

# Supporting Information

Supporting Information Corrected December 17, 2012

Van den Berg et al. 10.1073/pnas.1117465109

## SI Text

Convention: For convenience, we have mapped both orientation and color space to  $[0, 2\pi)$  in all equations.

**Fisher Information as Resource.** In the item-limit (IL) model, an item is encoded either perfectly or not at all. All other models we tested contain a notion of noise. Therefore, we have to specify the relationship between “amount of resource” and the level of noise.

Intuitively, resource is something that is allocated to an item to improve the quality of its encoding. The traditional notion of resource is that of a very large pool of available observations made of the stimulus, also called samples (1, 2). Each observation is corrupted by independent, zero-mean Gaussian noise with the same SD, and the observer’s eventual measurement,  $x$ , is the mean of these observations. Then the variance of the measurement decreases inverse proportionally to the number of observations, and precision increases proportionally.

In this paper, we instead identify resource with Fisher information, denoted  $J$ . Fisher information determines the best possible performance of any estimator, through the Cramér-Rao bound (3). Fisher information is defined in terms of the noise distribution, which is the distribution of the observations conditioned on the stimulus  $s$ ,

$$J(s) = - \left\langle \frac{\partial^2}{\partial s^2} \log p(\text{observations} | s) \right\rangle, \quad [\text{S1}]$$

where  $\langle \rangle$  denotes an expected value over  $p(\text{observations}|s)$ .

If  $x$  follows a Gaussian distribution with mean  $s$  and SD  $\sigma$ , it is easily verified from the definition, Eq. S1, that Fisher information is equal to the inverse variance,  $J = \frac{1}{\sigma^2}$ , recovering the earlier relationship. This equation is an improvement over the “number of observations” argument because  $J$  is defined on a continuum and readily neurally interpretable. At the neural level, Fisher information is proportional to the gain of a population when neural variability is Poisson-like (4).

A slight complication arises from the fact that the stimulus spaces we use (orientation and color) are circular, so that the Gaussian distribution is no longer appropriate. Instead, we assume that the measurement follows a Von Mises distribution:

$$p(x | s) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(x-s)} \equiv \text{VM}(x; s, \kappa).$$

$I_0$  is the modified Bessel function of the first kind of order zero (5) and serves as a normalization. The concentration parameter  $\kappa$  controls the width of the noise distribution. When it is large, the Von Mises distribution resembles a Gaussian distribution with variance  $1/\kappa$ . When  $\kappa = 0$ ,  $p(x|s)$  is the uniform distribution. It is important that the Gaussian distribution is a special case of the Von Mises distribution, because the maximum-likelihood estimate has an asymptotically Gaussian distribution.

We calculate Fisher information from its definition, Eq. S1,

$$J = \langle \kappa \cos(x-s) \rangle = \frac{\kappa}{2\pi I_0(\kappa)} \int \cos(x-s) e^{\kappa \cos(x-s)} d\hat{s} = \kappa \frac{I_1(\kappa)}{I_0(\kappa)}, \quad [\text{S2}]$$

where  $I_1(\kappa)$  is the modified Bessel function of the first kind of order one (5). This equation relates Fisher information in a one-to-one fashion to the concentration parameter of the Von Mises

distribution. We use it in all models except for the IL model. One can think of Fisher information as precision, by analogy to the Gaussian case. We write the inverse relationship of Eq. S2 as

$$\kappa = \Phi(J). \quad [\text{S3}]$$

The inverse function  $\Phi$  is not analytical but can be computed numerically.

**Equal-precision model.** In the equal-precision (EP) model, we assume

$$J = \frac{J_1}{N^\alpha}, \quad [\text{S4}]$$

where  $J_1$  is the Fisher information at set size 1. Using Eq. S3, the concentration parameter at set size  $N$  is

$$\kappa(N) = \Phi\left(\frac{J_1}{N^\alpha}\right). \quad [\text{S5}]$$

**Slots-plus-averaging model.** The slots-plus-averaging (SA) model (6) is similar to the IL model, with the modification that when  $N < K$ , multiple chunks of resource can be assigned to a single item. This modification gives it some characteristics of the EP model. Specifically, the assumption is that the amount of resource is proportional to the number of assigned chunks,  $S$ . Zhang and Luck (6) do not mention the exact relationship between amount of resource and the concentration parameter of the Von Mises distribution, but we assume that they used the correct relationship, Eq. S2. Then, the concentration parameter as a function of  $S$  is

$$\kappa = \Phi(SJ_1), \quad [\text{S6}]$$

where  $J_1$  is now the Fisher information corresponding to having one chunk ( $S = 1$ ). When  $N > K$ , an item receives 0 chunks or 1 chunk, with probabilities  $K/N$  and  $1 - K/N$ , respectively. This allocation is the same as in the IL model. When  $N \leq K$ , all items receive at least one chunk and it is assumed that the chunks are distributed as equally as possible over all items. For example, if  $K = 4$  and  $N = 3$ , two items get assigned one chunk each and one item gets two chunks. From this, it follows that the number of chunks an item receives,  $S$ , is equal to

