Title: Quantifying the importance of guessing, decision noise, and variable precision in explaining behavioral variability in visual perception

Authors: Shan Shen[1] and Wei Ji Ma[2*]

[1] Department of Neuroscience, Baylor College of Medicine

[2] Center for Neural Science and Department of Psychology, New York University

[*] Formerly at Department of Neuroscience, Baylor College of Medicine

[1]To whom correspondence should be addressed: Shan Shen, Department of Neuroscience, Baylor College of Medicine, TX, 77054, USA

E-mail: shans@bcm.edu

Conflict of Interest: The authors declare no competing financial interests.

ABSTRACT

When presented with the same sensory stimuli and performing the same task, people do not always make the same response. Such behavioral variability can have different causes, including sensory noise, decision noise, and guessing. In addition, recent work has proposed that the precision of sensory encoding is itself variable, both driven by the stimulus (heteroskedasticity) and independent of the stimulus. We analyzed data of 3 published and 8 new visual decision-making tasks with a single relevant feature, orientation. In modeling each experiment, we considered four factors: guessing (lapses), decision noise, orientation-dependent variable precision (oblique effect), and orientation-independent variable precision (inspired by visual working memory models). Modern computational power allows us to test all combinations of these factors; in a given model, each factor could be present or absent. To quantify the importance of each factor in explaining the data, we introduce three metrics: factor knock-in, factor knock-out and factor posterior probability. Across all experiments, we found strong evidence for guessing and for orientation-dependent variable precision. We found evidence for decision noise in only one experiment, and for orientation-independent variable precision only when distractors are variable across trials. On a methodological note, the factor importance metrics can be applied widely in factorial model comparison.

Keywords: visual perception, computational modeling, noise, variable precision, Bayesian inference

INTRODUCTION

When presented with the same stimuli in the same perceptual task, human observers do not always make the same response. The factors that affect such trial-to-trial behavioral variability have been investigated almost since the birth of psychophysics. A productive way to identify potential factors that affect behavioral variability is by following the standard schema of a perceptual process model: encoding, decision, and action (Fig. 1A). Encoding is the mapping from the stimulus to the internal representation. This mapping is known to be noisy at the neural level (Faisal, Selen, & Wolpert, 2008; London, Roth, Beeren, Häusser, & Latham, 2010; Tolhurst, Movshon, & Dean, 1983), and has long been modeled as noisy in behavioral models (Fechner, 1860; Green & Swets, 1966; Thurstone, 1927). It is common to assume that such sensory or encoding noise follows a zero-mean Gaussian distribution in the stimulus space (Green & Swets, 1966), or a Von Mises distribution when the stimulus variable is circular (Wilken & Ma, 2004; Zhang & Luck, 2008). Noise might also occur in the mapping from the internal representation to the decision; this is sometimes called decision noise (Mueller & Weidemann, 2008). Model mismatch (Orhan & Jacobs, 2013), statistical inefficiency (Burgess, Wagner, Jennings, & Barlow, 1981; Liu, Knill, & Kersten, 1995), or systematically suboptimal inference (Beck, Ma, Pitkow, Latham, & Pouget, 2012) could mimic decision noise. Decision noise has been modeled using a softmax function (Daw, O'Doherty, Dayan, Seymour, & Dolan,

2006; Soltani, 2006), as Gaussian noise on the log posterior ratio (Drugowitsch, Wyart, Devauchelle, & Koechlin, 2016; Keshvari, van den Berg, & Ma, 2012, 2013), or as Gaussian noise on the decision criterion (Mueller & Weidemann, 2008). Finally, there could be noise in the mapping from the decision to the motor action. This noise includes the motor noise (Trommershäuser, Maloney, & Landy, 2003a, 2003b; Wolpert & Landy, 2012) and lapses of attention that occasionally produce random responses (Wichmann & Hill, 2001).

Previous studies have tried to discriminate between different factors that affect behavioral variability. Using signal detection theory, some work focused on distinguishing "internal noise" and "statistical inefficiency" (Burgess et al., 1981; Liu et al., 1995; Pelli & Farell, 1999). These experiments were designed such that both factors would have qualitatively different effects on a psychometric curve, such as a threshold-versus-contrast curve. However, their approach has limitations: a) internal noise could include both sensory and decision components; b) the experiments were not set up to estimate variability in precision. Nowadays, these limitations can be overcome using quantitative model comparison, which allows one to further break down factors that affect behavioral variability. However, this approach requires much more computational power: for example, the main model fitting done for the present paper (11 experiments, 48 total subjects, 215400 total trials, 16 models fitted per experiment and per subject) took approximately 30×24×25 processor hours on a cluster of 3.2 GHz processors. Assuming Moore's law (Moore, 1998), the same code would have taken approximately 711 years to complete at the time of the Pelli & Farell paper. Basically, the kind of modeling we do here would not have been possible at that time.

One recent study took full advantage of this computational power to revisit the factors that affect behavioral variability (Drugowitsch et al., 2016). Using a paradigm that contains sensory encoding, accumulation of evidence, and binary choice, they were able to separate noise in the sensory, inference, and decision stages, and found that noise in the inference stage was most important to explain the data. The current study uses a similar quantitative model comparison approach, but is different in the following three aspects:

1) We consider variability in encoding precision. In recent years, a number of studies have found evidence for variability in encoding precision. The idea is that the encoding precision – the inverse of the variance of the sensory noise – is itself a random variable. Variable-precision models have been used to model visual short-term memory (Devkar & Wright, 2015; Fougnie, Suchow, & Alvarez, 2012; Keshvari et al., 2012, 2013; Salahub & Emrich, 2016; van den Berg, Awh, & Ma, 2014; van den Berg, Shin, Chou, George, & Ma, 2012) and visual attention data (Bhardwaj, Van Den Berg, Ma, & Josic, 2016; Mazyar, van den Berg, & Ma, 2012; Mazyar, Berg, & Seilheimer, 2013). A related concept appears in the beta-binomial model for the psychometric curve (Schütt, Harmeling, Macke, & Wichmann, 2016), where an extra parameter is used to capture variability in the probability of a binary response. At the neural level, variable precision has a parallel in double stochasticity in neural spike counts (Churchland et al., 2011; Goris, Movshon, & Simoncelli, 2014).

Variability in precision could be driven by variations in the stimulus itself or be independent of the stimulus. Stimulus-dependent variable precision is also called heteroskedasticity (unequal variances across the stimulus space). For example, cardinal orientations (horizontal or vertical) are encoded with higher precision than oblique orientations, which is also called the "oblique effect" (Appelle, 1972; Girshick, Landy, & Simoncelli, 2011; Pratte, Park, Rademaker, & Tong, 2017). Heteroskedasticity has also been characterized in color perception and color visual short-term memory (Bae, Olkkonen, Allred, & Flombaum, 2015; Bae, Olkkonen, Allred, Wilson, & Flombaum, 2014).

Stimulus-dependent variable precision could be due to the non-uniform distribution of the preferred stimuli of visual cortical neurons (Li, Peterson, & Freeman, 2003), which in turn could be related to efficient coding (Ganguli & Simoncelli, 2014; Wei & Stocker, 2015). Stimulus-independent contributions to variability in precision could be due to fluctuations in attention (Adam, Mance, Fukuda, & Vogel, 2015; Cohen & Maunsell, 2009; Luck, Chelazzi, Hillyard, & Desimone, 1997) or stochastic memory decay (Fougnie et al., 2012). Although the interpretations of the two factors that affect precision variability are quite different, only one paper has attempted to separate them (Pratte et al., 2017). However, this was in the realm of visual short-term memory, not perception. In color perception and visual working memory, stimulus-dependent variable precision can mimic stimulus-independent variable precision (Bae et al., 2014), but the factors have not been disentangled. Moreover, to our knowledge, no studies have tried to distinguish either form of variability in precision from guessing and decision noise.

