seeming inevitability

when dealing with categorical, high-level objects such as letters (Pashler, 1988). Limited-capacity models state that a few items are encoded noiselessly and all others are not encoded at all. Decades of work in psychophysics and neuroscience have convincingly demonstrated that the noiseless encoding of stimuli is patently absurd (Aldo Faisal, Selen, & Wolpert, 2008). In partial recognition of this absurdity and faced with contradictory evidence (Wilken & Ma, 2004), limited-capacity advocates recently dropped the noiselessness assumption by proposing the slots-plus-averaging model (Zhang & Luck, 2008). However, this model contained the equally unsubstantiated notion that noise level in working memory is controlled by the number of discrete "slots" assigned to a single item. Moreover, they failed to recognize the implications of variability for tasks with targets and distractors, such as change detection. When the representation of individual stimuli is variable, then whether an item is a target also becomes subject to uncertainty and has to be inferred probabilistically. In the visual search community, signal detection modelers have, for many years, been waging a battle to explain exactly this (e.g., Eckstein, Thomas, Palmer, & Shimozaki, 2000; Palmer et al., 2000; Verghese, 2001). In the realms of attentional tracking and working memory, the same realization is taking longer to materialize. The present work demonstrates that a proper probabilistic treatment of the decision process in a change discrimination task points to a continuous resource limitation rather than an item or slot limit.

Bayesian models are known to explain behavior in numerous tasks that require observers to infer one feature of one stimulus. Perceptual decisions are often more complex, requiring the combination of information from multiple stimuli to extract a global, higher-level variable (like C here). Our results provide new evidence that the brain also uses Bayesian inference in such judgments, however sometimes under a constraint of uncertainty increasing with set size. Moreover, our findings emphasize the parallel importance of feature and positional uncertainty. An important open question is why the attentional tracking task studied here is subject to an increase of uncertainty with set size, while visual search tasks do not seem to be (see also Palmer, 1990).

Relation to signal detection theory

The Bayesian models presented here should be viewed in the tradition of signal detection theory, which takes the noisy representation of stimuli as a given. However, the decision rules we derived differ from those typically used in signal detection theory models, such as the max and the sum rules. The Bayesian decision rules are optimal in any condition, while the max and sum rules are only optimal in certain limits and certain conditions. For example, the sum rule is optimal in Experiment 1 (see Equation 11; it still needs to be combined with a spike constraint to fit the data), and we pointed out below Equation 20 that the Bayesian decision rule in Experiment 3 is, in a limit, well approximated by the signed-max rule.

In general, an "easier" decision rule might often be a decent and convenient approximation to the optimal decision rule, as was pointed out long ago (Nolte & Jaarsma, 1966). The goodness of such approximations in the decisions we modeled is an interesting and central issue. These approximations might even be indicative of the approach the brain takes (Verghese, 2001). However, this argument should be treated with caution, as there is no reason why an approximation that is convenient for a signal detection modeler is convenient for a realistic neural network.

It can be argued that signal detection theory, although probabilistic across many trials, is ultimately a theory of point estimates. On a single trial, a measurement of a single one-dimensional stimulus is represented by a single number. In contrast, Bayesian theory states that on a single trial, the brain represents a full posterior distribution (degree of belief versus hypothesized stimulus value) over each stimulus (Knill & Pouget, 2004). This allows for the automatic encoding of uncertainty, through the width of this probability distribution. This uncertainty, σ, appears in many optimal decision rules, such as those in cue combination (Knill & Richards, 1996). In the present paper, Equations 18 and 20 are examples. Encoding uncertainty on a trial-to-trial and item-to-item basis is especially critical when uncertainty differs between items (e.g., some dots have lower contrast than the others) or between trials (e.g., contrast is changed between trials). This was not explored in the experiments modeled here, but we made a prediction about it (Prediction 2) and it constitutes an important future direction. (Point estimates of uncertainty might be used instead of posterior distributions, but this proposal requires more details, and clues as to how it generalizes to non-Gaussian distributions.)

Limitations

The constrained Bayesian model differs from the data in several places. The increase of threshold with set size in Experiment 2 (Figure 2f) is not as steep as the data. Percentage correct at $N = 1$ and $\Delta = 19°$ is predicted to be below 100% in Experiment 3 (Figure 9f), in contrast to the data. Effective capacity at higher deviation angles is overestimated in Experiment 4 (Figure 11f).

