Optimal inference of sameness
Supporting information

Contents

1 Decision rules of the optimal observer

Since equal reliabilities (Experiment 1) are a special case of unequal reliabilities (Experiment 2), we treat the latter first.

1.1 Unequal reliabilities

The generative model of the task is given in Figure 2A. The variables are as follows: C is a binary variable that denotes sameness (+1 for same, –1 for different), $\mathbf{s}$ denotes the vector of N orientations presented, and $\mathbf{x}$ denotes the corresponding vector of N internal representations. Each x_i is drawn from a Gaussian distribution with mean s_i and standard deviation σ_i. The Gaussian assumption is common, and reflects the presence of many independent, unbiased sources of noise (1). Although orientation space is circular, we can treat it as the real line, because the internal representations are close to the true stimulus and σ_i is much smaller than the

circle circumference. We refer to $1/\sigma_i^2$ as the reliability of the i^{th} observation. When the stimuli are the same, each orientation s_i is equal to μ. When the stimuli are different, each s_i is drawn independently from a Gaussian distribution with mean μ and standard deviation σ_s. Regardless of sameness, the value of μ is drawn from a uniform distribution on $[-L,L)$ (one could take $L=\pi/2$ here, but any L works).

Optimal (Bayesian) observers base their decision on the posterior probability distribution over the variable of interest, here C, given single-trial observations, here $\mathbf{x}=(x_1,\ldots,x_N)^{\mathrm{T}}$. Since C is a binary random variable, this posterior reduces to a single number, which can be expressed in many ways. A particularly convenient way is to express it as a log posterior ratio:

$$d = \log\frac{p(C=1\,|\,\mathbf{x})}{p(C=-1\,|\,\mathbf{x})} = \log\frac{p(\mathbf{x}\,|\,C=1)}{p(\mathbf{x}\,|\,C=-1)} + \log\frac{p(C=1)}{p(C=-1)}. \tag{S1}$$

Evaluating the likelihoods in this expression, $p(\mathbf{x}|C)$ for $C=\pm 1$, requires marginalization over the stimulus orientations, $\mathbf{s}=(s_1,\ldots,s_N)^{\mathrm{T}}$, and their mean, μ:

$$p(\mathbf{x}\,|\,C) = \iint p(\mathbf{x}\,|\,\mathbf{s})\,p(\mathbf{s}\,|\,C,\mu)\,p(\mu)\,d\mathbf{s}\,d\mu. \tag{S2}$$

We assume that the standard deviation, σ_i, of the noise associated with a stimulus is known to the observer for each stimulus and each trial, as would be the case in a Poisson-like population code (2). Therefore, we do not need to marginalize over σ_i, but can treat it as a known parameter. We now evaluate Eq. (S2) by substituting the known distributions. If $\mathbf{1}$ denotes the column vector of length N with all entries equal to 1, and $\delta(x)$ is the Dirac Delta distribution, then

$$\begin{aligned} p(\mathbf{x}\,|\,C=1) &= \frac{1}{2L}\iint p(\mathbf{x}\,|\,\mathbf{s})\,\delta(\mathbf{s}-\mu\mathbf{1})\,d\mathbf{s}\,d\mu \\ &= \frac{1}{2L}\int p(\mathbf{x}\,|\,\mathbf{s}=\mu\mathbf{1})\,d\mu \\ &= \frac{1}{2L}\int \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right)d\mu \\ &= \frac{1}{2L}\int \frac{1}{\prod_{i=1}^{N}\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma_i^2}\right)d\mu. \end{aligned} \tag{S3}$$

In the rest of this document, all sums and products run from 1 to N unless mentioned otherwise. We rewrite the sum in the exponent as

$$\begin{aligned}\sum_i \frac{(x_i-\mu)^2}{\sigma_i^2} &= \sum_i \frac{x_i^2}{\sigma_i^2} - 2\mu\sum_i \frac{x_i}{\sigma_i^2} + \mu^2 \sum_i \frac{1}{\sigma_i^2} \\ &= \left(\sum_i \frac{1}{\sigma_i^2}\right)\left(\mu - \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}\right)^2 + V_1,\end{aligned} \tag{S4}$$

with

$$V_1 = \sum_i \frac{x_i^2}{\sigma_i^2} - \frac{\left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2}{\sum_i \frac{1}{\sigma_i^2}}.$$

If all reliabilities are equal, so that $\sigma_i = \sigma$, then V_1 simplifies to

$$V_1 = \frac{1}{\sigma^2}\sum_i x_i^2 - \frac{1}{N\sigma^2}\left(\sum_i x_i\right)^2 = \frac{N\mathrm{Var}\,\mathbf{x}}{\sigma^2}.$$

Here, Var $\mathbf{x}$ is defined as $\mathrm{Var}\,\mathbf{x} = \frac{1}{N}\sum_i x_i^2 - \frac{1}{N^2}\left(\sum_i x_i\right)^2$. For the general case, in which each item comes with its own reliability, we substitute Eq. (S4) into Eq. (S3) to find

$$\begin{aligned}p(\mathbf{x}\,|\,C=1) &= \frac{\exp\left(-\frac{V_1}{2}\right)}{2L(2\pi)^{\frac{N}{2}}\prod_i \sigma_i}\int \exp\left(-\frac{1}{2}\left(\sum_i \frac{1}{\sigma_i^2}\right)\left(\mu - \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}\right)^2\right) d\mu \\ &= \frac{\exp\left(-\frac{V_1}{2}\right)}{2L(2\pi)^{\frac{N-1}{2}}\prod_i \sigma_i}\left(\sum_i \frac{1}{\sigma_i^2}\right)^{-\frac{1}{2}},\end{aligned}$$

where in the second equality, we assumed that the domain of μ is large compared to $\left(\sum_i \frac{1}{\sigma_i^2}\right)^{-\frac{1}{2}}$, and so the integral can be evaluated over the entire real line. This assumption is justified here, because in our task the maximum-likelihood estimate of a stimulus is distributed around the true value of the stimulus in a range that is much narrower than the entire circular space.