2) We examine the importance of different behavioral variability factors in a factorial way. Different from Drugowitsch et al. (2016), in which only models with single factors are tested and compared, we test models with all combinations of behavioral variability factors, and also introduce three metrics to evaluate the importance of each factor: factor knock-in, factor knock-out, factor posterior probability. These approaches hopefully provide a more comprehensive characterization of the importance of each behavioral variability factor.

3) We examine task dependency of the importance of different behavioral variability factors. Most of the studies trying to identify different sources of variability focused on one or two tasks, but the importance of different behavioral variability factors might vary depending on the stimulus context and features of the task. Here we vary the experimental design systematically and try to link the importance of different behavioral variability factors to explain the data to the features in stimulus context and task type. We analyzed data of 11 perceptual experiments (8 new and 3 previously published by our lab), and tested the following behavioral variability factors in all these experiments: stimulus-dependent variable precision, stimulus-independent variable precision, decision noise and lapses (guessing).

TASKS

We conducted eight new target discrimination (categorization) experiments to distinguish the possible factors that might account for behavioral variability, and analyzed the results of three

previously published experiments (Table 1). The previously published experiments are numbered Experiment 7 (was Experiment 1 in Shen & Ma, 2016), Experiment 8 (was Experiment 2 in Mazyar et al., 2013), and Experiment 11 (was Experiment 1 in Mazyar et al., 2013).

All experiments were identical in the following aspects:

- Stimuli were Gabors, with orientation the only relevant feature.
- Presentation times was brief (50 or 83 ms).
- None required visual short-term memory: there was little or no delay between the stimulus display and the response.
- Subjects fixated and all stimuli were presented at the same eccentricity (5° of visual angle).
- The task was a binary choice.
- There were no intertrial dependencies.

The experiments differed in:

- task type (discrimination versus detection),
- set size,
- set size context (a single set size or multiple set sizes in the experiment),
- distractor context (homogeneous or heterogeneous).

Apparatus and stimuli. Subjects were seated at a viewing distance of approximately 60 cm. All stimuli were displayed on a 21-inch LCD monitor with a refresh rate of 60 Hz and a resolution of 1280×1024 pixels. The stimulus displays were composed of Gabor patches shown on a grey background. In Experiments 1-7, 9, and 10, background luminance was 29.3 cd/m^2, and the Gabors had the following settings: peak luminance 35.2 cd/m^2, spatial frequency 3.1 cycles per degree, standard deviation of the Gaussian envelope 8.2 pixels, phase 0 for the cosine pattern. Settings were different in Experiments 8 and 11 (see Mazyar et al., 2013), background luminance was 33.1 cd/m and the Gabors had the following settings: peak luminance 122 cd/m^2, spatial frequency 1.6 cycles per degree, standard deviation of the Gaussian envelope 10 pixels, phase 0 for the cosine pattern.

Experimental procedure. Each trial started with a fixation dot on a blank screen (500 ms), followed by a stimulus display (50 ms in Experiments 1-7, 9, and 10; 83 ms in Experiments 8 and 11). Then, a blank screen was shown until the subject made a response by pressing a button. Response time was not limited. Experiments 1-7, 9, and 10 were visual discrimination tasks. In these experiments, except in Experiment 2, the subject reported whether the target stimulus/stimuli was/were tilted to the left or to the right relative to vertical. In Experiment 2, the subject reported whether the target stimulus was tilted clockwise or counterclockwise relative to a simultaneously presented, randomly drawn reference. Experiments 8 and 11, the subject reported whether a vertical target was present or not. After the response, correctness feedback was given through a change of color of the fixation dot (green for correct, red for incorrect, 500

ms; Figs. 3-13A). The nature of the distractors differed across experiments and is described in detail in Results (Figs. 3-13A and B).

Experiments 1-7, 9, and 10 each consisted of three sessions on different days. Each session consisted of five blocks, and each block contained 200 trials, for a total of 3×5×200=3000 trials per subject. In Experiments 8 and 11, each subject completed 1400 trials (for detailed session and block information, see Mazyar et al., 2013).

Experiment	Number of subjects	Number of stimuli	Number of targets	Task	Distractors
1	6	1	1	Target discrimination relative to vertical	None
2	5	2	1	Target discrimination relative to reference (stimulus on the right)	None
3	6	4	all	Target discrimination relative to vertical	None
4	6	1, 2, 4, 8	all	Target discrimination relative to vertical	None
5	6	4	1	Target discrimination relative to vertical	Vertical
6	6	1, 2, 3, 4	1	Target discrimination relative to vertical	Vertical
7	10	4	1	Target discrimination relative to vertical	Homogeneous, variable
8	13	1, 2, 4, 8	0, 1	Target detection, vertical target	Homogeneous, variable
9	6	4	1	Target discrimination relative to vertical	Heterogeneous, variable
10	11	1, 2, 4, 8	1	Target discrimination relative to vertical	Heterogeneous, variable
11	6	1, 2, 4, 8	0, 1	Target detection, vertical target	Heterogeneous, variable

Table 1: Overview of experiments. For distractors, we use “homogeneous” and “heterogeneous” to indicate that the distractors were identical to (or different from, respectively) each other within a display; we use “variable” to indicate variability across trials. Experiment 7 was previously published as Experiment 1 in Shen & Ma (2016). Experiments 8 and 11 were previously published as Experiments 2 and 1 in Mazyar et al. (2013), respectively.

THEORY

We build Bayesian models for the data (Fig. 1A). A Bayesian model consists of three main steps, and different contributors to behavioral variability appear in each step:

1. The *generative model*, which is a statistical description of both the noisy internal measurements of the stimuli and of what the observer believes about how the stimuli were generated (which may or may not be how they were actually generated). The part of

the generative model describes the noisy measurements. We consider two kinds of variability in the precision of measurement noise in this step: orientation-dependent variable precision (O) and orientation-independent variable precision (V).

2. The observer's decision model. We assume an optimal decision rule, which produces a decision rule by inverting the generative model. We allow for the outcome of this decision rule to be corrupted by decision noise (D). Systematic suboptimality in the decision rule might also cause variability in behavior (Beck et al., 2012), but such suboptimality will be very task-dependent and therefore we do not consider it as a general model factor. As an example, we explore suboptimal rules for Experiment 7 in Results.
3. Predictions for subject responses on a trial-by-trial basis. We consider the guessing (G) in this step.

We first describe the measurement component of the generative model, since it is shared across all experiments. We then describe the experiment-specific component of the generative model, and finally derive the observer's decision rules.