Some of these discrepancies may stem from the fact that many task details have been ignored to keep the model simple. We conveniently collapse the observer's judgment into a static one, even though the time course of the trajectories is probably relevant. By not modeling temporal dynamics, errors resulting from confusing trajectories where they intersect are possibly underestimated (e.g., in Figure 2f). Also, possible effects of pursuit eye

movements, eccentricity, perceptual grouping (Yantis, 1992), and noise correlations between items are not incorporated. Taking these aspects into account would require a model with many more parameters. Our purpose here has been to capture the essence of the task with a minimal set of assumptions, and a number of parameters that is equal or comparable to the alternative models. Our model accounts for a large set of both near-threshold and suprathreshold data despite having only two free parameters. We believe that more detailed, task-specific models could provide somewhat better fits to the data but would not change the qualitative conclusions.

Priors

As is common in Bayesian modeling, priors are both a strength and a weakness. Priors reflect background knowledge and should be incorporated, but it is hard to know precisely what prior distribution an observer uses. Sometimes, one can assume reasonable priors that correspond, or may correspond, to natural statistics (e.g., light-from-above prior, prior for low speeds). In other cases, the best approach is to use priors that correspond to experimental frequencies of stimuli in the task at hand. Here, we have taken the second approach since we do not know of natural statistics for variables like number of deviations or deviation angle. Moreover, the observers in these experiments often had knowledge of the experimental parameters. It will be interesting to explore whether the priors have a significant impact on the model predictions, and if so, whether the prior probabilities used by the observer can be manipulated experimentally. Tripathy and colleagues (2007) already started doing this by varying the frequencies of the different deviation angles in Experiment 5, but they found no effect.

Neural implementation

Even though the Bayesian decision variables and decision rules can get rather complex, this does not mean they cannot be implemented in a neurally plausible way. At the behavioral level, computations are performed on probability distributions over task-relevant variables. At the neural level, computations are performed on population patterns of activity. The mapping between the operations at both levels is by no means trivial and has only begun to be explored (Beck et al., 2008; Huys, Zemel, Natarajan, & Dayan, 2007; Ma et al., 2006). It is likely that complex decision rules that feature a combinatorial explosion, like in Equation A7, will have to be implemented through an approximate algorithm. Work on the neural basis of Bayes-optimal visual search (Ma, Navalpakkam, Beck, & Pouget, 2008; Vincent et al., 2009) might provide clues, as it is another task where a global binary variable is inferred from multiple stimuli.

We have argued that the neural constraint leading to the $\sqrt{N}$ increase of uncertainty with set size can be implemented using divisive normalization, a mechanism believed to play a central role in attentional processing (Reynolds & Heeger, 2009). A full neural model will have to integrate this mechanism, which presumably acts at the input level, with the mechanism of the probabilistic inference process.

Standard multiple-object tracking

There are differences between the task modeled here and the standard multiple-object tracking paradigm (Pylyshyn & Storm, 1988). In the standard paradigm, a number of objects are labeled as targets at the start of a trial and the task is to report later whether a probed item is a target or not. Non-target items should be ignored. In the task we model, all items have to be tracked during the first half of the trial, since it is not known which ones will deviate. Another difference is that trajectories are linear in the task we model, and much more complex in the standard paradigm. Since these differences might be important, we make no claims about the standard multiple-object tracking paradigm. However, even in the standard multiple-object tracking paradigm, which has complex temporal dynamics, approximate Bayesian models with a flexible resource constraint might still explain behavior best (Vul, Frank, Alvarez, & Tenenbaum, 2009).

Attention or memory?

Change detection is mostly used as a paradigm for studying working memory since past and present information has to be combined over periods of several seconds. Unlike working memory experiments, the task we model does not feature a delay period; however, working memory might still plays a role. Similarities have been noted between standard multiple-object tracking and working memory (Cavanagh & Alvarez, 2005). In this context, it is relevant that there exists strong evidence for flexible-resource theories for visual working memory (Bays & Husain, 2008; Wilken & Ma, 2004). Alternatively, it is possible that the task studied here reveals limitations of sensory (or iconic) memory rather than attention (Alvarez & Franconeri, 2007; Narasimhan, Tripathy, & Barrett, 2009). However, this would not alter the theories we compare. In fact, the approach presented here makes a model-based comparison between attentional tracking and various memory systems possible.

Appendix A

Miscellaneous model predictions

Bayesian model

Experiment 4: Suprathreshold, D of N, blocked

Following the same logic as in the Theory and methods section (Experiment 2), we find for the Δ-conditioned likelihood

$$p(\mathbf{x},\mathbf{y}|C,\Delta) = \\ = \alpha\binom{N}{D}^{-1}\left(\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-x_i)^2}{2\sigma^2}}\right)e^{-\frac{D\Delta^2}{2\sigma^2}}\sum_{|\mathbf{I}|=D}\prod_{j\in\mathbf{I}}e^{\frac{C\Delta(y_j-x_j)}{\sigma^2}}, \tag{A1}$$

where $\mathbf{I}$ is a set of trajectory indices, $|\mathbf{I}|$ is the number of elements in this set, and $\Sigma_{|\mathbf{I}|=D}$ is shorthand for "sum over all index sets $\mathbf{I}$ of size D." The log odds are

$$d = \log\sum_{|\mathbf{I}|=D}\prod_{j\in\mathbf{I}}e^{\frac{\Delta(y_j-x_j)}{\sigma^2}} - \log\sum_{|\mathbf{I}|=D}\prod_{j\in\mathbf{I}}e^{-\frac{\Delta(y_j-x_j)}{\sigma^2}}. \tag{A2}$$