$$S = \begin{cases} \left\lfloor \frac{K}{N} \right\rfloor & \text{with probability } 1 - \frac{K \bmod N}{N}; \\ \left\lfloor \frac{K}{N} \right\rfloor + 1 & \text{with probability } \frac{K \bmod N}{N}, \end{cases}$$

where  $\lfloor x \rfloor$  denotes the largest integer smaller than  $x$  (floor function). Using Eq. S6, these two values of  $S$  correspond to two values of the concentration parameter  $\kappa$ , which we denote by  $\kappa_{\text{low}}$  and  $\kappa_{\text{high}}$ , respectively:

$$\begin{aligned} \kappa_{\text{low}}(N) &= \Phi\left(\left\lfloor \frac{K}{N} \right\rfloor J_1\right), \\ \kappa_{\text{high}}(N) &= \Phi\left(\left(\left\lfloor \frac{K}{N} \right\rfloor + 1\right) J_1\right). \end{aligned} \quad [\text{S7}]$$

In the example above, two items would be memorized with concentration parameter  $\kappa_{\text{low}}$  and the third one with  $\kappa_{\text{high}}$ .

**Variable-precision model.** In the variable-precision model, precision is variable across items and trials. We assume that each precision is drawn independently from a gamma distribution with mean precision  $\bar{J}$  and scale parameter  $\tau$ ,

$$p(g|\bar{g};\tau) = \text{Gamma}(J;\bar{J},\tau). \quad [\text{S8}]$$

The variance of  $J$  is equal to  $\bar{J}\tau$ . The gamma distribution is a common distribution on the positive real line. We assume that mean precision depends on set size in the following way:

$$\bar{J} = \frac{\bar{J}_1}{N^\alpha}. \quad [\text{S9}]$$

**Model Predictions for Delayed Estimation. Item-limit model.** The item-limit model assumes that the memory of a stored item is perfect; thus  $\hat{s} = s$ . However, we allow for the possibility that response noise (e.g., motor noise) corrupts the subject's response. Therefore, we assume that the response, denoted  $r$ , follows a Von Mises distribution centered on the true stimulus with concentration parameter  $\kappa_r$ . For  $N \leq K$ , we then have  $p(r|s) = \text{VM}(r;s,\kappa_r)$ . If  $N > K$ , there is a probability of  $K/N$  that the probed item was memorized and a probability of  $1 - K/N$  that it was not memorized, in which case the subject will make a random guess. Hence, the response distribution is a mixture of a Von Mises distribution and a uniform (guessing) distribution:

$$p(r|s) = \frac{K}{N} \text{VM}(r;s,\kappa_r) + \left(1 - \frac{K}{N}\right) \frac{1}{2\pi}. \quad [\text{S10}]$$

This model has two free parameters:  $K$  and  $\kappa_r$ .

**Equal-precision model.** In the presence of encoding noise, the best estimate of the stimulus is equal to the measurement,  $\hat{s} = x$ . The estimate distribution predicted by the EP model is then

$$p(\hat{s}|s) = \text{VM}(\hat{s};s,\kappa(N))$$

with  $\kappa(N) = \Phi\left(\frac{J_1}{\sqrt{\alpha}}\right)$  (Eq. S5). Including response noise with concentration parameter  $\kappa_r$ , the response distribution is

$$p(r|s) = \int_0^{2\pi} \frac{1}{2\pi I_0(\kappa(N))} e^{\kappa(N)\cos(\hat{s}-s)} \frac{1}{2\pi I_0(\kappa_r)} e^{\kappa_r \cos(r-\hat{s})} d\hat{s}.$$

A lengthy but straightforward calculation gives

$$p(r|s;N) = \frac{I_0\left(\sqrt{\kappa(N)^2 + \kappa_r^2 + 2\kappa(N)\kappa_r \cos(r-s)}\right)}{2\pi I_0(\kappa_r)I_0(\kappa(N))}. \quad [\text{S11}]$$

The EP model for the delayed-estimation task has three free parameters:  $J_1$ ,  $\alpha$ , and  $\kappa_r$ .