Starting from Eq. (S2), we repeat this calculation for the hypothesis that the stimuli are different, that is $C = -1$:

$$
\begin{aligned}
p(\mathbf{x}\mid C=-1) &= \frac{1}{2L}\int_{-L}^{L}\int\left(\prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(x_i-s_i)^2}{2\sigma_i^2}\right)\right)\left(\prod_i \frac{1}{\sqrt{2\pi\sigma_s^2}}\exp\left(-\frac{(s_i-\mu)^2}{2\sigma_s^2}\right)\right)d\mathbf{s}d\mu \\
&= \frac{1}{2L}\int_{-L}^{L}\prod_i\left(\int \frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(x_i-s_i)^2}{2\sigma_i^2}\right)\frac{1}{\sqrt{2\pi\sigma_s^2}}\exp\left(-\frac{(s_i-\mu)^2}{2\sigma_s^2}\right)ds_i\right)d\mu \\
&= \frac{1}{2L}\int_{-L}^{L}\prod_i\left(\frac{1}{\sqrt{2\pi\left(\sigma_i^2+\sigma_s^2\right)}}\exp\left(-\frac{(x_i-\mu)^2}{2\left(\sigma_i^2+\sigma_s^2\right)}\right)\right)d\mu \\
&= \frac{\exp\left(-\frac{V_{-1}}{2}\right)}{2L(2\pi)^{\frac{N}{2}}\prod_i\sqrt{\sigma_i^2+\sigma_s^2}}\int_{-L}^{L}\exp\left(-\frac{1}{2}\left(\sum_i\frac{1}{\sigma_i^2+\sigma_s^2}\right)\left(\mu-\frac{\sum_i\frac{x_i}{\sigma_i^2+\sigma_s^2}}{\sum_i\frac{1}{\sigma_i^2+\sigma_s^2}}\right)^2\right)d\mu \\
&= \frac{\exp\left(-\frac{V_{-1}}{2}\right)}{2L(2\pi)^{\frac{N-1}{2}}\prod_i\sqrt{\sigma_i^2+\sigma_s^2}}\left(\sum_i\frac{1}{\sigma_i^2+\sigma_s^2}\right)^{-\frac{1}{2}},
\end{aligned}
$$

where

$$
V_{-1} = \sum_i\frac{x_i^2}{\sigma_i^2+\sigma_s^2} - \frac{\left(\sum_i\frac{x_i}{\sigma_i^2+\sigma_s^2}\right)^2}{\sum_i\frac{1}{\sigma_i^2+\sigma_s^2}}.
$$

The log likelihood ratio, the first term on the right-hand side of Eq. (S1), now becomes

$$
\begin{aligned}
\log\frac{p(\mathbf{x}\mid C=1)}{p(\mathbf{x}\mid C=-1)} &= -\frac{V_1}{2}+\frac{V_{-1}}{2}+\frac{1}{2}\sum_i\log\frac{\sigma_i^2+\sigma_s^2}{\sigma_i^2}-\frac{1}{2}\log\frac{\sum_i\frac{1}{\sigma_i^2}}{\sum_i\frac{1}{\sigma_i^2+\sigma_s^2}} \\
&= \frac{1}{2}\sum_i\left(\tilde{w}_i-w_i\right)x_i^2+\frac{\left(\sum_i w_i x_i\right)^2}{2\sum_i w_i}-\frac{\left(\sum_i \tilde{w}_i x_i\right)^2}{2\sum_i \tilde{w}_i}+\frac{1}{2}\sum_i\log\frac{w_i}{\tilde{w}_i}-\frac{1}{2}\log\frac{\sum_i w_i}{\sum_i \tilde{w}_i} \\
&= \frac{1}{2}\left(\sum_i\left(\tilde{w}_i-w_i\right)x_i^2+\frac{\sum_{ij} w_i w_j x_i x_j}{\sum_i w_i}-\frac{\sum_{ij}\tilde{w}_i\tilde{w}_j x_i x_j}{\sum_i \tilde{w}_i}+\sum_i\log\frac{w_i}{\tilde{w}_i}-\log\frac{\sum_i w_i}{\sum_i \tilde{w}_i}\right),
\end{aligned}
$$

where $\mathbf{w}$ and $\tilde{\mathbf{w}}$ are column vectors with entries

$$w_i = \frac{1}{\sigma_i^2}, \quad \tilde{w}_i = \frac{1}{\sigma_i^2 + \sigma_s^2}.$$

Therefore, the log posterior ratio, d, is a quadratic form in $\mathbf{x}$. In matrix notation, we can write

$$\boxed{d = \frac{1}{2}\left(-\mathbf{x}^{\mathrm{T}} \mathbf{A}(\boldsymbol{\sigma}) \mathbf{x} + \sum_i \log \frac{w_i}{\tilde{w}_i} - \log \frac{\sum_i w_i}{\sum_i \tilde{w}_i} \right) + \log \frac{p_{\text{same}}}{1 - p_{\text{same}}},} \tag{S5}$$

where $p(C{=}1){=}p_{\text{same}}$ and $\mathbf{A}(\boldsymbol{\sigma})$ is a symmetric, reliability-dependent matrix with entries

$$\boxed{A_{ij} = (w_i - \tilde{w}_i)\delta_{ij} + \frac{\tilde{w}_i \tilde{w}_j}{\sum_k \tilde{w}_k} - \frac{w_i w_j}{\sum_k w_k}.} \tag{S6}$$

Maximum-a-posteriori estimation amounts to responding "same" whenever d is positive. This leads to Eq. (2) in the main text.