Step 1a: Generative model: Noisy measurements and variability in precision

We assume that the observer makes a noisy measurement x_i of each physical orientation s_i, where i =1,…, N labels the stimuli in a given display (N is the set size). We denote the vector of physical orientations of the stimuli by **s** and the vector of orientation measurements by **x**. Throughout, we will assume that the measurements are independent given the stimuli,

$$p(\mathbf{x}|\mathbf{s}) = \prod_{i=1}^{N} p(x_i|s_i).$$

We assume that the distribution of x_i given s_i is either Gaussian,

$$p(x_i|s_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i - s_i)^2}{2\sigma_i^2}},$$

or Von Mises (circular Gaussian),

$$p(x_i|s_i) = \frac{1}{2\pi I_0(\kappa_i)} e^{\kappa_i \cos 2(x_i - s_i)}, \quad (1)$$

where I_0 is the modified Bessel function of the first kind of order 0. Noise level or precision is controlled by the standard deviation σ_i (Gaussian) or by the concentration parameter κ_i (Von Mises). The factor 2 in the exponent is present because orientation space is [0, π) instead of [0,

2π). In the limit of large κ_i, the Von Mises distribution converges to the Gaussian distribution, with $\kappa_i = \frac{1}{4\sigma_i^2}$. We use the Gaussian distribution when the range of orientations (Experiments other than Experiment 2) or the orientation difference (Experiment 2) in the experiment was small compared to the full range [0, π) (in Experiments 1 to 8), and Von Mises otherwise (in Experiments 9 to 11).

We consider the variability of the measurement noise, that is, the parameter that characterizes the precision, σ_i or κ_i, is also variable. This variability can be orientation-dependent or orientation-independent.

The main characteristic of orientation-dependent precision variability is that encoding noise is less for cardinal (horizontal or vertical) orientations (Appelle, 1972; Girshick et al., 2011). For Gaussian noise, the standard deviation of the noise can be modeled as a rectified sin function of the stimulus orientation:

$$\sigma_i = \sigma_0\left(1+\beta\left|\sin\left(2s_i\right)\right|\right),$$

where σ_0 is the baseline noise level and β is the amplitude parameter of orientation dependence. σ_i is fixed when β is equal to 0. Therefore, we obtain for precision J_i (Fig. 1B):

$$\begin{aligned} J_i &= \frac{1}{\sigma_0^2\left(1+\beta\left|\sin\left(2s_i\right)\right|\right)^2} \\ &= \frac{J_0}{\left(1+\beta\left|\sin\left(2s_i\right)\right|\right)^2}, \end{aligned} \tag{2}$$

where J_0 is the baseline precision. We use the latter equation also for Von Mises noise.

For orientation-independent precision variability, one successful empirical description of behavior has used Fisher information as a starting point (Cover & Thomas, 2005). Fisher information, denoted by J, is related to the precision parameters through

$$\begin{aligned} J &= \frac{1}{\sigma^2}\ \left(\text{Gaussian}\right) \\ J &= \frac{4\kappa I_1\left(\kappa\right)}{I_0\left(\kappa\right)}\ \left(\text{Von Mises}\right), \end{aligned} \tag{3}$$

where I_1 is the modified Bessel function of the first kind of order. It is then assumed that J follows a Gamma distribution:

$$p(J) = \text{Gamma}\left(J; \frac{\bar{J}}{\tau}, \tau\right), \tag{4}$$

where $\bar{J}$ is the mean precision, and τ is called the scale parameter (Fig. 1C).

In previous work (Keshvari et al., 2012; Mazyar et al., 2012; Mazyar et al., 2013; van den Berg et al., 2012), we did not include the factor of 2 in Eq. (1) and the factor of 4 in Eq. (3), but instead rescaled orientations from [0, π) to [0, 2π) before doing any analysis. This rescaling is mathematically equivalent to inserting those factors, but here, we opted against the rescaling, so that we can compare the results of Gaussian-based analysis to those of Von Mises-based analysis with minimal confusion (e.g. in Fig. 16).

In all experiments, we tested all four combinations of the two factors for precision variability: base model with fixed precision (Base), orientation-dependent variable precision only (O), orientation-independent variable precision only (V), and a combination of orientation-dependent and independent variable precision (OV). For the base model with fixed precision, J_i is the same across all i and across all trials. For the O model, J_i is computed from Eq. (2). For the V model, J_i is drawn independently across i and across trials from a gamma distribution with the mean $\bar{J}$ and the scale parameter τ (Eq. (4)). For the OV model, we first computed $\bar{J}$ from Eq. (2), then draw J from a gamma distribution with $\bar{J}$ and the scale parameter τ.

In experiments with multiple set sizes, we allowed J (fixed-precision models), J_0, or $\bar{J}$ (variable-precision models) to vary with set size; we did not impose a parametric form but fitted the parameter independently at each set size.

Step 1b: Generative model: Experimental statistics

The generative model of an experiment consists not only of the distribution $p(x_i|s_i)$, which we discussed in Step 1a, but also of the experimental statistics. Relevant variables are category (target tilted left or right in the discrimination experiments, target present or absent in the detection experiments), and the stimuli on a given trial. We specify their joint distribution. This distribution is determined by the experimental design, with two exceptions:

- The two categories were always presented with probability 0.5. However, we did not assume that subjects would believe this probability to be exactly 0.5. Instead, we used a free parameter to characterizing the observer's prior probability that the stimulus was tilted right (p_{right}) in the discrimination experiments, or that the stimulus was present (p_{present}) in the detection experiments.
- In Experiments 1, 3, 4, 5, 6, we used discrete stimulus values, e.g. 19 values spaced linearly between -15° and 15° (Experiments 1 and 3), between -5° and 5° (Experiment 4), between -20° and 20° (Experiments 5 and 6). We did not assume that subjects had detailed knowledge of these values, but we instead assumed that the observers believed this distribution was Gaussian with the same mean and standard deviation as the actual distribution. In the cases that we examined, the model predictions only depend weakly on

the assumed stimulus distribution, and our eventual results are similar between the actual distribution and the Gaussian approximation.

Step 2: Decision model: Bayesian observer and decision noise

The Bayesian observer "inverts" the generative model to obtain a probability distribution over the variable of interest (here category, C=1 or C=-1 given the noisy measurements $\mathbf{x}$ on a given trial. The Bayesian decision variable, denoted by d, is the log of the ratios of the probabilities of C=1 and C=-1 given $\mathbf{x}$:

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

The Bayesian observer reports C=1 if d is positive. The derivations of the Bayesian decision rules for all experiments are given in Appendix 1.

We test models with decision noise (D), where on each trial, the actual decision variable $\tilde{d}$ is drawn from a Gaussian distribution with a mean of d and a standard deviation of σ_d. The observer then instead reports C=1 if $\tilde{d}$ is positive.

Even without decision noise, the Bayesian observer is not strictly optimal, because we made two modifications in Step 1b (for a detailed distinction between the terms Bayesian and optimal, see Ma, 2012). An additional deviation from optimality is hidden in our assumption that in orientation-dependent models, the subject knows the noise, but does not know the relationship between the noise and the orientation and therefore does not infer the orientation from the noise. In practice, this means that we assumed σ to be known in Step 2, and we do not take its s dependence into account when marginalizing over s.

Step 3: Predictions: Sampling of measurements and guessing rate

Step 2 produces a mapping from a set of measurements, $\mathbf{x}$, to an estimate of category, $\hat{C}$. However, we are ultimately interested in the probability that on a given trial, the observer will make either category response, that is, $p(\hat{C}|\mathbf{s})$, where $\mathbf{s}$ are the physical stimuli on that trial. This distribution is obtained as an average (marginalization) over measurement vectors $\mathbf{x}$:

$$p(\hat{C}|\mathbf{s}) = \int p(\hat{C}|\mathbf{x})\, p(\mathbf{x}|\mathbf{s})\, d\mathbf{x}. \quad (5)$$

Here, $p(\hat{C}|\mathbf{x})$ is deterministic and given by Step 2, and $p(\mathbf{x}|\mathbf{s})$ is given by the measurement distributions in Step 1a. To approximate this integral, we sampled, for each trial in the experiment, a large number of measurement vectors $\mathbf{x}$ based on the physical stimuli $\mathbf{s}$ on that trial. For each $\mathbf{x}$, we applied the decision rule from Step 2, and counted the outcomes. The

proportions of either category response serve as our approximation of $p(\hat{C}|\mathbf{s})$. The number of samples of **x** needs to be sufficiently large for the approximation to be good. Based on an earlier test that showed convergence near 256 samples in a similar task (van den Berg et al., 2012, Appendix), we chose 2000 samples.