**Slots-plus-averaging model.** The estimate distribution is a mixture of a Von Mises and a uniform distribution,

$$p(\hat{s}|s) = \frac{K}{N} \text{VM}(\hat{s};s,\kappa_1) + \left(1 - \frac{K}{N}\right) \frac{1}{2\pi},$$

with  $\kappa_1 = \Phi(J_1)$ . With response noise, the response distribution becomes

$$p(r|s) = \frac{K}{N} \frac{I_0\left(\sqrt{\kappa_1^2 + \kappa_r^2 + 2\kappa_1\kappa_r \cos(r-s)}\right)}{2\pi I_0(\kappa_1)I_0(\kappa_r)} + \left(1 - \frac{K}{N}\right) \frac{1}{2\pi}. \quad [\text{S12}]$$

The estimate distribution for  $N \leq K$  is a mixture of two Von Mises distributions:

$$p(\hat{s}|s) = \frac{K \bmod N}{N} \frac{1}{2\pi I_0(\kappa_{\text{high}}(N))} e^{\kappa_{\text{high}}(N)\cos(\hat{s}-s)} + \left(1 - \frac{K \bmod N}{N}\right) \frac{1}{2\pi I_0(\kappa_{\text{low}}(N))} e^{\kappa_{\text{low}}(N)\cos(\hat{s}-s)}.$$

With response noise, the response distribution for  $N \leq K$  is

$$p(r|s) = \frac{K \bmod N}{N} \frac{I_0(\kappa_{c,\text{high}}(N))}{2\pi I_0(\kappa_{\text{high}}(N))I_0(\kappa_r)} + \left(1 - \frac{K \bmod N}{N}\right) \frac{I_0(\kappa_{c,\text{low}}(N))}{2\pi I_0(\kappa_{\text{low}}(N))I_0(\kappa_r)}, \quad [\text{S13}]$$

with

$$\begin{aligned} \kappa_{c,\text{high}}(N) &= \sqrt{\kappa_{\text{high}}(N)^2 + \kappa_r^2 + 2\kappa_{\text{high}}(N)\kappa_r \cos(\hat{s}-s)}, \\ \kappa_{c,\text{low}}(N) &= \sqrt{\kappa_{\text{low}}(N)^2 + \kappa_r^2 + 2\kappa_{\text{low}}(N)\kappa_r \cos(\hat{s}-s)}. \end{aligned} \quad [\text{S14}]$$

The SA model for the delayed-estimation task has three free parameters:  $K$ ,  $J_1$ , and  $\kappa_r$ .

**Variable-precision model.** The estimate distribution corresponding to a fixed precision  $J$  is  $p(\hat{s}|s;J) = \text{VM}(\hat{s};s,\Phi(J))$ . When precision is variable, the estimate distribution is a mixture of the estimate distributions associated with individual values of precision, with mixture proportions equal to the frequencies of those values,  $p(J|\bar{J};\tau) = \text{Gamma}(J;\bar{J},\tau)$ , with  $\bar{J}$  given by Eq. S9. Therefore,

$$\begin{aligned} p(\hat{s}|s;\bar{J},\tau) &= \int p(\hat{s}|s;J)p(J|\bar{J};\tau)dJ \\ &= \int \text{VM}(\hat{s};s,\Phi(J))\text{Gamma}(J;\bar{J},\tau)dJ, \end{aligned} \quad [\text{S15}]$$

This distribution is a mixture of an infinite set of Von Mises distributions. We approximate the mixture by sampling 500 values of  $J$  from the gamma distribution and averaging the Von Mises distributions corresponding to these samples. We examined the effect of the number of samples on the model predictions and found that 500 is a sufficiently large number to give robust results (Fig. S11A). Response noise is added as above, by convolving  $p(\hat{s}|s;\bar{J},\tau)$  with a Von Mises distribution with concentration parameter  $\kappa_r$ . Thus, the variable-precision (VP) model for the delayed-estimation task has four free parameters:  $\bar{J}_1$ ,  $\alpha$ ,  $\tau$ , and  $\kappa_r$ .

**Model Predictions for Change Localization.** In change localization, the variables in the task are the location of the change,  $L$ , the magnitude of the change,  $\Delta$ , the vector of stimuli in the first display,  $\theta = (\theta_1, \dots, \theta_N)$ , and the vector of stimuli in the first display,  $\varphi = (\varphi_1, \dots, \varphi_N)$ . Each  $L$  has a probability of  $1/N$ . The probability density of  $\Delta$  is flat at  $\frac{1}{2\pi}$ , and the one over  $\theta$  is flat at

$\left(\frac{1}{2\pi}\right)^N$ . The relation between  $\theta$  and  $\varphi$  is  $\varphi = \theta + \Delta \mathbf{1}_L$ , where  $\mathbf{1}_L$  is the vector of zeros with a 1 at the  $L$ th entry.

**Item-limit model.** According to the item-limit model, the probability of being correct is equal to  $1 - \varepsilon$  when  $N \leq K$  and to  $\frac{K}{N} + \left(1 - \frac{K}{N}\right)\frac{1}{N-K} = \frac{K+1}{N}$  when  $N > K$ . These probabilities are independent of  $\theta$ ,  $\varphi$ , and  $\Delta$ . We introduced  $\varepsilon$  because without it (i.e.,  $\varepsilon = 0$ ), the data would have probability zero under the model. The IL model for the change localization task has two free parameters:  $K$  and  $\varepsilon$ .