1.2 Equal reliabilities

When all reliabilities are equal, $\sigma_i{=}\sigma$ and therefore $w_i{=}w$ and $\tilde{w}_i = \tilde{w}$, as in Experiment 1, the entries of $\mathbf{A}$ reduce to $A_{ij} = (w - \tilde{w})\left(\delta_{ij} - \frac{1}{N}\right)$, and the quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ becomes $N(w - \tilde{w})\operatorname{Var}\mathbf{x}$. The decision rule, which describes when the optimal observer responds "C=1", takes the form

$$\boxed{N \operatorname{Var}\mathbf{x} < \frac{1}{w - \tilde{w}}\left[(N-1)\log\frac{w}{\tilde{w}} + 2\log\frac{p_{\text{same}}}{1 - p_{\text{same}}} \right].} \tag{S7}$$

Thus, the decision rule amounts to comparing the variance to a criterion that depends in a specific way on the common reliability of the items (hidden in w and $\tilde{w}$) and on set size, N.

2 Response probabilities of the optimal observer

In the previous section, we derived the decision rules of the optimal observers for both experiments. However, these decision rules cannot be tested directly, since as experimenters we do not have access to the internal observations **x**. We only know the presented stimuli, **s**, and the subject's response on each trial. To obtain testable model predictions, we compute the theoretical probability of a "same" response given by the optimal observer, when the observations **x** are drawn from the generative model for given **s**. In the resulting model predictions, **x** is no longer present. The complete model thus consists of two parts "merged together": the generative model, producing a theoretical (or simulated) set of observations, and the inference model, which describes the decision made by the optimal observer based on those observations.

Formally, we denote the observer's estimate of C by $\hat{C}$; for the optimal observer, it equals the sign of d. The optimal observer responds according to this estimate. Eq. (S5) determines the optimal decision rule on a single trial. We denote the probability of the optimal observer responding $\hat{C}$ when the experimenter presents a specific stimulus set **s** by $p\left(\hat{C} \mid \mathbf{s}\right)$. It is obtained by averaging over the hypothetical observations **x** generated by **s**:

$$p\left(\hat{C} \mid \mathbf{s}\right) = \int p\left(\hat{C} \mid \mathbf{x}\right) p\left(\mathbf{x} \mid \mathbf{s}\right) d\mathbf{x} = \int \delta_{\hat{C},\mathrm{sgn}\left(d(\mathbf{x})\right)} p\left(\mathbf{x} \mid \mathbf{s}, \boldsymbol{\sigma}\right) d\mathbf{x}\,, \tag{S8}$$

where δ is the Kronecker delta. In words, the probability of responding $\hat{C} = 1$ is equal to the proportion of observations drawn from $p(\mathbf{x}|\mathbf{s})$ for which $d(\mathbf{x})$ is positive. For many psychophysical tasks, an integral like that given in Eq. (S8) is analytically intractable, and needs to be approximated, for example, by Monte Carlo simulation. In the present task, an analytical treatment is possible.

2.1 Equal reliabilities

We first consider the case of Experiment 1, with the decision rule given by Eq. (S7). We are interested in the probability that this inequality is satisfied for **x** drawn from a multivariate normal distribution with mean **s** and covariance matrix $\sigma^2\mathbf{I}$, denoted $\mathbf{x} \sim \mathcal{N}(\mathbf{s}, \sigma^2\mathbf{I})$, with **I** the identity matrix. We write Var $\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{A}'\mathbf{x}$, where the components of $\mathbf{A}'$ are $A'_{ij} = \delta_{ij}/N - 1/N^2$. This matrix has $N-1$ eigenvalues equal to 1, and one eigenvalue equal to 0. Therefore, for an orthogonal matrix **M** and diagonal matrix $\mathbf{D} = \mathrm{diag}(1,1,\ldots,1,0)$, we can write $\mathbf{M}^{\mathrm{T}}\mathbf{A}'\mathbf{M} = \mathbf{D}$. Let **Y** be $\mathbf{Y} = \mathbf{M}^{\mathrm{T}}\mathbf{x}$. Since **M** is orthogonal, $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, with $\boldsymbol{\mu} = \mathbf{M}^{\mathrm{T}}\mathbf{s}$. Since

$$\sum_{i=1}^{N-1} \frac{y_i^2}{\sigma^2} = \frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}\mathbf{D}\mathbf{Y} = \frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{D}\mathbf{M}^{\mathrm{T}}\mathbf{x} = \frac{1}{\sigma^2}\mathbf{x}^{\mathrm{T}}\mathbf{A}'\mathbf{x}\,, \tag{S9}$$

it follows that $\mathbf{x}^{\mathrm{T}}\mathbf{A}'\mathbf{x}/\sigma^2$ follows a non-central chi-squared distribution with $N-1$ degrees of

freedom. This variable is equal to $wN\mathrm{Var}\,\mathbf{x}$. The noncentrality parameter of the sum on the left of Eq. (S9) is $\sum_{i=1}^{N-1}\frac{\mu_i^2}{\sigma^2}$, which can be rewritten as

$$\sum_{i=1}^{N-1}\frac{\mu_i^2}{\sigma^2}=\frac{1}{\sigma^2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{D}\boldsymbol{\mu}=\frac{1}{\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{M}\mathbf{D}\mathbf{M}^{\mathrm{T}}\mathbf{s}=\frac{1}{\sigma^2}\mathbf{s}^{\mathrm{T}}\mathbf{A}'\mathbf{s}=wN\,\mathrm{Var}\,\mathbf{s}\,.$$

Therefore, we have

$$wN\,\mathrm{Var}\,\mathbf{x}\sim\chi^2_{N-1}\left(wN\,\mathrm{Var}\,\mathbf{s}\right). \tag{S10}$$