The decision rule is to report "C = 1" if

$$\sum_{|\mathbf{I}|=D}\left[\exp\left(\frac{\Delta}{\sigma^2}\sum_{j\in\mathbf{I}}(y_j-x_j)\right)-\exp\left(-\frac{\Delta}{\sigma^2}\sum_{j\in\mathbf{I}}(y_j-x_j)\right)\right]>0. \tag{A3}$$

Experiment 5: Suprathreshold, D of N, interleaved

We assume that the deviation angle Δ is drawn from a discrete distribution with A possible values, $\Delta_1,\ldots,\Delta_A$, all with equal probability, $1/A$. Similarly, D is drawn from a discrete distribution with B possible values, $D_1,\ldots,D_B$, also all with equal probability, $1/B$. The derivation starts with Equation 5 applied to a discrete distribution, $p(\Delta)$:

$$p(\mathbf{x},\mathbf{y}|C) = \frac{1}{A}\sum_{a=1}^{A}p(\mathbf{x},\mathbf{y}|C,\Delta_a). \tag{A4}$$

Next, we integrate over the other top-level variable, D:

$$p(\mathbf{x},\mathbf{y}|C) = \frac{1}{AB}\sum_{a=1}^{A}\sum_{b=1}^{B}p(\mathbf{x},\mathbf{y}|C,\Delta_a,D_b). \tag{A5}$$

Evaluating this leads to the following log odds:

$$d = \log\frac{\sum_{a=1}^{A}\sum_{b=1}^{B}\binom{N}{D_b}^{-1}e^{-\frac{D_b\Delta_a^2}{2\sigma^2}}\sum_{|\mathbf{I}|=D_b}\prod_{i\in\mathbf{I}}e^{\frac{\Delta_a(y_i-x_i)}{\sigma^2}}}{\sum_{a=1}^{A}\sum_{b=1}^{B}\binom{N}{D_b}^{-1}e^{-\frac{D_b\Delta_a^2}{2\sigma^2}}\sum_{|\mathbf{I}|=D_b}\prod_{i\in\mathbf{I}}e^{-\frac{\Delta_a(y_i-x_i)}{\sigma^2}}}. \tag{A6}$$

The decision rule is

$$\sum_{a=1}^{A}\sum_{b=1}^{B}\binom{N}{D_b}^{-1}e^{-\frac{D_b\Delta_a^2}{2\sigma^2}}\sum_{|\mathbf{I}|=D_b}\left[e^{\frac{\Delta_a}{\sigma^2}\sum_{i\in\mathbf{I}}(y_i-x_i)}-e^{-\frac{\Delta_a}{\sigma^2}\sum_{j\in\mathbf{I}}(y_i-x_i)}\right]>0. \tag{A7}$$

This decision rule is applied to permuted vectors $\mathbf{x}_{\text{new}}$ and $\mathbf{y}_{\text{new}}$, as discussed in the Theory and methods section. When $A = 1$ and $B = 1$, Equation A7 is the same as Equation A3.

Prediction 1

In a threshold paradigm, Δ is averaged out because it is a random variable of unknown value. However, the calculation of the Bayesian decision variable is the same as that in Experiment 4, up until the point where that average is taken. Therefore, we start with Equation A1 and average over Δ, assuming a uniform prior distribution $p(\Delta)$. The result of this is:

$$p(\mathbf{x},\mathbf{y}|C) = \alpha\beta\binom{N}{D}^{-1}\left(\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-x_i)^2}{2\sigma^2}}\right)\sigma\sqrt{\frac{\pi}{2}} \\ \cdot\sum_{|\mathbf{I}|=D}e^{\frac{1}{2D\sigma^2}\left(\sum_{j\in\mathbf{I}}(y_j-x_j)\right)^2}\left(1+\operatorname{erf}\frac{\sum_{j\in\mathbf{I}}(y_j-x_j)}{\sigma C\sqrt{2D}}\right). \tag{A8}$$

The Bayesian decision rule, $d > 0$, simplifies to

$$\sum_{|\mathbf{I}|=D}e^{\frac{1}{2D\sigma^2}\left(\sum_{j\in\mathbf{I}}(y_j-x_j)\right)^2}\operatorname{erf}\frac{\sum_{j\in\mathbf{I}}(y_j-x_j)}{\sigma\sqrt{2D}}>0. \tag{A9}$$

Slots-plus-averaging model

Experiments 4 and 5

We assume that when multiple deviating trajectories receive a slot, the observer averages over all of them.