**Bayesian decision rule.** All models except for the IL model have noise in the measurements and probabilistic inference is needed to estimate the location of the change. The  $i$ th measurement in the first display,  $x_i$ , is drawn independently from a Von Mises distribution with mean  $\theta_i$  and concentration parameter  $\kappa_i$ . The  $i$ th measurement in the second display, denoted  $y_i$ , is drawn from a Von Mises distribution with mean  $\varphi_i$  and concentration parameter  $\kappa_i$  (it is possible to allow  $\kappa_i$  to be different between the two displays but we chose not to do so). The relations between the variables are shown

in the graphical model in Fig. S1B. To model how the observer decides on the basis of the measurements  $\mathbf{x} = (x_1, \dots, x_N)$  and  $\mathbf{y} = (y_1, \dots, y_N)$ , we use a Bayesian-observer model. The Bayesian observer computes a probability distribution over the location of the change,  $p(L | \mathbf{x}, \mathbf{y})$ , and then reports the location with the highest probability. The posterior distribution over  $L$  is proportional to the joint distribution,  $p(\mathbf{x}, \mathbf{y}, L)$ , which in turn is evaluated as an integral over the remaining variables, namely  $\Delta$ ,  $\theta$ , and  $\varphi$ ,

$$p(\mathbf{x}, \mathbf{y}, L) = \iiint p(\mathbf{x}, \mathbf{y}, \theta, \varphi, \Delta, L) d\Delta d\theta d\varphi \\ = \iiint p(L) p(\Delta) p(\theta) p(\varphi | L, \theta) p(\mathbf{x} | \theta) p(\mathbf{y} | \varphi) d\Delta d\theta d\varphi,$$

where in going from the first to the second line we have used the structure of the generative model in Fig. S1B. Substituting distributions and evaluating the integral over  $\varphi$  gives

$$p(\mathbf{x}, \mathbf{y}, L) = \frac{1}{N} \left( \frac{1}{2\pi} \right)^{N+1} \int \prod_{i=1}^N \left( \int p(x_i | \theta_i) p(y_i | \varphi_i = \theta_i + \Delta \delta_{L,i}) \right) d\Delta, \quad [\text{S16}]$$

where  $\delta_{L,i} = 1$  when  $L = i$  and 0 otherwise. Because we are interested only in the dependence on  $L$ , we can freely divide by the  $L$ -independent product  $\prod_{i=1}^N (\int p(x_i | \theta_i) p(y_i | \varphi_i = \theta_i))$ , leaving only integrals pertaining to the  $L$ th location:

$$p(\mathbf{x}, \mathbf{y}, L) \propto \frac{\iint p(x_L | \theta_L) p(y_L | \varphi_L = \theta_L + \Delta) d\theta_L d\Delta}{\int p(x_L | \theta_L) p(y_L | \varphi_L = \theta_L)}. \quad [\text{S17}]$$

$$\begin{cases} (L \text{ encoded}) & \frac{1}{2\pi} \frac{\iint p(x_L | \theta_L) p(y_L | \varphi_L = \theta_L + \Delta) d\theta_L d\Delta}{\int p(x_L | \theta_L) p(y_L | \varphi_L = \theta_L)} \\ (L \text{ not encoded}) & 1. \end{cases} = \frac{1}{2\pi \int p(x_L | \theta_L) p(y_L | \varphi_L = \theta_L)}$$

This probability evaluates to

$$p(\mathbf{x}, \mathbf{y}, L) \propto \frac{I_0(\kappa_L)^2}{I_0(\kappa_L \sqrt{2 + 2 \cos(x_L - y_L)})}.$$

Thus, the maximum a posteriori (MAP) estimate of the location of the change is

$$\hat{L} = \underset{L}{\operatorname{argmax}} \frac{I_0(\kappa_L)^2}{I_0(\kappa_L \sqrt{2 + 2 \cos(x_L - y_L)})}. \quad [\text{S18}]$$

The distribution of the MAP estimate for given  $L$ ,  $\Delta$ ,  $N$ , and  $\kappa$ , denoted  $p(\hat{L} | L, \Delta, \kappa; N)$ , depends on the model but is computed through Monte Carlo simulation for all models (using 10,000 samples of  $\mathbf{x}$  and  $\mathbf{y}$ ). Note that the estimate distribution is characterized by a single number, namely probability correct.

**Equal-precision model.** In the equal-precision model, we take  $\kappa_i = \Phi(\frac{J_i}{\sqrt{N}})$  for all  $i$ . We assume equality between first and second displays, because the concentration parameters in both displays can essentially not be fitted independently (compare with the sum of two normally distributed random variables: the variances sum and cannot be estimated individually). The EP model for the change localization task has two free parameters:  $J_1$  and  $\alpha$ .