We allowed for the possibility that the subject guesses on some proportion of trials. To this end, we introduced a guessing rate λ, so that the probability of reporting $\hat{C}$ given **s** becomes

$$p_{\text{with lapse}}(\hat{C}|\mathbf{s}) = 0.5\lambda + (1-\lambda)\, p(\hat{C}|\mathbf{s}). \quad (6)$$

Thus, in each experiment, we tested a total of 16 models, in a factorial manner (van den Berg et al., 2014). We will denote the factors by G, D, O, and V (see Table 2).

Notation	**Added factor**
Base	None (Base model: no guessing, no decision noise, fixed precision)
G	Guessing (lapse rate)
D	Decision noise
O	Orientation-dependent precision
V	Orientation-independent variable precision
Full	All (Full model: contains all factors)

Table 2: Notation for the model factors that we consider.

MODELING METHODS

Model fitting

We fitted each model to each individual subject's data using maximum-likelihood estimation. The log likelihood of a given parameter combination is the logarithm of the probability of all of the subject's responses given the model and each parameter combination:

$$\begin{aligned}\log L_M(\text{parameters}) &\equiv \log p(\text{data}|M,\ \text{parameters}) \\ &= \log \prod_{j=1}^{N_{\text{trials}}} p(\hat{C}_j|\mathbf{s}_j, M,\ \text{parameters}) \\ &= \sum_{j=1}^{N_{\text{trials}}} \log p(\hat{C}_j|\mathbf{s}_j, M,\ \text{parameters})\end{aligned}$$

where j is the trial index, N_{trials} is the number of trials, $\mathbf{s}_j$ is the set of orientations presented on the j^{th} trial, $\hat{C}_j$ is the subject's response on the j^{th} trial, and we have assumed that there are no sequential dependencies between trials. The probability of the subject response,

$p\left(\hat{C}_j \middle| \mathbf{s}_j, M, \text{ parameters}\right)$, is obtained from Eqs. (5) or (6). To find the values of parameters that maximize log L_M(parameters), we used a novel Bayesian optimization algorithm named Bayesian Adaptive Direct Search (BADS, Acerbi & Ma, 2017). We initialized BADS with random initial values. After BADS returned a parameter combination, we recomputed the log likelihood 10 times with that combination and took the mean, to reduce sampling noise. We performed this process for 10 different initializations and took the maximum of the log likelihoods as the maximum log likelihood for the model, $\mathrm{LL}_{\max}(M)$.

Model comparison metrics

We used Alkaike Information Criterion (AIC) as a measure of goodness-of-fit that takes into account differences in number of parameters between the models (Akaike, 1974): $\mathrm{AIC}(M) = 2k_M - 2\mathrm{LL}_{\max}(M)$, where k_M is the number of parameters of the given model. Another metric is the Bayesian Information Criterion (BIC; Schwarz, 1978), $\mathrm{BIC}(M) = \log(N_{\text{trials}})k_M - 2\mathrm{LL}_{\max}(M)$. AIC penalizes each parameter by 2 points, while BIC penalizes each parameter by 8.0 points in Experiments 1-7, 9, and 10, and by 7.2 points in Experiment 8 and 11. We use the AIC as the main metric because it has been argued to be a good metric for model selection (Burnham & Anderson, 2002, Chapter 2.2). We also show the summary results using BIC in the Appendix (Fig. A1). Choosing BIC does not change our major conclusions in this study.

Model comparison and quantification of factor importance

We have four model factors, each with two levels, for a total of 16 models. We first compared the goodness-of-fit of each model to the full model, which contains all factors (*all from full*, Fig. 2A).

In addition, we would also like to draw conclusions about the importance of each factor regardless of model. In van den Berg et al. (2014), this was done by calculating the proportion of subjects for whom all models in a given model family are rejected (according to the AIC), as a function of the rejection criterion. This method has two disadvantages: a) it works at the population level and cannot be applied when the number of subjects is small; b) it outputs a curve (function) rather than a number. Therefore, we here introduce three new methods: factor knock-in analysis, factor knock-out analysis, and factor posterior probabilities. We expect these methods to be useful in any study that performs factorial model comparison.

We can use a graphical representation to illustrate our model comparison and factorial analysis methods (Fig. 2). In the graph, each dimension represents a factor and each vertex represents a model. For example, if we consider 3 factors that can each be absent or present, we get 8 models in total, which are represented by the 8 vertices of a cube. In the actual analysis, we have 4 factors and 16 models.

Factor knock-in analysis

In the factor knock-in analysis, we compute by how much the AIC of a model decreases (goodness-of-fit improves) when adding one factor at a time (Fig. 2B). Specially, we estimate the "value" of factors guessing (G), decision noise (D), orientation-dependent variable precision (O), orientation-independent variable precision (V), or both O and V, by comparing AICs of models G, D, O, V, OV with the Base model. Drugowitsch et al. (2016) applied a similar analysis.

Factor knock-out analysis

In the factor knock-out analysis, we quantify the contribution of each factor by comparing the AIC of the full model that contains all four factors (GDOV) with the models that lack a certain factor (DOV, GOV, GDV, or GDO), or both O and V (GD) (Fig. 2C). This method shows how "necessary" a factor is to explain the data.

Factor posterior probabilities

Finally, we quantify the importance of a factor by estimating the posterior probability of its existence given the data, $p(F=1|\text{data})$, which we will refer to as the *factor posterior probability* (FPP). This quantity is the most principled one we can compute, as it reflects most objectively the degree of belief in the factor (Van Horn, 2003), but in practice, additional assumptions are needed. First, the computation involves marginalizing (averaging) over all models that contain the factor F; here, we assume that the 8 models we have per factor value (F=1 or -1) are representative of that model space. Second, we have to assume priors over factor values and models; here, we assumed that both values of each factor and all models for a given factor are a priori equally probable. Third, we approximate the log marginal likelihood of a given model by -0.5 times the AIC of that model (Burnham and Anderson, 2002, Chapter 2.9). Combining the assumptions, we find the *marginal likelihood* of a factor to be:

$$\begin{aligned} p(\text{data}|F=1) &= \sum_M p(\text{data}|M)\, p(M|F=1) \\ &= \frac{1}{2^{N_{F=1}}} \sum_{M:F=1} p(\text{data}|M) \\ &= \frac{1}{2^{N_{F=1}}} \sum_{M:F=1} e^{-0.5\,\text{AIC}(M)}, \end{aligned}$$

where $N_{F=1}$ denotes the number of models that contains factor F. The posterior ratio is the ratio of the marginal likelihoods times the prior ratio, but according to our assumption, the latter is 1. Thus,

$$\frac{p(F=1|\text{data})}{p(F=-1|\text{data})} = \frac{\sum_{M:F=1} e^{-0.5\,\text{AIC}(M)}}{\sum_{M:F=-1} e^{-0.5\,\text{AIC}(M)}}.$$