We first consider the case $N \geq K$. Then each trajectory can receive no more than one slot, and we have to calculate the probability that M deviating trajectories receive a slot. This is the same problem as: you draw K balls from a vase that contains D blue and $N - D$ red balls. What is the probability of drawing M blue balls and $K - M$ red ones? The answer is given by the hypergeometric probabilities, $\binom{D}{M}\binom{N-D}{K-M} / \binom{N}{K}$. Here, M is restricted to be $\max(K + D - N, 0) \leq M \leq \min(K, D)$. When M deviating trajectories receive a slot, uncertainty is $\sigma_K/\sqrt{M}$, therefore proportion correct is $\frac{1}{2} + \frac{1}{2}\,\mathrm{erf}(\Delta/\sigma_K\sqrt{M/2})$. Overall proportion correct is this value averaged over M, with weight factors given by the hypergeometric probabilities:

$$\mathrm{PC}(\Delta, N \geq K, D) = \\ = \frac{1}{2} + \frac{1}{2}\binom{N}{K}^{-1} \sum_{M=\max(K+D-N,0)}^{\min(K,D)} \binom{D}{M}\binom{N-D}{K-M}\mathrm{erf}\,\frac{\Delta}{\sigma_K}\sqrt{\frac{M}{2}}. \quad \text{(A10)}$$

When $D = 1$, this is equal to Equation 25. Next, we consider the case $N \leq K$. In this case, each trajectory receives at least one slot. Specifically, the average number of slots per trajectory is K/N. As a consequence, the uncertainty in the representation of a single trajectory will be $\sigma_N = \sigma_K\sqrt{N/K}$. Moreover, all D deviating trajectories will be represented with this uncertainty. Therefore, averaging over these D observations yields a representation with standard deviation $\sigma_K\sqrt{N/KD}$. Percentage correct is then

$$\mathrm{PC}(\Delta, N \leq K, D) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\,\frac{\Delta}{\sigma_K}\sqrt{\frac{KD}{2N}}. \quad \text{(A11)}$$

This equation is valid regardless of whether different values of Δ and D are blocked (as in Experiment 4) or interleaved (as in Experiment 5).

Prediction 1

For Prediction 1, we just use the proportions correct derived above. For $N \leq K$, it follows from Equation A11 that $\Delta_{\mathrm{thr}}(N \leq K, D) = \sigma_K\sqrt{N/KD}$. For $N \geq K$, we have Equation A10, so $\Delta_{\mathrm{thr}}(N \leq K, D)$ cannot be computed directly. Instead, we simply find the value of Δ for which percentage correct exceeds 84.1%. If this never happens (as always in a limited-capacity model when N is sufficiently large), then threshold is infinite. Figure 13b was obtained in this way, with $\sigma_{K=1} = 3°$.

Bayesian model with capacity limit

If $N \leq K$, all items are attended and performance is identical to that of the unconstrained Bayesian model (denoted by a subscript "UB"), $\Delta_{\mathrm{thr}}(N \leq K, D) = \Delta_{\mathrm{thr,UB}}(N \leq K, D)$. We therefore only consider the case $N \geq K$.

Experiment 1

When all trajectories are deviating, but only K of them are attended, the deviation threshold at $N \geq K$ should be equal to that at $N = K$, which is obtained from the unconstrained Bayesian model: $\Delta_{\mathrm{thr}}(N \geq K, N) = \Delta_{\mathrm{thr,UB}}(K, N)$.

Experiments 2 and 3

The deviating trajectory is attended with probability K/N. When it is attended, proportion correct is equal to that in the Bayesian model at set size K. Otherwise, performance is at chance. This yields the following expression for proportion correct for the Bayesian model with limited capacity:

$$\mathrm{PC}(\Delta, N \geq K, 1) = \frac{K}{N}\mathrm{PC_{UB}}(\Delta, K, 1) + \frac{1}{2}\left(1 - \frac{K}{N}\right), \quad \text{(A12)}$$