**Slots-plus-averaging model.** In the slots-plus-averaging model,  $\kappa_i$  is given by Eq. S7. When  $N \leq K$ , Eq. S18 applies. When  $N > K$ , inference is based on only  $K$  of  $N$  measurements in each display,  $\mathbf{x} = (x_1, \dots, x_K)$  and  $\mathbf{y} = (y_1, \dots, y_K)$ , yet the change could have occurred at any location. We first evaluate the joint probability of  $\mathbf{x}$ ,  $\mathbf{y}$  and that the change occurred at a location  $L$

that is among the encoded ones. In analogy to Eq. S16, this probability is

$$(L \text{ encoded}) p(\mathbf{x}, \mathbf{y}, L) = \frac{1}{N} \left( \frac{1}{2\pi} \right)^{K+1} \\ \times \int \prod_{i=1}^K \left( \int p(x_i | \theta_i) p(y_i | \varphi_i = \theta_i + \Delta \delta_{L,i}) \right) d\Delta. \quad [\text{S19}]$$

Now we evaluate the joint probability of  $\mathbf{x}$ ,  $\mathbf{y}$  and that the change occurred at a location  $L$  that is not among the encoded ones. This probability is equal to

$$(L \text{ not encoded}) p(\mathbf{x}, \mathbf{y}, L) = \iint p(\mathbf{x}, \mathbf{y}, \theta, \varphi, L) d\theta d\varphi \\ = \iint p(L) p(\theta) p(\varphi | L, \theta) p(\mathbf{x} | \theta) p(\mathbf{y} | \varphi) d\theta d\varphi \\ = \frac{1}{N} \left( \frac{1}{2\pi} \right)^K \prod_{i=1}^K \left( \int p(x_i | \theta_i) p(y_i | \varphi_i = \theta_i) \right). \quad [\text{S20}]$$

As one would expect, this probability does not depend on  $L$ . Because we are interested only in the location  $L$  for which  $p(\mathbf{x}, \mathbf{y}, L)$  is largest (i.e., the argmax), we divide both Eqs. S19 and S20 by Eq. S20. Then, in analogy to Eq. S17, we have to take the argmax of

Evaluating the integral, the estimate of location is

$$\hat{L} = \underset{L}{\operatorname{argmax}} \frac{I_0(\kappa_L)^2}{I_0(\kappa_L \sqrt{2 + 2 \cos(x_L - y_L)})} \quad [\text{S21}]$$

when the value of this maximum exceeds 1, and we randomly guess from among the nonencoded items when it does not. The SA model for the change localization task has two free parameters:  $J_1$  and  $K$ .

**Variable-precision model.** In the variable-precision model, every  $J_i$  is independently drawn from a gamma distribution with mean  $\bar{J}$  (given by Eq. S9) and scale parameter  $\tau$ . Then, the estimate distribution is

$$p(\hat{L} | L, \Delta; \bar{J}, \tau) = \int \dots \int p(\hat{L} | L, \Delta; \mathbf{J}) \left( \prod_{i=1}^N \text{Gamma}(J_i; \bar{J}, \tau) \right) dJ_1 \dots dJ_N.$$

This distribution is obtained through Monte Carlo simulation of  $\mathbf{J}$ , using 10,000 samples. The VP model for the change localization task has three free parameters:  $\bar{J}_1$ ,  $\alpha$ , and  $\tau$ .

**Experimental Details. Experiment 1: Delayed estimation with color stimuli.** Observers briefly viewed and memorized a set of colors and reported the color of a randomly chosen target disk (Fig. 2A).

**Stimuli.** Stimuli were displayed on a 21-inch cathode ray tube monitor at a viewing distance of  $\sim 60$  cm. The stimulus array consisted of  $N$  colored discs ( $N = 1, \dots, 8$ ) with a diameter of  $2^\circ$  of visual angle, with their centers lying on an imaginary circle of radius  $4.5^\circ$ . The locations of the discs were randomly selected from

eight fixed positions equally spaced along the circle, including the positions corresponding to the cardinal directions with respect to fixation. The colors of the discs were drawn independently from 180 color values uniformly distributed along a circle of radius 60 in CIE 1976 ( $L^*$ ,  $a^*$ ,  $b^*$ ) color space. This circle had constant luminance ( $L^* = 70$ ) and was centered at the point ( $a^* = 10$ ,  $b^* = 10$ ). The stimuli were presented on a midlevel gray background (128 on an 8-bit grayscale) of luminosity 8.1 cd/m<sup>2</sup>.