From Eq. (S7), the probability of responding "same" equals the probability that wN Var $\mathbf{x}$ is less than $\frac{w}{w-\tilde{w}}\left[(N-1)\log\frac{w}{\tilde{w}}+2\log\frac{p_{\mathrm{same}}}{1-p_{\mathrm{same}}}\right]$. Using Eq. (S10), this probability is obtained from a cumulative chi-squared distribution:

$$\boxed{p\left(\hat{C}=1\,|\,\mathbf{s}\right)=\Pr\left(\chi^2_{N-1}\left(wN\,\mathrm{Var}\,\mathbf{s}\right)<\frac{w}{w-\tilde{w}}\left[(N-1)\log\frac{w}{\tilde{w}}+2\log\frac{p_{\mathrm{same}}}{1-p_{\mathrm{same}}}\right]\right)}. \tag{S11}$$

2.2 Unequal reliabilities

We next consider the case of Experiment 2, where reliability can differ between stimuli. Then $\mathbf{x}$ follows a multivariate normal distribution with mean $\mathbf{s}$ and covariance matrix $\boldsymbol{\Sigma}=\mathrm{diag}(\sigma_1^2,\ldots\sigma_N^2)$. We define a random variable $Q(\mathbf{x})=\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$, where $\mathbf{B}$ is a non-negative definite and symmetric matrix. We denote mean and standard deviation of Q by μ_Q and σ_Q, respectively. Liu, Tang, and Zhang (3) have found the following approximation to the distribution of Q:

$$\Pr\left(\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}<k\right)\approx\Pr\left(\chi_l^2\left(\delta\right)<t\sigma_\chi+\mu_\chi\right) \tag{S12}$$

where the variables l, δ, t, σ_χ, and μ_χ are defined as follows:

$$t = \frac{k - \mu_Q}{\sigma_Q}$$

$$l = 3a^2 - 2r_1 a^3$$

$$\delta = r_1 a^3 - a^2$$

$$\mu_\chi = 2a^2 - r_1 a^3$$

$$\sigma_\chi = a\sqrt{2}$$

$$a = \begin{cases} \dfrac{1}{r_1 - \sqrt{r_1^2 - r_2}}, & \text{if } r_1^2 > r_2 \\ \dfrac{1}{r_1}, & \text{if } r_1^2 \le r_2 \end{cases}$$

$$r_1 = \frac{c_3}{{c_2}^{3/2}}$$

$$r_2 = \frac{c_4}{{c_2}^2}$$

$$c_j = \mathrm{Tr}\left(\left(\mathbf{B\Sigma}\right)^j\right) + j\mathbf{s}^{\mathrm{T}}\left(\mathbf{B\Sigma}\right)^{j-1}\mathbf{Bs}$$

Numerical simulation shows that for our purposes, this is a very good approximation. In order to apply this approximation to $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ from Eq.(S5), **A** has to be both symmetric and non-negative definite. Since **A** is obviously symmetric, what remains to be shown is that in our case, **A** is non-negative definite.

2.3 Proof that A is non-negative definite in the unequal-reliabilities case

The matrix **A** specified by Eq. (S6) can be written as

$$\mathbf{A} = \mathrm{diag}\left(\mathbf{w} - \tilde{\mathbf{w}}\right) + \frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^{\mathrm{T}}}{\tilde{\mathbf{w}}^{\mathrm{T}}\mathbf{1}} - \frac{\mathbf{w}\mathbf{w}^{\mathrm{T}}}{\mathbf{w}^{\mathrm{T}}\mathbf{1}}.$$

To show that A is non-negative definite (positive semidefinite), we prove that its eigenvalues are non-negative. In Experiment 2, m of the σ_i's are equal to σ_{low}, and $N–m$ are equal to σ_{high}. Without loss of generality, we arrange them as $\boldsymbol{\sigma}=(\sigma_{\mathrm{low}},\ldots,\sigma_{\mathrm{low}}, \sigma_{\mathrm{high}},\ldots,\sigma_{\mathrm{high}})$, *i.e.*, the first m correspond to low-reliability stimuli, and the last $N–m$ to high-reliability ones. Consequently, $w_1=\ldots=w_m=w_{\mathrm{low}}=1/{\sigma_{\mathrm{low}}}^2$ and $w_{m+1}=\ldots=w_N=w_{\mathrm{high}}=1/{\sigma_{\mathrm{high}}}^2$. The $(i,j)^{\mathrm{th}}$ term of **A** is given by Eq. (S6). Note that

$$\left(\mathbf{A1}\right)_i = \sum_{j=1}^{N} A_{ij} = w_i - \tilde{w}_i + \sum_{j=1}^{N}\left(\frac{\tilde{w}_i\tilde{w}_j}{\sum_k \tilde{w}_k} - \frac{w_i w_j}{\sum_k w_k}\right) = 0\,, \tag{S13}$$

we see that $\mathbf{1}=(1,\ldots,1)^{\mathrm{T}}$ is an eigenvector with eigenvalue 0. We now define a set of $m-1$ vectors $\mathbf{v}^k$, with $k=2,\ldots,m$, as follows:

$$v_i^k = \delta_{i1} - \delta_{ik}\,.$$

For example, $\mathbf{v}^3=(1,0,-1,0,\ldots 0)^{\mathrm{T}}$. These vectors are eigenvectors of $\mathbf{A}$ with eigenvalues $w_{\mathrm{low}} - \tilde{w}_{\mathrm{low}}$, since