Then the FPP becomes:

$$p\left(F=1\middle|\text{data}\right)=\frac{\sum_{M:F=1} e^{-0.5\,\mathrm{AIC}(M)}}{\sum_{M:F=1} e^{-0.5\,\mathrm{AIC}(M)}+\sum_{M:F=-1} e^{-0.5\,\mathrm{AIC}(M)}}. \tag{7}$$

We now discuss an important special case. If a factor is completely "neutral", which means that the model containing the factor has the exact same $\mathrm{LL}_{\max}$ as the corresponding model without that factor, the its FPP is:

$$\begin{aligned} p\left(F=1\middle|\text{data}\right) &= \frac{\sum_{M:F=1} e^{\mathrm{LL}_{\max}(M)-k_M}}{\sum_{M:F=1} e^{\mathrm{LL}_{\max}(M)-k_M}+\sum_{M:F=-1} e^{\mathrm{LL}_{\max}(M)-k_M}} \\ &= \frac{e^{-1}}{e^{-1}+1} \\ &= 0.27. \end{aligned} \tag{8}$$

The fact that this is lower than 0.5 is due to the AIC penalty for the extra parameter. If we use -0.5BIC as an approximation of log marginal likelihood, this number would become 0.018 for Experiments 1-7, 9 and 10, and 0.027 for Experiments 8 and 11. Adding a factor would have to afford a substantially better fit for the posterior probability of the existence of the factor to cross 0.5. In theory, the FPP of a factor should always be higher than 0.27, but in practice, it is possible to be slightly lower because of the simulation noise. We mark this baseline as reference in all plots showing the FPPs in Results.

RESULTS

We are interested in how different factors that affect behavioral variability in different perceptual tasks. To answer this question, we tested models varying in four factors: guessing (G), decision noise (D) , orientation-dependent variable precision (O), and orientation-independent variable precision (V). We test all combinations of the above factors, for a total of 16 models.

Below, we describe the details of each of the 11 experiments. Whenever we write "randomly", we mean "randomly from a uniform distribution over the possible values". For the *angular positions* of the stimuli on the screen, we use the positive horizontal axis as 0º, and positive values are counterclockwise (as is convention for polar coordinates). For stimulus *orientations*, we use the vertical orientation as 0º, and positive values are clockwise; this is most natural given our stimulus distributions.

Experiment 1: Single stimulus, four possible locations

The subject reported the tilt with respect to vertical (0°) of a single oriented stimulus (Fig. 3A). On each trial, a single stimulus appeared in one of four angular positions: -135º, -45º, 45º, and 135º. Stimulus orientation was drawn randomly from 19 values equally spaced between -15º and 15º (Fig. 3B).

Models that contain guessing (G), orientation-dependent variable precision (O), orientation-independent variable precision (V), or a combination of these factors fit the data best and have nearly equal AICs amongst each other, with mean differences across subjects being less than 5.3. AICs of the models with none of the factors G, O, or V (Base and D) are higher than those best-fitting models, for example by 59 ± 28 (mean ± s.e.m.), 62 ± 28, relative to GDOV, respectively (Fig. 3C).

Knocking in factors G, O, or V decreases the AIC of the Base model by 63 ± 27, 60 ± 28, and 62 ± 28, respectively, but knocking in decision noise (D), yields no benefits, with a slight increase in AIC of 2.57 ± 0.49 (Fig. 3D). Knocking out any individual factor barely affects the AIC of the "full" (GDOV) model, with mean increases across subjects being less than 1, suggesting that none of the factors is necessary to explain the data (Fig. 3E). The factor posterior probabilities (FPPs) of factors G, D, O, and V are 0.66 ± 0.12 and 0.243 ± 0.011, 0.455 ± 0.085, and 0.49 ± 0.10, respectively, indicating some evidence for G, and little or no evidence for D, O or V (Fig. 3F). Consistent with the FPP of factor G, the G model fits the psychometric curves better than the Base model (Fig. 3G).

We also examine the importance of the combination of factors O and V. Starting with the Base model, knocking in factors O and V together decreases the AIC by 61 ± 28, which is similar to the effect of knocking in factors G, O, or V (Fig. 3D). Knocking out both factors hardly affects the AIC (Fig. 3E). The FPP of both factors is 0.41 ± 0.14, suggesting little or no evidence for the combination of factors O and V (Fig. 3F). The model fits of GD and GDOV models are equally good, consistent with the AIC results and factor importance analyses (Fig. 3G).

In summary, in this simple orientation discrimination task, knocking in any factors G, O or V improves the Base model, but none of the factors is necessary to explain the data. There is substantial evidence for factor G.

Experiment 2: Discrimination of a single target with respect to a variable reference orientation

In Experiment 1, the subject reported the tilt relative to vertical. We then wondered whether making the reference orientation variable would change the importance of the four factors. In Experiment 2, the stimulus display consisted of two stimuli, placed on the horizontal axis left and right to the fixation (Fig. 4A). The stimulus on the right was the reference stimulus, whose orientation s_{ref} was drawn from a uniform distribution over the entire orientation space. The stimulus on the left was the target stimulus, whose orientation was drawn from a Von Mises distribution centered at s_{ref} with a concentration parameter of 10 (Fig. 4B). The subject reported

whether the target was oriented clockwise or counterclockwise with respect to s_{ref}. Experiment 2 was different from Experiment 1 in two aspects: first, the reference orientation in Experiment 2 was variable; second, the stimulus range covered the entire space of orientation. We expect to find evidence for factor O in this experiment according to similar experiments in previous literature (Andrews, 1967; Girshick et al., 2011).

Indeed, models that contain factor O (O, GO, DO, GDO, OV, GOV, DOV, GDOV) have small or moderate differences in AIC amongst each other, with mean differences across subjects being less than 8.0. The AICs of models without factor O (Base, G, D, GD, V, GV, DV, GDV) are higher, for example by 34 ± 11, 35 ± 11, 26 ± 11, 28 ± 10, 27 ± 10, 30 ± 10, 28 ± 10, and 30 ± 10 relative to GDOV, respectively (Fig. 4C).

Knocking in factor O decreases the AIC of the Base model by 33 ± 11, indicating that factor O is very beneficial. Knocking in factors G or V decreases the AIC by 8.3 ± 2.0, or 6.6 ± 1.7, respectively, while knocking in factor D yields no benefits, with a slight increase in AIC of 1.29 ± 0.18 (Fig. 4D). Knocking out factor O from GDOV increases its AIC by 30 ± 10, while knocking out any other factor does not increase the AIC, indicating that factor O is the only factor that is necessary to explain the data (Fig. 4E). The FPPs of factors G, D, O, V are 0.671 ± 0.095, 0.269 ± 0.019, 0.86 ± 0.14, 0.392 ± 0.064, respectively, indicating strong evidence for factor O, some evidence for factor G, and little or no evidence for factors D or V (Fig. 4F). In all three analyses, the combination of factors O and V has similar effects as factor O by itself (Fig. 4D-F).

Although there is strong evidence for factor O or a combination of factors O and V, there are no visible differences in model fits to the psychometric curves – proportion of reporting "clockwise" versus orientation of target relative to the reference – between model O and Base, or between models GDOV and GD (Fig. 4G). However, if we plot the accuracy as a function of the reference orientation, there is a clear "oblique effect" in the subject data (Fig. 4H). That is, the accuracy is higher for cardinal orientations (0° and 90°) than oblique orientations (45° and 135°). These results are similar to the literature (Andrews, 1965, 1967). Consistent with the high FPP of factor O, and of a combination of factors O and V, the models O and GDOV fit very well to the data, while models Base and GD clearly deviate (Fig. 4H).