**Procedure.** A trial sequence consisted of the presentation of a fixation cross, the stimulus array, and a delay period during which only the fixation cross was visible (Fig. 2A). Set size was chosen pseudorandomly and the colors of the items were drawn independently from a uniform distribution. The response screen consisted of white circles marking the circumferences of the discs in the stimulus array, with a thicker circle marking one randomly chosen disk. The subject's task was to report the color of the disk that had been present in the stimulus array at the marked location, by using the left and right arrow keys to scroll through all possible colors. After the first key press, a random color appeared within the thicker circle. After subsequent key presses, this color changed by moving either clockwise or counterclockwise through the color wheel. The association between left/right key presses and the direction in which the color wheel was traversed was randomized on each trial. To submit a response, the subject pressed the space bar.

The experiment consisted of three sessions on different days. Each session consisted of two blocks in which subjects responded using a color wheel condition (*SI Text*) and two blocks in which they responded by scrolling. Color wheel and scrolling blocks were interleaved in ABBA order, with A and B randomized for each subject. After every 24 trials, feedback was given in the form of a total score. The score per trial was 3 when the estimate was within five color values of the true value and was  $[3 - E/15]$  (floor function), with  $E$  the error, otherwise. The first two blocks were each preceded by 8 practice trials. If the total score across these trials was less than 3, subjects were asked to repeat the practice. The actual block consisted of 144 trials. In total, each subject completed  $3 \cdot 2 \cdot 2 \cdot 144 = 1,728$  testing trials.

Fourteen subjects participated in this experiment (age range, 18–50 y; 12 naive). Data of one subject were excluded, because her estimated value of  $w$  at set size 1 was extremely low ( $w = 0.72$ , compared with  $w > 0.97$  for every other subject).

**Experiment 2: Delayed estimation with orientation stimuli.** Experiment 2 differed from experiment 1 in the following ways. Set size was 2, 4, 6, or 8. All stimuli were displayed on a 19-inch liquid crystal display monitor at a viewing distance of  $\sim 50$  cm. The stimulus array was composed of Gabors with a Gaussian envelope of  $0.5^\circ$  and a wavelength randomly chosen from a uniform distribution on  $[0.3, 0.8]$  cycles per degree (Fig. 2B). The Gabor centers lay on an imaginary circle of radius  $8.2^\circ$ . Presentation time was 110 ms. A circle appeared around the location of the item whose orientation had to be reported. When subjects moved the mouse, a Gabor appeared inside that circle. They had to rotate it using the mouse to match the orientation of the Gabor that had been in that location. They pressed the space bar to submit their response. Feedback consisted of an integer score between  $-3$  and  $+3$  on every trial. When  $E$  is the error, the score was computed as  $3 - E/15$ , rounded to the nearest integer. Six subjects participated (four naive). Each subject completed four sessions of 640 trials each, for a total of 2,560 trials.

**Experiment 3: Change localization with color stimuli.** Observers briefly viewed two screens containing a set of colors, separated in time by a blank screen, and reported the location of the color change (Fig. 2C).

**Stimuli.** All stimuli were displayed on a 19-inch LCD monitor at a viewing distance of  $\sim 60$  cm. The first stimulus array was composed of  $N$  colored discs ( $n = 2, 4, 6, \text{ or } 8$ ) with a diameter of  $0.62^\circ$  of visual angle with their centers lying on an imaginary circle of radius  $7^\circ$  (Fig. 2C). The locations of the discs were

randomly selected from eight fixed positions equally spaced along the circle, including the positions corresponding to the cardinal directions with respect to fixation. The colors were drawn independently from 180 color values uniformly distributed along a circle of radius 60 in CIE 1976 ( $L^*$ ,  $a^*$ ,  $b^*$ ) color space. This circle had constant luminance ( $L^* = 58$ ) and was centered at the point ( $a^* = 12$ ,  $b^* = 13$ ). The stimuli were presented on a midlevel gray background (128 on an 8-bit grayscale) of luminosity 33.1 cd/m<sup>2</sup>.

**Procedure.** The trial sequence consisted of the presentation of a fixation cross (1,000 ms), the stimulus array (110 ms), a delay period during which only the fixation cross was visible (1,000 ms), another stimulus array in which one of the stimuli changed color (110 ms), and a response screen that consisted of empty circles at the locations where the stimuli were shown. In the first stimulus array, set size was chosen randomly and the color of each item was chosen randomly as described above. In the second stimulus array,  $N - 1$  stimuli were identical to those in the first display, and the color of the remaining stimulus was chosen randomly from the same uniform distribution. The location of the changing stimulus was chosen randomly. The subject's task was to click on the location of the stimulus that had changed color. The experiment consisted of four sessions on different days. Each session consisted of four blocks with 120 trials each. Hence, each subject completed  $4 \cdot 4 \cdot 120 = 1,920$  trials in total. Seven subjects participated in this experiment (age range, 21–32 y; five naive).

**Experiment 4: Change localization with orientation stimuli.** Experiment 4 differed from experiment 3 in the following ways. Stimuli were white, oriented ellipses with minor and major axes of  $0.41^\circ$  and  $0.94^\circ$  of visual angle, respectively, and a luminance of  $95.7 \text{ cd/m}^2$  (Fig. 2D). Eleven subjects participated (age range, 23–29 y; 9 naive).