where $\mathrm{PC_{UB}}(\Delta, N = K)$ is proportion correct in the unconstrained Bayesian model at deviation angle Δ and set size K. Just like in all limited-capacity models, asymptotic performance is $\mathrm{PC}(\Delta = \infty, N, 1) = \frac{1}{2}(1 + K/N)$ and $\Delta_{\mathrm{thr}}(N, 1) = \infty$ when $N > 1.46K$. When $N/K \in [1, 1.46)$, threshold is finite but higher than in the unconstrained Bayesian model. Since $\mathrm{PC}(\Delta = \infty, N, 1) < 1$, we can no longer fit a cumulative normal distribution to PC (Δ) in order to estimate threshold. Instead, we have to fit a rescaled function, $\mathrm{PC}(\Delta) = \frac{1}{2} + \frac{1}{2}\,K/N \cdot \mathrm{erf}(\Delta/(\alpha\sqrt{2}))$, which has the correct asymptote. Comparing this with Equation A12, we find $\mathrm{PC_{UB}}(\Delta, K, 1) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}(\Delta/(\alpha\sqrt{2}))$, from which it is clear that the parameter α is nothing but the deviation threshold of the unconstrained Bayesian model at set size K, $\Delta_{\mathrm{thr,UB}}(K, 1)$. We find threshold from Equation A12:

$$\mathrm{PC}(\Delta_{\mathrm{thr}}) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\,\frac{1}{\sqrt{2}} \\ \Delta_{\mathrm{thr}}(N \geq K, 1) = \Delta_{\mathrm{thr}},\mathrm{UB}(N = K, 1)\sqrt{2}\,\mathrm{erf}^{-1}\left(\frac{N}{K}\mathrm{erf}\,\frac{1}{\sqrt{2}}\right). \quad \text{(A13)}$$

A sanity check: we find $\Delta_{\mathrm{thr}}(N, 1) = \Delta_{\mathrm{thr,UB}}(N, 1)$ if $N = K$; this is correct, since the capacity limit has no effect when all items are attended.

Experiments 4 and 5

When D of N trajectories are deviating, the logic is very similar to that of the slots-plus-averaging model. K

trajectories are picked at random to be attended, and of these, M will be deviating, where M is constrained by $\max(K + D - N, 0) \le M \le \min(K, D)$. When M of K trajectories are attended, performance is given by that of the unconstrained Bayesian model for M-of-K deviating trajectories:

$$\mathrm{PC}(\Delta, N \ge K, D) = \\ = \binom{N}{K}^{-1} \sum_{M=\max(K+D-N,0)}^{\min(K,D)} \binom{D}{M} \binom{N-D}{K-M} \mathrm{PC}_{UB}(\Delta, K, M). \quad \text{(A14)}$$

It is not immediately obvious that this reduces to Equation A12 when $D = 1$, but it does. (Note that $\mathrm{PC}_{\mathrm{UB}}(\Delta, N, 0) = \frac{1}{2}$ for any Δ and N.) Asymptotic performance is $\mathrm{PC}(\Delta = \infty, N \ge K, D) = 1 - \frac{1}{2}\binom{N-D}{K} / \binom{N}{K}$, again below 100%.

Appendix B

Parameters

This appendix lists parameter values used by model. Wherever possible, parameters were taken from the actual experimental settings (Tripathy & Barrett, 2004; Tripathy et al., 2007).

Bayesian models

In Experiment 1, set size took values $N = 1, 2, 3, 4, 6, 8$. All trajectories deviated, $D = N$. Directional uncertainty at set size 1 was chosen $\sigma_1 = 2.8°$ ($\sigma_{\mathrm{pre},1} = \sigma_{\mathrm{pre},1} = 2°$); this value is comparable to the single-trajectory threshold reported in Figure 2b of Tripathy and Barrett (2004) at the dot speed used, 32 deg/s. Positional uncertainty and the jitter in the initial motion directions are not relevant since the decision rule is to average over all trajectories.

In Experiment 2, set size took values $N = 1, 2, 3, 4, 5$. One trajectory deviated, $D = 1$. Directional and positional uncertainty were $\sigma_1 = 2.8°$ and $\sigma_{\mathrm{pos},1} = 21'$, respectively. Vertical distances between the midpoints (relevant for positional uncertainty) were taken to be $10'$, with $10'$ uniform jitter, as in the experiment. Initial motion directions were drawn from a uniform distribution on $[-32°, 32°]$. In the model, the mean initial motion direction is irrelevant.

In Experiment 3, $N = 1, 2, 3, 4, 6, 8$; $D = 1$; $\Delta = 19°, 38°, 76°$. For directional uncertainty at set size 1, we took $\sigma_1 = 11.3°$. This is larger than in Experiments 1 and 2 because dot speed is lower, 4 deg/s. The value is again comparable to, though somewhat higher than the single-trajectory threshold reported in Figure 2b of Tripathy and Barrett (2004) (observer DB). Positional uncertainty at set size 1 was chosen $\sigma_{\mathrm{pos},1} = 21'$. Vertical distances between the midpoints: $40'$ with $5'$ jitter. Initial motion directions were drawn from a uniform distribution on $[-80°, 80°]$.