In summary, adding factor O largely improves the Base model, and factor O is also the only factor that is necessary to explain the data. Consistently, there is strong evidence for factor O, some evidence for factor G, and little or no evidence for factors D or V. These results suggest that when the stimulus orientation covers a wide range of the orientation space, there is evidence for orientation-dependent variable precision.

Experiment 3: Discrimination with all stimuli being targets

In Experiment 1 and 2, there was one target. We wondered how the importance of the four factors would change with a larger number of targets. In Experiment 3, we test this idea using an orientation discrimination task with four identical stimuli shown on each trial; both the angular

positions and the stimulus orientation were the same as in Experiment 1 (Fig. 5A and B). All four stimuli were targets, and the subject reported the tilt of their common orientation.

All results are quite similar to those in Experiment 1. Models that contain factor G, O, V, or a combination of these factors have small or moderate differences in AIC amongst each other, with mean differences across subjects being less than 6.7. The AICs of the models with none of the factors G, O or V (Base and D) are higher than the AICs of the above models, for example by 67 ± 28, 69 ± 28 relative to GDOV, respectively (Fig. 5C). Knocking in factors G, O, or V decreases the AIC of the Base model by 65 ± 23, 67 ± 28, and 68 ± 27, respectively, but knocking in D yields no benefits, with a slight increase in the AIC of 2.17 ± 0.37 (Fig. 5D). Knocking out any individual factor barely affects the AIC, with mean increases in AIC across subjects being less than 0.11, suggesting that none of the factors is necessary to explain the data (Fig. 5E). The FPPs of factors G, D, O, and V are 0.64 ± 0.12, 0.273 ± 0.009, 0.0523 ± 0.092, and 0.511 ± 0.070, respectively, indicating there is some evidence for factor G, little or no evidence for factors D, O, or V (Fig. 5F). Consistent with the FPP of factor G, the G model fits the psychometric curves better than the Base model (Fig. 5G).

Knocking in both factors O and V has similar effect as adding any factors G, O or V (Fig. 5D). Knocking out both factors hardly affects the AIC of the GDOV model (Fig. 5E). The FPP of both factors is 0.51 ± 0.12 (Fig. 5F). Consistently, the model fits of GD and GDOV are almost identical, showing that there is little evidence for variability in precision (Fig. 5G).

In summary, the results of this experiment are quite similar to those of Experiment 1: adding any of the factors G, O or V improves the Base model; none of the factors is necessary to explain the data, although there is substantial evidence for factor G or factor V. These results suggest that having multiple targets does not change the importance of the four factors we test.

Experiment 4: Discrimination with all stimuli being targets and multiple set sizes

We next tested how the importance of different factors change when the set size varied across trials – perhaps, this would make attentional allocation less stable. Experiment 4 had the same paradigm as Experiment 3, but the set size was 1, 2, 4, or 8, drawn randomly on each trial. At set size 8, we used all 8 angular positions. At set sizes 1, 2, and 4, we placed the first stimulus at a random angular position. At set size 2, we placed the second stimulus diametrically opposite to the first. At set size 4, we placed the remaining stimuli at every other position. Stimulus orientation was drawn randomly from 19 values equally spaced between -5º and 5º; we chose this range narrower than in Experiment 2 because we were concerned that the task would otherwise be too easy at set size 8 (Fig. 6A and B).

In each model, we treat precision or mean precision at a given set size as a free parameter; we do the same in later experiments that use multiple set sizes. We assume that all other parameters (prior, guessing rate, orientation dependence of noise, and scale parameter of the gamma distribution) are shared among all set sizes.

All models have small or moderate differences in AIC amongst each other, with mean differences across subjects being less than 8.2 (Fig. 6C). Knocking in factors G, O, or V slightly

decreases the AIC of the Base model by 5.8 ± 2.9, 3.1 ± 2.7, and 3.6 ± 2.2, respectively, and knocking in factor D yields no benefits, with a slight increase in AIC of 2.34 ± 0.54 (Fig. 6D). Knocking out any individual factor barely affects the AIC, with mean increases in AIC across subjects being less than 0.089, suggesting that none of the factors is necessary to explain the data (Fig. 6E). The FPPs of factors G, D, O, and V are 0.61 ± 0.13, 0.239 ± 0.041, 0.370 ± 0.085, and 0.379 ± 0.081, respectively, indicating there is some evidence for factor G, little or no evidence for factors D, O, or V (Fig. 6F). Differences in the model fits to the psychometric curves between Base and G models are hardly visible (Fig. 6G), consistent with the fact that knocking in factor G only slightly decreases the AIC of the Base model. There is no evidence of the combination of factors O and V in any of the three analyses (Fig. 6D-F). Consistently, the model fits of GD and GDOV are almost identical (Fig. 6G).

Different from Experiments 1 and 3, knocking in factor G does not improve the Base model either in AIC (Fig. 6D) or the model fits (Fig. 6G). This might be because the range of stimuli (-5º to 5º) was narrower than in the previous two experiments (-15º to 15º). Therefore, in Experiments 1 and 3, a mistake on an easy trial (tilts close to 15º) is very harmful to the goodness-of-fit of the Base model.

In summary, in this experiment with multiple targets and multiple set sizes, we found no strong evidence for any of the factors.

Experiment 5: Discrimination of a single target with a fixed number of vertical distractors

In Experiments 3 and 4, although multiple stimuli were presented simultaneously, all stimuli were targets. We next examine whether replacing targets by distractors – in other words, adding a visual search component to the discrimination task – changes the importance of different factors. Experiment 5 was an orientation discrimination task (Fig. 7A). Set size was 4 and the angular positions were the same as in Experiment 1. Three of the stimuli were vertical; these were the distractors. The fourth stimulus, whose position was drawn randomly, was the target. Target orientation was drawn randomly from 19 values equally spaced between -20º and 20º (Fig. 7B).

Models that contain factor G, O, V, or a combination of these factors have small or moderate differences in AIC amongst each other, with mean differences across subjects being less than 12. AICs of the models with none of the factors G, O, or V (Base, and D) are much higher than the above models, for example by 44 ± 12, 46 ± 13 relative to GDOV, respectively (Fig. 7C). Knocking in factors G, O, or V decreases the AIC of the Base model by 49 ± 12, 39 ± 10, and 44 ± 11, respectively, but knocking in D yields no benefits, with a slight increase in AIC of 2.2 ± 1.6 (Fig. 7D). Knocking out any individual factor barely affects the AIC of the GDOV model, with mean increases in AIC across subjects being less than 3.0, suggesting that none of the factors is necessary to explain the data (Fig. 7E). The FPPs of factors G, D, O, and V are 0.844 ± 0.074, 0.316 ± 0.035, 0.311 ± 0.022, and 0.374 ± 0.046, respectively, indicating strong evidence for factor G and little or no evidence for other factors (Fig. 7F). Consistent with the high FPP of factor G, the G model fits the psychometric curves better than the Base model (Fig.

7G). Knocking in both factors O and V decreases the AIC of the Base model by 42 ± 11 (Fig. 7D), but knocking out both factors hardly affects the AIC of the GDOV model (Fig. 7E). The FPP of the combination of both factors is 0.215 ± 0.038 (Fig. 7F), indicating no evidence for the combination of factors O and V. Consistently, the model fits of models GD and GDOV are almost identical, showing that there is no evidence for variable precision (Fig. 7G).