**Details of Data Analysis in Experiments 1 and 2.** In experiments 1 and 2, to remove bias, we circularly subtracted, for each subject separately, the circular mean across all trials from the subject's reports before any analyses.

**Computing the summary statistics  $w$  and CSD.** In delayed estimation, the raw data consist of the distributions of the estimation error,  $\Delta s$ , at each set size (Fig. S2). The summary statistics  $w$  and CSD (Fig. 4) were obtained by fitting a mixture of a Von Mises distribution and a uniform distribution:

$$p_{\text{fit}}(\Delta s; w, \kappa_{\text{fit}}) = \frac{w}{2\pi I_0(\kappa_{\text{fit}})} e^{\kappa_{\text{fit}} \cos \Delta s} + \frac{1-w}{2\pi}. \quad [\text{S22}]$$

The circular SD is defined as  $\text{CSD} = \sqrt{1 - \frac{I_1(\kappa_{\text{fit}})}{I_0(\kappa_{\text{fit}})}}$  (7). We fitted

this mixture separately for each subject and each set size, both to the data and to the error distributions predicted by each of the models. Fitting was done through maximum-likelihood estimation, which means choosing the values of the parameters of Eq. S22,  $w$  and  $\kappa_{\text{fit}}$ , that maximize the probability of the data given the parameters. This is equivalent to maximizing the log-likelihood function

$$\begin{aligned} \log L(w, \kappa_{\text{fit}}) &= \log p(\text{data} | w, \kappa_{\text{fit}}) = \log \prod_{i=1}^{N_{\text{trials}}} p_{\text{fit}}(\Delta s_i | w, \kappa_{\text{fit}}) \\ &= \sum_{i=1}^{N_{\text{trials}}} \log p_{\text{fit}}(\Delta s_i | w, \kappa_{\text{fit}}), \end{aligned}$$

where  $N_{\text{trials}}$  is the number of trials. We use `fminsearch` in Matlab to perform the maximization.

**Nontarget reports in experiment 1.** It has been argued that in the color wheel condition, guessing is confounded with nontarget reports, in the sense that the fitted uniform component includes a substantial amount of reports of nontarget colors (8). To test for this, we

fitted two modified mixtures to the data. The first is the one that assigns a probability to reporting the color of a nontarget disk (8),

$$p_{\text{fit}}(\hat{s} | s; w, \kappa_{\text{fit}}) = \frac{w_{\text{guess}}}{2\pi} + w\text{VM}(\hat{s}, s, \kappa_{\text{fit}}) + (1 - w - w_{\text{guess}}) \frac{1}{N-1} \sum_{j=1}^{N-1} \text{VM}(\hat{s}, s_j, \kappa_{\text{fit}}),$$

where  $\hat{s}$  is the reported value,  $s$  is the target value,  $s_j$  is the  $j$ th nontarget value,  $w_{\text{guess}}$  is the guessing rate, and the sum runs over all nontarget items. This model has one parameter more than Eq. S22. The second modified mixture reflects the possibility that the nontarget weight depends on the distance (along the circle) between the target and the nontarget location,

$$p_{\text{fit}}(\hat{s} | s; w, \kappa_{\text{fit}}) = \frac{w_{\text{guess}}}{2\pi} + \frac{1 - w_{\text{guess}}}{2\pi I_0(\kappa_{\text{fit}})} \frac{\sum_{j=1}^N w_{d_j} e^{\kappa_{\text{fit}} \cos(\hat{s} - s_j)}}{\sum_{j=1}^N w_{d_j}},$$

where  $d_j$  is the distance along the circle between the target and the  $j$ th item in units of the minimum distance. It can take integer values from 0 to 4, with 0 corresponding to the target. The normalization of the weights in the second term is needed because items occupy different sets of locations on different trials, but the overall distribution must always be normalized; therefore, the weights can only be relative. This mixture model has a total of six free parameters.

We compared the original descriptive mixture fit, Eq. S22, to its two variations. We applied the Bayesian information criterion to correct for the number of free parameters. When  $\log L_{\text{max}}$  is the maximum log likelihood of a model, the Bayesian information criterion (9) is  $\text{BIC} = \log L_{\text{max}} - \frac{k}{2} \log N_{\text{trials}}$ , where  $k$  is the number of free parameters (two, three, or six) and  $N_{\text{trials}}$  is the number of trials.