$$\begin{aligned}
\left(\mathbf{A}\mathbf{v}^k\right)_i &= \sum_{j=1}^{N} A_{ij} v_j^k = \sum_{j=1}^{N} A_{ij}\left(\delta_{j1} - \delta_{jk}\right) \\
&= \sum_{j=1}^{N}\left(\left(w_i - \tilde{w}_i\right)\delta_{ij} + \frac{\tilde{w}_i\tilde{w}_j}{\sum_k \tilde{w}_k} - \frac{w_i w_j}{\sum_k w_k}\right)\left(\delta_{j1} - \delta_{jk}\right) \\
&= \sum_{j=1}^{N}\left(\left(w_i - \tilde{w}_i\right)\delta_{ij}\left(\delta_{j1} - \delta_{jk}\right) + \left(\frac{\tilde{w}_i\tilde{w}_j}{\sum_k \tilde{w}_k} - \frac{w_i w_j}{\sum_k w_k}\right)\left(\delta_{j1} - \delta_{jk}\right)\right) \\
&= \left(w_i - \tilde{w}_i\right)\left(\delta_{i1} - \delta_{ik}\right) + \frac{\tilde{w}_i\left(\tilde{w}_1 - \tilde{w}_k\right)}{\sum_k \tilde{w}_k} - \frac{w_i\left(w_1 - w_k\right)}{\sum_k w_k} \\
&= \left(w_{\mathrm{low}} - \tilde{w}_{\mathrm{low}}\right)\left(\delta_{i1} - \delta_{ik}\right) = \left(w_{\mathrm{low}} - \tilde{w}_{\mathrm{low}}\right) v_i^k\,.
\end{aligned}$$

Similarly, the $N-m-1$ vectors $\mathbf{v}^k$ with $k=m+2,\ldots,N$, whose i^{th} entries are $v_i^k = \delta_{i,m+1} - \delta_{ik}$ are eigenvectors with eigenvalues $w_{\mathrm{high}} - \tilde{w}_{\mathrm{high}}$. That leaves one eigenvalue to find. Consider the vector $\mathbf{v}=(N-m,\ldots, N-m,-m,\ldots,-m)^{\mathrm{T}}$. We will show that this vector is an eigenvector. We first calculate the product of any of the first m rows of A with $\mathbf{v}$, making use of Eq. (S13)

$$
\begin{aligned}
\left(\mathbf{Av}\right)_i &= \left(N-m\right)\sum_{j=1}^{m} A_{ij} - m\sum_{j=m+1}^{N} A_{ij} \\
&= \left(N-m\right)\left(w_{\text{low}} - \tilde{w}_{\text{low}}\right) + m\left(N-m\right)\left(\frac{\tilde{w}_{\text{low}}^2}{\sum_k \tilde{w}_k} - \frac{w_{\text{low}}^2}{\sum_k w_k}\right) - m\left(N-m\right)\left(\frac{\tilde{w}_{\text{low}}\tilde{w}_{\text{high}}}{\sum_k \tilde{w}_k} - \frac{w_{\text{low}}w_{\text{high}}}{\sum_k w_k}\right) \\
&= -\left(N-m\right)\left(N-m\right)\left(\frac{\tilde{w}_{\text{low}}\tilde{w}_{\text{high}}}{\sum_k \tilde{w}_k} - \frac{w_{\text{low}}w_{\text{high}}}{\sum_k w_k}\right) - m\left(N-m\right)\left(\frac{\tilde{w}_{\text{low}}\tilde{w}_{\text{high}}}{\sum_k \tilde{w}_k} - \frac{w_{\text{low}}w_{\text{high}}}{\sum_k w_k}\right) \\
&= N\left(N-m\right)\left(\frac{w_{\text{low}}w_{\text{high}}}{\sum_k w_k} - \frac{\tilde{w}_{\text{low}}\tilde{w}_{\text{high}}}{\sum_k \tilde{w}_k}\right) \\
&= N\left(N-m\right)\left(\frac{1}{m\sigma_{\text{low}}^2 + \left(N-m\right)\sigma_{\text{high}}^2} - \frac{1}{m\left(\sigma_{\text{low}}^2 + \sigma_{\text{s}}^2\right) + \left(N-m\right)\left(\sigma_{\text{high}}^2 + \sigma_{\text{s}}^2\right)}\right) \\
&= \left(N-m\right)\left(\frac{N^2\sigma_s^2}{\left(m\sigma_{\text{low}}^2 + \left(N-m\right)\sigma_{\text{high}}^2\right)\left(m\sigma_{\text{low}}^2 + \left(N-m\right)\sigma_{\text{high}}^2 + N\sigma_{\text{s}}^2\right)}\right).
\end{aligned}
$$

Similarly, for any of the last N–m rows of A, we get

$$
\left(\mathbf{Av}\right)_i = -m\left(\frac{N^2\sigma_{\text{s}}^2}{\left(m\sigma_{\text{low}}^2 + \left(N-m\right)\sigma_{\text{high}}^2\right)\left(m\sigma_{\text{low}}^2 + \left(N-m\right)\sigma_{\text{high}}^2 + N\sigma_{\text{s}}^2\right)}\right).
$$

Thus, $\dfrac{N^2\sigma_{\text{s}}^2}{\left(m\sigma_{\text{low}}^2 + \left(N-m\right)\sigma_{\text{high}}^2\right)\left(m\sigma_{\text{low}}^2 + \left(N-m\right)\sigma_{\text{high}}^2 + N\sigma_{\text{s}}^2\right)}$ is the final eigenvalue of A. Since all eigenvalues are non-negative, A is non-negative definite, and therefore we may apply the approximation discussed in Section 2.2.