The evidence for factor G is higher than that in Experiments 1, 3, and 4 (Fig. 7F). This might be because the stimulus range used in experiment is wider, from -20° to 20°. Therefore there are more easy trials than in previous experiments. Factor G is the best factor to explain the mistakes in those easy trials, and is less replaceable by other factors here than in Experiments 1, 3 and 4.

In summary, in this experiment with vertical distractors, adding any of the factors G, O, or V improves the Base model; although none of the factors is necessary to explain the data, we found strong evidence for factor G, and little or no evidence for the other factors.

Experiment 6: Discrimination of a single target with a variable number of vertical distractors

We next tested how the conjunction of distractors and variable set sizes across trials would change the importance of factors. Experiment 6 was identical to Experiment 5, except for the following differences (Fig. 8A). The set size was 1, 2, 3, or 4, drawn randomly on each trial. Angular positions were drawn randomly. Target orientations were drawn randomly from 19 values equally spaced between -20° and 20° (Fig. 8B).

All models with factor G (G, GD, GO, GDO, GV, GDV, GOV, GDOV) fit the data best, and have very similar AICs amongst each other, with mean differences across subjects being less than 5.2. The AICs of the remaining models (Base, D, O, DO, V, DV, OV, GOV, DOV) are higher than those of the best-fitting models, for example by 65 ± 21, 68 ± 22, 15.4 ± 8.6, 15.6 ± 8.1, 10.2 ± 7.9, 12.3 ± 7.2, 10.3 ± 6.8, and 9.8 ± 5.6, relative to GDOV, respectively (Fig. 8C). Among these models, models with factors O, V, or both are better than the remaining models.

Knocking in factors G, O, or V decreases the AIC of the Base model by 69 ± 21, 50 ± 17, and 55 ± 18, respectively, but knocking in D yields no benefits, with a slight increase in AIC of 2.83 ± 0.85 (Fig. 8D). Knocking out factor G increases the AIC of the GDOV model by 9.8 ± 5.6, while knocking out other factors barely affects the AIC, suggesting that factor G, but not other factors, is necessary to explain the data (Fig. 8E). The FPPs of factors G, D, O, and V are 0.910 ± 0.070, 0.388 ± 0.065, 0.347 ± 0.044, and 0.398 ± 0.096 respectively, indicating strong evidence for factor G and little or no evidence for other factors (Fig. 8F). Consistent with the high FPP of factor G, the G model fits the psychometric curves better than the Base model (Fig. 8G). Knocking in both factors O and V decreases the AIC of the Base model by 55 ± 19 (Fig. 8D), but knocking out both factors hardly affects the AIC of the GDOV model (Fig. 8E). The FPP of the combination of both factors is 0.32 ± 0.12 (Fig. 8F), indicating little evidence for the combination of factors O and V. Consistently, the model fits of models GD and GDOV are almost identical, showing that there is no evidence for variable precision (Fig. 8G).

Similar to Experiment 5, the evidence for factor G is strong in this experiment. Factor G is even necessary to explain the data. This might be because the range of target stimulus -20° to 20° is wide and there are a decent number of easy trials, especially when the set size is small. The mistakes on those easy trials could be well explained by factor G and is not replaceable by other factors.

In summary, we found in this experiment with vertical distractors and multiple set sizes, that adding any of the factors G, O or V improves the Base model, but factor G is the only factor that is necessary to explain the data. There is strong evidence for factor G and little evidence for the other factors.

Interim conclusion from Experiment 1-6:

So far, we found that in the experiment with stimulus orientations covering the entire orientation space (Experiment 2), there is strong evidence for factor O. In all other experiments, we found little or no evidence for factors D, O, or V, but the evidence for factor G is correlated with the proportion of easy trials. In Experiment 4 with few easy trials (stimulus range from -5° to 5°), the evidence for factor G is weak. In Experiments 1 and 3 with wider stimulus range (-15° to 15°) and more easy trials than in Experiment 4, the evidence for factor G is stronger. In Experiments 5 and 6, with even wider stimulus range (-20° to 20°) and more easy trials than Experiments 1 and 3, the evidence for factor G is even stronger. This is intuitive because the strongest evidence for guessing comes from errors on easy trials. Consistently, replacing factor G by O or V produces a worse fit in Experiment 6, but an almost equally good one in Experiments 1, 3, and 5. Taking factor G as a priori more probable than factor O or V, we did not find convincing evidence for factor O or V. We also did not find any convincing evidence for factor D in any of the experiments above.

Experiment 7: Discrimination of a single target with a fixed number of homogeneous distractors

We next tested the importance of the factors in a task where distractors are not just present, but also vary from trial to trial. We used the published data from an experiment in which distractors were identical within a trial but varied across trials (Shen & Ma, 2016; Fig. 9A). The set size was 4 and the angular positions were the same as in Experiment 1. Each stimulus display contained one target and three distractors; target position was drawn randomly. On each trial, the target orientation and the common distractor orientation were drawn independently from the same Gaussian distribution, which had a mean of 0° and a standard deviation of 9.06°. Subjects reported the tilt of the target (the unique stimulus) (Fig. 9A and B).

Different from Experiments 1-6, the GV, GDV, GOV, DOV, and GDOV models fit the data best, with AICs similar amongst each other (mean differences across subjects less than 2.8) (Fig. 9C). Four of these models contain both factors G and V. All other models are worse than the above models, but to varying degrees. The second-best models are models GO, DO, GDO and DV, with AICs higher than those of the besting-fitting models, for example by 11.4 ± 4.7,

16.1 ± 4.3, 12.3 ± 5.2, 11.1 ± 5.4 relative to GDOV, respectively. The next best models are G, GD, V and OV, with AICs higher than those of the best-fitting models, for example by 30 ± 11, 31 ± 11, 39 ± 24, and 39 ± 23 relative to GDOV, respectively. These models contain either factor G or factor V, but not both. The AICs of the remaining models – Base, D and O – are higher than those of the above models, for example by 106 ± 19, 73 ± 11, and 66 ± 23, relative to GDOV, respectively.

Knocking in factors G, O, or V decreases the AIC of the Base model by 76 ± 26, 33 ± 20, 40 ± 11, and 67 ± 11, respectively (Fig. 9D). Knocking in factors G or V causes larger improvements in AIC than knocking in factors O or D. Knocking out factor V increases the AIC of the GDOV model by 12.3 ± 5.2, while knocking out other factors barely affects the AIC, suggesting that factor V, but not other factors, is necessary to explain the data (Fig. 9E). Although factor G is also important in our knock-in analysis, it could be replaced by a combination of factors D, O and V, since the DOV model has an AIC as low as that of the GDOV model. The FPPs of factors G, D, O, and V are 0.59 ± 0.11, 0.44 ± 0.10, 0.327 ± 0.075, and 0.870 ± 0.063 respectively, indicating strong evidence for factor V, some evidence for factor G, and little or no evidence for factors D or O (Fig. 9F).

Knocking in both factors O and V decreases the AIC of the Base model by 55 ± 19, similar to knocking in V only, indicating that adding factor O on top of factor V does not yield more benefits (Fig. 9D). Knocking out both factors O and V increases the AIC of the GDOV model by 31 ± 11, larger than knocking out factor V only (Fig. 9E). This is because the GD model is worse than the GDO model. The FPP of factors O and V is 0.868 ± 0.080, similar to that of factor V only (Fig. 9F).