In Experiment 4, parameters were $N = 6,8$; $D = 1, 2, 3, 5, 6, 8$; $\Delta = 19°, 38°, 76°$; $\sigma_1 = 11.3°$; and $\sigma_{\mathrm{pos},1} = 21'$. Vertical distances between the midpoints: $30'$ with $5'$ jitter. Initial motion directions were drawn from a uniform distribution on $[-80°, 80°]$. In Experiment 5, parameters were identical except for $N = 10$; $D = 1, 2$.

In Prediction 1, we used $D = 1,\ldots, 5$. Other parameters were as in Experiment 2. In Prediction 2, $N = 6$ and $\Delta = 38°$. The number of high-uncertainty trajectories took values $H = 0, 1, 2,\ldots, 6$. Other parameters: $\sigma_{1,\mathrm{low}} = 11.3°$ (directional uncertainty of low-uncertainty trajectory at $N = 1$); $\sigma_{1,\mathrm{high}} = 22.6°$; $\sigma_{\mathrm{pos},1,\mathrm{low}} = 21'$; and $\sigma_{\mathrm{pos},1,\mathrm{high}} = 42'$. The vertical distances between midpoints and the initial motion directions were chosen as in Experiment 3.

In the Bayesian model with capacity limit, parameters were identical to those above. The capacity limit, K, was chosen to be 3.

Limited-capacity model

This model has only one free parameter, K, which was chosen to be $K = 3$ or $K = 4$ in this paper. $K = 3$ often fits the data best (though still not very well), and $K = 4$ is believed to be the capacity limit in standard multiple-object tracking (Pylyshyn & Storm, 1988). The effect of varying K is explored in Figures 6 and 9b.

Slots-plus-averaging model

As in the limited-capacity model, we used $K = 3$. The effect of changing K is explored for the near-threshold experiments in Figure 7b. For the suprathreshold experiments, changing K does not improve the resemblance of the model predictions to the data in Figures 9, 10, 11, and 12. We adjusted the uncertainty parameter of the model, σ_K, for different K to give the model a fair chance, as explained below Equation 27. In all of these except Figure 7b, $\sigma_{K=1} = 11.3°$, as in the Bayesian model (keep in mind that $\sigma_{K=1} = \sigma_{N=1}$). Therefore, $\sigma_{K=3} = 11.3° \cdot \sqrt{3} = 19.6°$. For Experiments 1 and 2 (Figures 7 and 8) and Prediction 1 (Figure 13), $\sigma_{K=1} = 3°$, and $\sigma_{K=3} = 3° \cdot \sqrt{3} = 5.2°$. In Prediction 2, we used $N = 6$; $K = 3$; $\Delta = 38°$. The number of high-uncertainty items took values $H = 0, 1, 2,\ldots, 6$. Directional uncertainty was $\sigma_{K,\mathrm{low}} = 11.3° \cdot \sqrt{3} = 19.6°$ and $\sigma_{K,\mathrm{high}} = 22.6° \sqrt{3} = 39.1°$.

Averaging model

In the averaging model, the only free parameter, σ_1, was chosen $\sigma_1 = 2.8°$ in Experiments 1 and 2, and $\sigma_1 = 11.3°$ in Experiments 3–5.

Appendix C

A higher power?

The constrained Bayesian model assumes a square root dependency of uncertainty on set size, $\sigma \propto \sqrt{N}$. This relationship was derived from a neural argument, and it accounts well for the behavioral data. However, in a recent short-term memory experiment, a power law with a higher exponent was found, $\sigma \propto N^{\alpha}$ with $\alpha \approx 0.74$ (Bays & Husain, 2008). A higher exponent might also be consistent with earlier data (Wilken & Ma, 2004). A higher power could arise from several causes, such as (a) feature uncertainty might be confounded with positional uncertainty; (b) the total amount of spikes expended decreases with set size; and (c) the form of neural variability is different from Poisson-like, leading to a different relationship between neural gain and uncertainty. The value of the exponent, and its origins, deserve further study. No data about α are available for attentional tracking. Here, we only examine the consequences of a higher power on the constrained Bayesian model.

In Experiment 1, the assumption $\sigma \propto \sqrt{N}$ caused threshold to be independent of set size, since this increase in uncertainty exactly canceled out the benefit from averaging N independent observations. In this scenario, the number of trajectories a subject attends to does not affect performance. This changes when $\sigma \propto N^{\alpha}$, with $\alpha > 0.5$. Attending to all trajectories would lead to an increase in threshold with set size, namely $\Delta_{\mathrm{thr}} = \sigma/\sqrt{N} \propto N^{\alpha}/\sqrt{N} \propto N^{\alpha-1/2}$. Then it becomes beneficial to attend to only a single trajectory, independent of set size. This means that the prediction of the constrained Bayesian model for Experiment 1 (green line in Figure 8f) does not change. However, it now becomes important to verify whether a similar strategy works if D of N trajectories deviate ($D < N$).