3 Suboptimal models

3.1 Response probabilities for the single-criterion and blockwise-criterion models

Calculating a model's predictions for the probability of responding "same" given a set of presented orientations **s** consists of two steps: to determine the decision rule for responding "same", and to apply this decision rule to the set of internal representations on each individual trial. In non-optimal observer models, the first step (decision rule) is different, but the second step (calculating the probability with which the decision is satisfied) is identical. In particular, as long as the decision rule is of the form

$$\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} < k\,, \tag{S14}$$

with **B** a non-negative definite, symmetric matrix, we can still use Eq. (S12) to approximate the model response probabilities. The only difference from the optimal model is that different expressions must be substituted for **B** and/or k. In the special case that the decision rule is Var $\mathbf{x}<k$ and the reliabilities are equal, the model response probabilities are given by an exact expression analogous to Eq. (S11),

$$p\left(\hat{C}=1\,|\,\mathbf{s}\right)=\Pr\left(\chi^2_{N-1}\left(wN\,\mathrm{Var}\,\mathbf{s}\right)<wk\right).$$

3.2 Single-criterion model for Experiment 2

In Experiment 2, the SC observer assumes in the decision rule that all reliabilities are equal, $\sigma_i=\sigma_{\mathrm{assumed}}$ for all i. In other words, this observer does not weight the observations by their correct respective reliabilities. We will show here that this model is equivalent to one in which the observer compares the sample variance to a single criterion.

The decision rule is the same one as in Experiment 1, Eq. (S7), but with $\sigma_{\mathrm{assumed}}$ instead of σ. This rule can be rewritten as

$$\mathrm{Var}\,\mathbf{x} < \frac{\sigma^2_{\mathrm{assumed}}}{N}\left(1+\frac{\sigma^2_{\mathrm{assumed}}}{\sigma^2_{\mathrm{s}}}\right)\left[\left(N-1\right)\log\left(1+\frac{\sigma^2_{\mathrm{s}}}{\sigma^2_{\mathrm{assumed}}}\right)+2\log\frac{p_{\mathrm{same}}}{1-p_{\mathrm{same}}}\right]. \tag{S15}$$

Since $\sigma_{\mathrm{assumed}}$ is a free parameter, the right-hand side can assume any value on the real line. To see this, first note that this expression is continuous. Then compute the two limits

$$\lim_{\sigma^2_{\text{assumed}}\to\infty} \frac{\sigma^2_{\text{assumed}}}{N}\left(1+\frac{\sigma^2_{\text{assumed}}}{\sigma^2_{\text{s}}}\right)\left[(N-1)\log\left(1+\frac{\sigma^2_{\text{s}}}{\sigma^2_{\text{assumed}}}\right)+2\log\frac{p_{\text{same}}}{1-p_{\text{same}}}\right]=\infty$$

$$\lim_{\sigma^2_{\text{assumed}}\to 0} \frac{\sigma^2_{\text{assumed}}}{N}\left(1+\frac{\sigma^2_{\text{assumed}}}{\sigma^2_{\text{s}}}\right)\left[(N-1)\log\left(1+\frac{\sigma^2_{\text{s}}}{\sigma^2_{\text{assumed}}}\right)+2\log\frac{p_{\text{same}}}{1-p_{\text{same}}}\right]$$

$$=\frac{N-1}{N}\lim_{\sigma^2_{\text{assumed}}\to 0}\sigma^2_{\text{assumed}}\log\left(1+\frac{\sigma^2_{\text{s}}}{\sigma^2_{\text{assumed}}}\right)=0$$

Therefore, we can simply replace the entire right-hand side by a single free parameter k, i.e., Var $\mathbf{x}<k$, thereby justifying the terminology "single-criterion model". (Note that in order to obtain model predictions, we still need to fit the parameters σ_{low} and σ_{high}.) Thus, the SC model has only two variants: with and without lapse rate.

3.3 Blockwise-criterion model for Experiment 2

In Experiment 2, the BC model is the model in which σ_{assumed} may vary by block type (LOW, MIXED, or HIGH). Then, the decision rule is equivalent to Var $\mathbf{x}<k_{\text{block}}$, where each block type has its own free parameter k_{block}. This model is analogous to the BC model in Experiment 1, with reliability condition (LOW, MIXED, or HIGH) taking the place of set size. As in the SC model, there are only two variants: with and without lapse rate.

3.4 Maximum-absolute-difference models

For the MAD models, the response probabilities cannot be calculated analytically. We estimated these probabilities numerically, by simulating 10,000 sets of N internal representations for each of the 2700 experimental trials, and applying the model's decision rule to each set; the frequency of "same" responses was an estimate of the probability of responding "same" on the given trial.

4 Bayesian model comparison

We denote by $p(\hat{C}_i|\mathbf{s}_i,M,\boldsymbol{\theta})$ the probability predicted by model M with parameters $\boldsymbol{\theta}$ of the subject's *actual* response on the i^{th} trial, $\hat{C}_i$, when the presented stimuli are $\mathbf{s}_i$. We computed the log probability of the data given M by marginalizing over $\boldsymbol{\theta}$. For the prior over parameters, we assumed a uniform distribution, $p(\boldsymbol{\theta})=1/\text{Vol}_{\boldsymbol{\theta}}$, where $\text{Vol}_{\boldsymbol{\theta}}$ is the volume of parameter space. We calculated the parameter likelihood by assuming that the data are conditionally independent across trials: $p(\text{data}|M,\boldsymbol{\theta})=\prod_i p(\hat{C}_i|\mathbf{s}_i,M,\boldsymbol{\theta})$. For the MAD models, we approximated the marginalization through a Riemann sum; the parameter ranges were 1 to 40 for the decision criterion, 0.1° to 20.1° for σ, and 0 to 0.5 for the lapse rate, each in 35 steps. For the other models, since analytical expressions were available, we used the Laplace approximation (4); the sizes of the parameter ranges were 20° for σ, 20° for the assumed σ_{s}, 0.4 for p_{same}, and 0.4 for the lapse rate.