**Bayesian model comparison.** Bayesian model comparison is a powerful method to compare models, because it can use individual-trial responses instead of summary statistics and because it automatically penalizes models with more free parameters (10). We explain the method for delayed estimation; for change localization, it is analogous. Each model  $m$  produces a predicted error distribution  $p(\Delta s_i; m, N_i, \mathbf{t})$ , where  $\mathbf{t}$  denotes the model parameters. Bayesian model comparison consists of calculating for each model the probability of finding a subject's actual responses under this distribution, averaged over free parameters,

$$L(m) = p(\text{data} | m) = \int p(\text{data} | m, \mathbf{t}) p(\mathbf{t} | m) d\mathbf{t} = \int \left( \prod_{i=1}^{N_{\text{trials}}} p(\Delta s_i; m, N_i, \mathbf{t}) \right) p(\mathbf{t} | m) d\mathbf{t},$$

where  $\Delta s_i$  and  $N_i$  are the estimation error and set size on the  $i$ th trial, respectively. It is convenient to take the logarithm and rewrite it as

$$\log L(m) = \log L_{\text{max}}(m) + \log \int \exp(\log L(m; \mathbf{t}) - \log L_{\text{max}}(m)) p(\mathbf{t} | m) d\mathbf{t}, \quad [\text{S23}]$$

where  $\log L(m; \mathbf{t}) = \sum_{i=1}^{N_{\text{trials}}} \log p(\Delta s_i; m, N_i, \mathbf{t})$  and  $L_{\text{max}}(m) = \max_{\mathbf{t}} L(m; \mathbf{t})$ . This form prevents numerical problems, because the exponential in the integrand of Eq. S23 is now of order

1 near the maximum-likelihood value of  $\mathbf{t}$ . For the prior, we assume a uniform distribution across a plausible range (Table S1), whose size we denote  $S_j$  for the  $j$ th parameter. Then Eq. S23 becomes

$$\log L(m) = \log L_{\text{max}}(m) - \sum_{j=1}^{\dim \mathbf{t}} \log S_j + \log \int \exp(\log L(m; \mathbf{t}) - \log L_{\text{max}}(m)) d\mathbf{t},$$

where  $\dim \mathbf{t}$  is the number of parameters. We approximated the integral through a Riemann sum, with 25 bins in each parameter dimension (we verified that this is a sufficiently large number to give robust results, Fig. S11B). The ratio of likelihoods of two models is also known as a Bayes factor. As an alternative to Bayesian model comparison, the Bayesian information criterion

is  $\text{BIC} = \log L_{\text{max}} - \frac{\dim \mathbf{t}}{2} \log N_{\text{trials}}$ .

**Numerical robustness.** Because the model predictions for the VP models could not be computed analytically, we used Monte Carlo simulations. To obtain the model predictions, we drew 250 samples per combination of parameters, per stimulus. To verify whether 250 is a sufficiently high number, we checked how many samples are approximately needed for the model likelihoods to converge. The result shows that about 10 samples are needed (Fig. S11A). Hence, 250 samples is a sufficiently high number for obtaining reliable results.

For each model, we numerically approximated the integral over parameter space in the Bayesian model comparison by a Riemann sum. For all models, we discretized the parameter dimensions into 25 bins (except for  $K$  in the IL and SA models, because that parameter takes integer values only between 1 and 8). Results from running the analysis with different numbers of bins shows that about 15 bins are needed for convergence (Fig. S11B). Hence, 25 is a sufficiently high number of bins to obtain reliable results.

## SI Results

**Raw Data.** An example of the descriptive mixture fits (*Methods*) to the histograms of estimation error of a single subject at all set sizes is shown in Fig. S2.

**Scrolling vs. Color Wheel.** Two conditions were used in experiment 1: responding using a color wheel and responding using scrolling. In the color wheel condition, subjects responded by a mouse click on an annulus composed of all 180 colors that were used for the stimulus array and centered at the center of the screen with a radius of  $8.2^\circ$  and a width of  $2.2^\circ$  (Fig. S3A). The color wheel was randomly rotated on each trial. In the scrolling condition, subjects used the left and right arrow keys to scroll through all possible colors. After the first key press, a random color appeared within the thicker circle. We find that  $w$  declines with set size,  $N$ , in both conditions (Fig. S3B). Using a two-way repeated-measures ANOVA with factors set size (1–8) and response modality (color wheel or scrolling), we find that  $w$  is significantly different between response modalities [main effect of set size,  $F(2, 24) = 64.3$ ,  $P < 0.001$ ; main effect of response modality,  $F(1, 12) = 22.5$ ,  $P < 0.001$ ]. At all set sizes except 1 and 8, a paired  $t$  test shows a significant difference between response modalities ( $P < 0.01$ ). Estimated capacities are  $4.5 \pm 0.3$  and  $3.5 \pm 0.3$  in scrolling and color wheel conditions, respectively, constituting a significant difference [two-tailed paired  $t$  test,  $t(12) = -2.94$ ,  $P < 0.05$ ]. This result shows that the rate of (apparent) guessing increases with set size but is substantially smaller in the scrolling condition than in the color wheel condition.