To examine this question, we consider the example of $N = 5$ and $D = 3$ in the near-threshold paradigm. Attending to a subset is equivalent to imposing a capacity limit which may depend on N and D. We compute performance of an observer with a capacity limit $K \leq N$, who follows the constrained Bayesian model for the items selected on each trial. This is done in a manner analogous to the Bayesian model with capacity limit (see Theory and methods section), except that base performance is now from the Bayesian model with constraint. Analogous to Equation A14, we have

$$\mathrm{PC}(\Delta, N \geq K, D) = \\ = \binom{N}{K}^{-1} \sum_{M=\max(K+D-N,0)}^{\min(K,D)} \binom{D}{M} \binom{N-D}{K-M} \mathrm{PC}_{CB}(\Delta, K, M). \quad \text{(C1)}$$

The results are shown in Figure C1. When the power is higher than ½ and possibly even when it is equal to ½, there could be a small benefit in attending to a subset of the items. In Figure C1a, this follows from the fact that performance for $K < 5$ exceeds performance for $K = 5$ (attending to all items) over a range of Δ. However, this slight improvement comes at the cost of deteriorated performance in a different range of Δ. We have also ignored the fact that unattended items may get confused with attended ones because of positional uncertainty. Overall, we cannot say that attending to a subset is beneficial. Whether it is beneficial or not also depends on N and D, which might make this strategy impractical.

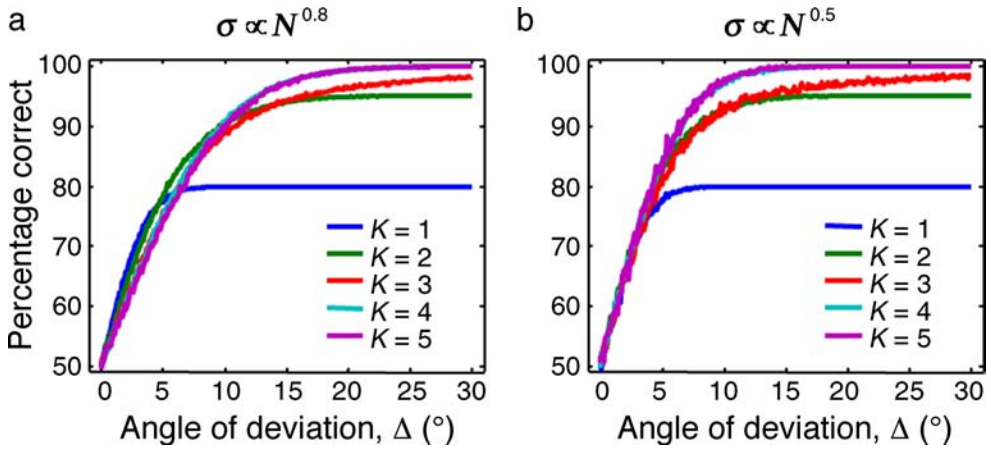

Figure C1. Effect of a higher power in the relationship between uncertainty and set size. We examine whether, in a near-threshold paradigm, a constrained Bayesian observer would benefit from only attending to K items. (a) Percentage correct as a function of deviation angle, for different values of the number of attended trajectories, K. We consider the case $N = 5$, $D = 3$, and assume $\sigma \propto N^{0.8}$ instead of $\sigma \propto N^{0.5}$. The case $K = 5$ corresponds to the constrained Bayesian model (attending to all trajectories). Parameters were as in Experiment 2 and a uniform prior over Δ was assumed. Attending to a subset slightly improves performance in some range of Δ, but at the cost of a deterioration at other values. This is qualitatively similar for other values of N and D that we tried. (b) As in panel a, but with $\sigma \propto N^{0.5}$. Attending to a subset ($K < 5$) unambiguously affects performance negatively.

Using the above parameters, attending to a subset always hurts when $\alpha = 0.5$ (Figure C1b). We conclude that even if uncertainty grows faster than the square root of set size, this does, within limits, not greatly affect the optimal strategy.

Acknowledgments

We would like to thank Patrick Wilken, Paul Bays, and two anonymous reviewers for helpful comments. This research was supported by seed funding from Baylor College of Medicine. Modeling code will be made available on http://neuro.bcm.edu/malab/code.

Commercial relationships: none.
Corresponding author: Wei Ji Ma.
Email: wjma@bcm.edu.
Address: 1 Baylor Plaza, Houston, TX 77030, USA.