5 Simulations of animal cognition experiments

To examine whether the proportion of "different" responses from the optimal observer correlates with the entropy of the stimulus set, we simulated the stimulus sets used by Young et al. (5) (see Table 2 in their paper) and used these as input to the optimal-observer model. These sets always contained 16 items, with several subsets of identical items (e.g., 4 subsets of 4 identical items each). In our simulations, for each subset, a random stimulus value was drawn from a Gaussian distribution (with σ_s=5) and assigned to all items in that subset. A total of 1000 trials were simulated per stimulus set. On each trial, stimulus observations were simulated by adding Gaussian noise to the stimulus orientations (σ=10) and a response was generated by applying the decision rule from the optimal-observer model. Entropy and "scaled logit of percent different responses" were computed as described by Young et al. (5). To examine the effect of set size on the proportion of different responses in the optimal-observer model, we simulated 1000 trials per set size, with σ_s=8 and σ=8. To examine the effect of stimulus visibility, we varied internal noise, σ, and simulated 1000 trials per noise level (with σ_s=5, N=8, p_{same}=0.6, and a guessing rate of 0.25). While model parameters were separately chosen for each experiment, the observed trends were robust under a wide range of parameters values.

6 Experiment 1 (color)

Methods

The methods for Experiment 1, color, were identical to the Experiment 1 for orientation, except for the following differences. Stimuli consisted of a set of colored discs with a radius of 0.4 deg. The colors of the discs were drawn independently from 180 color values uniformly distributed along a circle of radius 50 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 grey background of luminance 10 cd/m^2. Set size was chosen randomly on each trial. In the models, "blockwise criterion" was now defined as one criterion per set size, rather than per block. The experiment consisted of 3 sessions with 6 blocks of 150 trials. Two authors and five paid, naive subjects participated in the experiment.

Results

The SC models can be ruled out but not the BC models (Fig. S1A-B; log likelihood differences between the optimal model and the SC, BC, MAD-SC, and MAD-BC models were 8.7±4.9, –3.9±2.5, 27.3±6.3, and 0.8±2.3, respectively). Even though the BC model just edges out the optimal model, the decision criteria and noise levels from the best BC model are very close to those predicted by the best optimal-model variant (Fig. S1C; p>0.1 for all six t-tests). Finally, the noise level exhibits a weak dependence on set size (Fig. S1C; power law fit yields a power of 0.23±0.04).

7 Supplementary Tables

Table S1. Overview of model parameters. The models have the same parameters in both experiments, except that in Experiment 1, $\boldsymbol{\sigma}=(\sigma_{N=2}, \sigma_{N=4}, \sigma_{N=8})$ and $\mathbf{k}=(k_{N=2}, k_{N=4}, k_{N=8})$, while in Experiment 2, $\boldsymbol{\sigma}=(\sigma_{low}, \sigma_{high})$ and $\mathbf{k}=(k_{LOW}, k_{MIXED}, k_{HIGH})$.

Model	Free parameters
Optimal	$\boldsymbol{\sigma}$
Optimal	$\boldsymbol{\sigma}$, λ
Optimal	$\boldsymbol{\sigma}$, σ_s
Optimal	$\boldsymbol{\sigma}$, σ_s, λ
Optimal	$\boldsymbol{\sigma}$, p_{same}
Optimal	$\boldsymbol{\sigma}$, p_{same}, λ
Optimal	$\boldsymbol{\sigma}$, p_{same}, σ_s
Optimal	$\boldsymbol{\sigma}$, p_{same}, σ_s, λ
Single-criterion	$\boldsymbol{\sigma}$, k
Single-criterion	$\boldsymbol{\sigma}$, k, λ
Blockwise-criterion	$\boldsymbol{\sigma}$, $\mathbf{k}$
Blockwise-criterion	$\boldsymbol{\sigma}$, $\mathbf{k}$, λ
MAD with single criterion	$\boldsymbol{\sigma}$, k
MAD with single criterion	$\boldsymbol{\sigma}$, k, λ
MAD with blockwise criterion	$\boldsymbol{\sigma}$, $\mathbf{k}$
MAD with blockwise criterion	$\boldsymbol{\sigma}$, $\mathbf{k}$, λ

Table S2. Overview of maximum-likelihood parameter values of the best fitting optimal models in Experiment 1. An empty cell indicates that the respective parameter was not a free parameter in the best-fitting model variant.

	p_{same}	σ_s	λ
AK	0.58	6.55	0.06
DS	0.42		0.04
HB			
MH			0.12
ML	0.40	11.5	
RB		11.3	
RC			0.06
TR	0.46		0.11

Table S3. Overview of maximum-likelihood parameter values of the best-fitting optimal models in Experiment 2. An empty cell indicates that the respective parameter was not a free parameter in the best-fitting model variant.

	p_{same}	σ_s	λ
BN	0.53		0.08
DB			0.11
DS		19.5	0.22
HB	0.45		0.11
KJ	0.45		0.07
MH	0.44		0.11
ML	0.35	13.7	
MV	0.46		0.04
RB	0.54	7.73	
RC	0.55		

8 Supplementary Figures

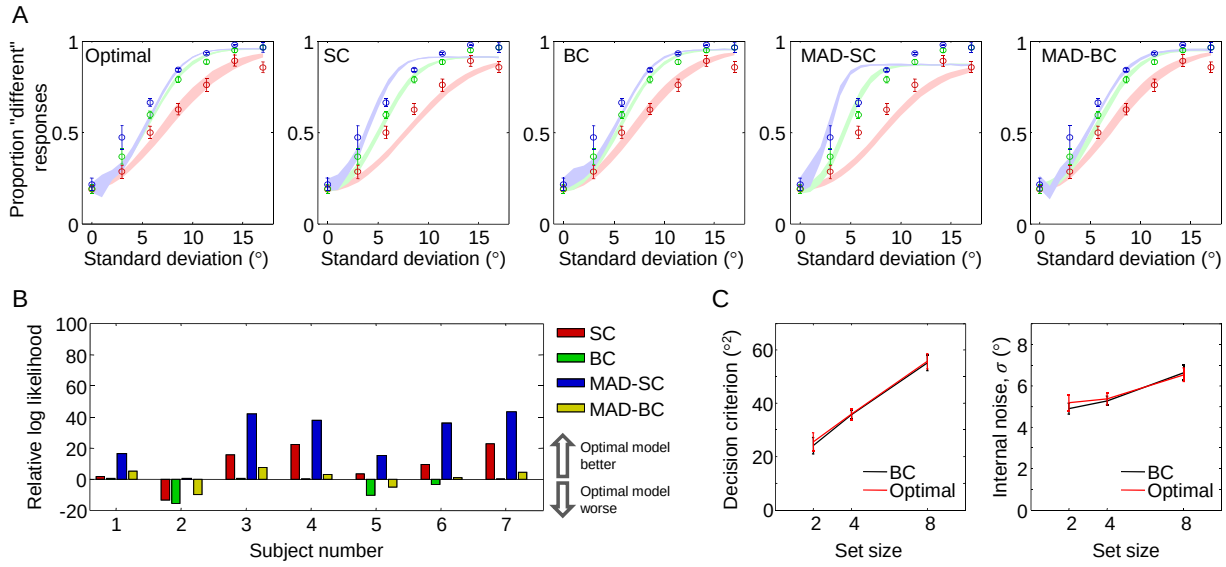

Fig. S1 Comparison of models in Experiment 3. Circles and error bars represent mean and s.e.m. of subject data. Shaded areas represent s.e.m. of model fits. (*A*) Proportion "different" responses as a function of sample standard deviation for optimal and suboptimal models. (*B*) Bayesian model comparison. Each bar represent the log likelihood of the optimal model minus that of a suboptimal model. (*C*) The decision criteria (left) and the internal noise levels (right) for the best-fitting BC model are nearly identical to those of the best-fitting optimal-observer model. This suggests that the human criteria are close to optimal.

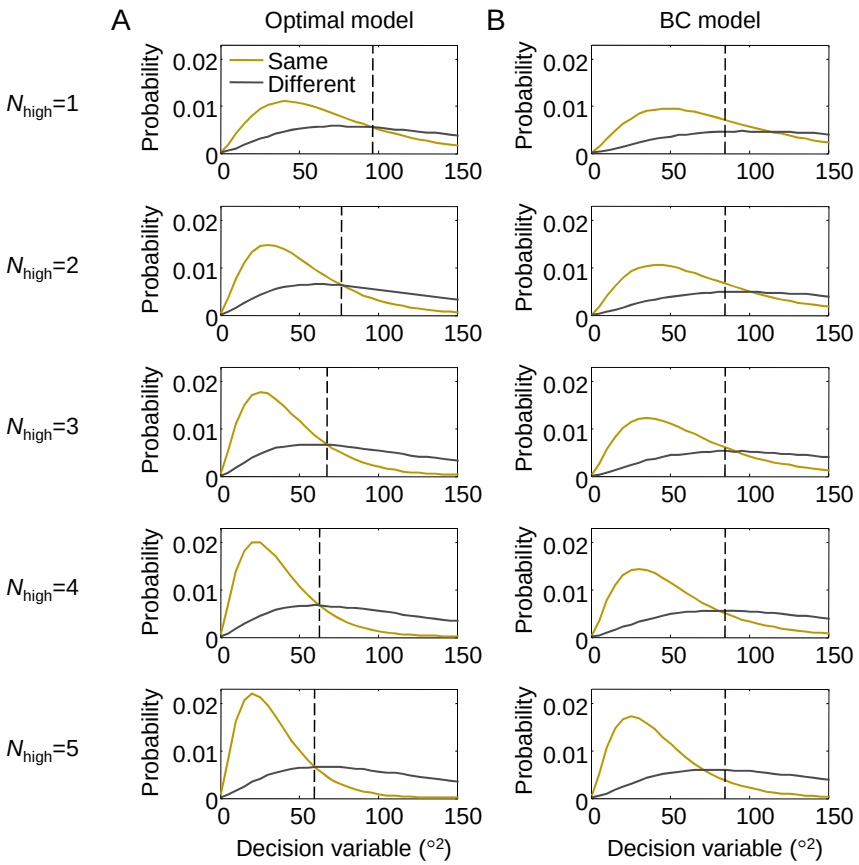

Fig. S2. A comparison of the decision strategies of the optimal and BC models in Experiment 2. (*A*) Distributions of the decision variable of the optimal model. As the number of high-reliability stimuli in the display, N_{high}, increases, the distributions become more separable, making the task easier. The optimal observer sets his criterion (dashed line) in such a way that performance is maximized. (*B*) Distributions of the decision variable of the BC model (Var **x**). This model uses the same criterion regardless of N_{high} (dashed line). Hence, it cannot maximize performance for all values of N_{high}. The same holds for the SC, MAD-SC, and MAD-BC models. The distributions differ between the models because the decision variables do.

1. Green DM & Swets JA (1966) *Signal detection theory and psychophysics* (John Wiley & Sons, Los Altos, CA).
2. Ma WJ, Beck JM, Latham PE, & Pouget A (2006) Bayesian inference with probabilistic population codes. *Nat Neurosci* 9(11):1432-1438.
3. Liu H, Tang Y, & Zhang HH (2009) A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comp Statistics and Data Analysis* 53:853-856.
4. MacKay DJ (2003) *Information theory, inference, and learning algorithms* (Cambridge University Press, Cambridge, UK).
5. Young ME & Wasserman EA (1997) Entropy detection by pigeons: response to mixed visual displays after same-different discrimination training. *J Exp Psychol Anim Behav Process* 23(2):157-170.