References

Aldo Faisal, A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews. Neuroscience, 9,* 292–303. [PubMed] [Article]

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, 1–10, http://journalofvision.org/7/13/14/, doi:10.1167/7.13.14. [PubMed] [Article]

Baldassi, S., & Verghese, P. (2002). Comparing integration rules in visual search. *Journal of Vision, 2*(8):3, 559–570, http://journalofvision.org/2/8/3/, doi:10.1167/2.8.3. [PubMed] [Article]

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

Bays, P. M., & Husain, M. (2009). Response to comment on "Dynamic shifts of limited working memory resources in human vision." *Science, 323,* 877.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T. D., Churchland, A. K., Roitman, J. D., et al. (2008). Bayesian decision-making with probabilistic population codes. *Neuron, 60,* 1142–1145.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences, 9,* 349–354.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioural Brain Science, 24,* 87–114. [PubMed]

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,* 425–451. [PubMed]

Graham, N., Kramer, P., & Yager, D. (1987). Signal detection models for multidimensional stimuli: Probability distributions and combination rules. *Journal of Mathematical Psychology, 31,* 366–409.

Green, C. S., & Bavelier, D. (2006). Enumeration versus multiple object tracking: The case of action video game players. *Cognition, 101,* 217–245. [PubMed]

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Los Altos, CA: John Wiley & Sons.

Hospedales, T., & Vijayakumar, S. (2009). Multisensory oddity detection as Bayesian inference. *PLoS ONE, 4,* e4205. [PubMed] [Article]

Hulleman, J. (2005). The mathematics of multiple object tracking: From proportions correct to number of objects tracked. *Vision Research, 45,* 2298–2309. [PubMed]

Huys, Q., Zemel, R. S., Natarajan, R., & Dayan, P. (2007). Fast population coding. *Neural Computation, 19,* 404–441. [PubMed]

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55,* 271–304. [PubMed]

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27,* 712–719. [PubMed]

Knill, D. C., & Richards, W. (Eds.) (1996). *Perception as Bayesian Inference*. New York: Cambridge University Press.

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2,* e943. [PubMed] [Article]

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

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9,* 1432–1438. [PubMed]

Ma, W. J., Navalpakkam, V., Beck, J. M., & Pouget, A. (2008). *Bayesian theory of visual search*. Paper presented at the Society for Neuroscience.

Narasimhan, S., Tripathy, S. P., & Barrett, B. T. (2009). Loss of positional information when tracking multiple

moving dots: The role of visual memory. *Vision Research, 49,* 10–27. [PubMed]

Nolte, L. W., & Jaarsma, D. (1966). More on the detection of one of M orthogonal signals. *Journal of the Acoustical Society of America, 41,* 497–505.

Oksama, L., & Hyona, J. (2008). Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive Psychology, 56,* 237–283. [PubMed]

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 332–350. [PubMed]

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

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44,* 369–378. [PubMed]

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3,* 179–197. [PubMed]

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7,* 17–42.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron, 61,* 168–185. [PubMed]

Rosenholtz, R. (2001). Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal detection theory models of search. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 985–999. [PubMed]

Sato, Y., Toyoizumi, T., & Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli. *Neural Computation, 19,* 3335–3355. [PubMed]

Seung, H., & Sompolinsky, H. (1993). Simple model for reading neuronal population codes. *Proceedings of National Academy of Sciences of the United States of America, 90,* 10749–10753. [PubMed] [Article]

Shaw, M. L. (1980). Identifying attentional and decision-making components in information processing. In R. S. Nickerson (Ed.), *Attention and performance* (vol. VIII, pp. 277–296). Hillsdale, NJ: Erlbaum.

Tripathy, S. P., & Barrett, B. T. (2004). Severe loss of positional information when detecting deviations in multiple trajectories. *Journal of Vision, 4*(12):4, 1020–1043, http://journalofvision.org/4/12/4/, doi:10.1167/4.12.4. [PubMed] [Article]

Tripathy, S. P., Narasimhan, S., & Barrett, B. T. (2007). On the effective number of tracked trajectories in normal human vision. *Journal of Vision, 7*(6):2, 1–18, http://journalofvision.org/7/6/2/, doi:10.1167/7.6.2. [PubMed] [Article]

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron, 31,* 523–535. [PubMed]

Vincent, B. T., Baddeley, R. J., Troscianko, T., & Gilchrist, I. D. (2009). Optimal feature integration in visual search [Abstract]. *Journal of Vision, 9*(5):15, 1–11, http://journalofvision.org/9/5/15/, doi:10.1167/9.5.15.

Vul, E., Frank, M., Alvarez, G. A., & Tenenbaum, J. B. (2009). *Statistical decision theory and the allocation of cognitive resources in multiple object tracking*. Paper presented at the Computational and Systems Neuroscience meeting.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12):11, 1120–1135, http://journalofvision.org/4/12/11/, doi:10.1167/4.12.11. [PubMed] [Article]

Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology, 24,* 295–340. [PubMed]

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika, 2,* 121–134.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233–235. [PubMed] [Article]