Consistent with the knock-in analysis, models G, D, O, V fit the psychometric curve better than the Base model. Among these models, G fits the best. A combination of factors G and V provides even better fits than the G model, and is as good as the full model GDOV. Consistent with the strong evidence we found for combination of factors O and V, the GD model fits worse than the full model GDOV (Fig. 9G).

In summary, in this experiment with distractors that vary across trials, we found that adding any of the four factors could improve the Base model, and models with different combinations of factors can fit the data equally well. We found strong evidence for factor V and some evidence for factor G, but V seems to be the only factor that is necessary to explain the data.

Experiment 8: Detection of a single target with a variable number of homogeneous distractors

We next examined the effect of task type on the importance of different factors. We reanalyzed data from a published experiment (Experiment 2 from Mazyar et al., 2013), which used the same stimuli as Experiment 6, but within a detection rather than a discrimination task (Fig. 10A).

The set size was 1, 2, 4, or 8, drawn pseudorandomly on each trial. At set size 8, all angular positions were used. At set sizes 1, 2, and 4, the first stimulus was placed at a random

angular position, and the remaining stimuli were placed at adjacent positions. The target orientation was vertical. Trial type was "target present" or "target absent", drawn pseudorandomly on each trial. On target-absent trials, all stimuli were distractors. On target-present trials, one stimulus was the target stimulus and the remaining stimuli were distractors; the position of the target stimulus was drawn randomly. The common orientation of the distractors was drawn from a Von Mises distribution centered at vertical, with a concentration parameter of 32 (Fig. 10B). The original study (Mazyar et al., 2013) only tested the Base and V models and focused on effects of set size on precision.

The AICs of models with factor V (V, GV, DV, OV, GOV, DOV, GDOV) are the lowest among all models, and almost identical to each other, with mean differences across subjects being less than 4.7. The AICs of other models (Base, G, D, GD, O, GO, DO, GOV) are higher, for example by 14.5 ± 4.1, 12.1 ± 4.6, 14.5 ± 4.1, 13.2 ± 4.2, 5.3 ± 3.1, 8.8 ± 3.7, 11.1 ± 3.4, 8.5 ± 2.9 relative to GDOV, respectively (Fig. 10C).

Knocking in factors O, or V decreases the AIC of the Base model by 9.2 ± 3.3, and 13.3 ± 4.4, respectively, while knocking in factors G or D hardly affects the AIC (Fig. 10D). This result is similar to that of Experiment 7, but knocking in factor G does not affect the AIC. The reason might be similar to what we mentioned in the Interim Conclusion of Experiments 1-6: the concentration parameter in Experiment 7 is 10, and that in Experiment 8 is 32, which means that Experiment 8 is more difficult. In Experiment 7, models without the factor G are penalized heavily on easy trials with incorrect responses. Knocking out factor V increases the AIC of the GDOV model by 8.5 ± 2.9, while knocking out other factors barely affects the AIC, suggesting that factor V, but not other factors, is necessary to explain the data (Fig. 10E). The FPPs of factors G, D, O, and V are 0.378 ± 0.068, 0.367 ± 0.067, 0.466 ± 0.064, and 0.786 ± 0.074 respectively, indicating strong evidence for factor V, and little or no evidence for other factors (Fig. 10F). Consistent with the low evidence for factor G and the high evidence for factor V, the G and Base models fit the psychometric curves similarly to each other, but worse than the V model (Fig. 10G).

Knocking in both factors O and V decreases the AIC of the Base model by 17.9 ± 3.9 (Fig. 10D). Knocking out both factors O and V increases the AIC of the GDOV model by 13.2 ± 4.2 (Fig. 10E). The FPP of both factors O and V is 0.803 ± 0.080 (Fig. 10F). In all three analyses, the evidence for the combination of factors O and V is slightly higher than that of factor V by itself. Consistently, GD and GDOV show clear differences in their fits to the psychometric curves (Fig. 10G).

In summary, we found that adding factors O or V improves the Base model, but factor V is the only factor that is necessary to explain the data. There is strong evidence for factor V and little or no evidence for the other factors.

Interim conclusion from Experiments 7 and 8

Experiments 7 and 8 show that with the same distractor context (homogeneous within a display but variable across trials), there is strong evidence for V in both discrimination and detection

tasks. A consistent explanation of the results of Experiment 1-8 would be that variability of distractors across trials causes orientation-independent variable precision (V). This would be consistent with effects of context or configuration on precision in visual short-term memory (Brady & Alvarez, 2016).

Experiment 9: Discrimination of a single target with a fixed number of heterogeneous distractors

We further examined the importance of the four factors in tasks with a more complex stimulus context. In Experiments 9 to 11, the stimulus display contained heterogeneous distractors variable both across trials and within a single trial. Experiment 9 and 10 were orientation discrimination tasks and Experiment 11 was a detection task.

In Experiment 9, the set size was 4 and the angular positions were the same as in Experiment 1 (Fig. 11A). Each stimulus display contained one target and three distractors; target position was drawn randomly. Target orientation was drawn from a Von Mises distribution with a mean of 0 and a concentration parameter of 10. Distractor orientations were drawn independently from a uniform distribution over the entire orientation space (Fig. 11B). The tasks in this experiment contain ambiguity, meaning that the correct answer was not clear to the subject even when there was no sensory noise; nevertheless, the subjects still did the task with reasonable performances (71.7 ± 1.6%).

Among all models we tested, DO, GDO, DOV and GDOV fit the data the best. Their AICs are similar amongst each other, with mean differences across subjects being less than 5.1. Models O, GO, V, GV, DV, GDV, OV and GOV are slightly worse than those best-fitting models, for example by 7.5 ± 7.7, 9.6 ± 7.9, 13 ± 11, 15 ± 11, 9.3 ± 6.0, 9.4 ± 5.7, 2.8 ± 7.3, and 5.2 ± 7.4 relative to GDOV, respectively. Note that the variability in AIC across subjects is high; this could mean that different subjects follow different models, but we do not have enough subjects to test for that. The other models that contain neither factor O nor factor V (Base, G, D, GD), are worse than the above models, with AICs higher than those of the best-fitting models, for example by 78 ± 26, 69 ± 21, 68 ± 19, and 54 ± 18 relative to GDOV, respectively (Fig. 11C).

Knocking in factors G, D, O, or V decreases the AIC of the Base model by 9.0 ± 6.5, 10 ± 12, 70 ± 21, and 65 ± 18, respectively (Fig. 11D). Knocking in factors O or V has a larger effect than knocking in factors G or D. Knocking out factors G or V barely affects the AIC of GDOV, with decreases of 1.76 ± 0.60 and 2.2 ± 1.3, respectively, indicating that these factors are not necessary to explain the data. Knocking out factor D decreases the AIC of GDOV by 5.2 ± 7.4, which we cannot conclude anything from. Knocking out factor O increases the AIC of GDOV by 9.4 ± 5.8 (Fig. 11E). The FPPs of factors G, D, O and V are 0.260 ± 0.028, 0.54 ± 0.12, 0.64 ± 0.16, and 0.44 ± 0.15, respectively (Fig. 11F), indicating there are some evidence for factors D and O, and little or evidence for factors G and V. Evidence for any of the factors is not very strong, because different combinations of factors could explain the data equally well. The evidence for factor O is intuitive because the distractor orientations cover the entire orientation

space, and are highly relevant to the task. In other words, every item provides important information to the subject about whether it is the target and whether the answer to the trial is "left" or "right".

Knocking in both factors O and V decreases the AIC of the Base model by 75 ± 22, slightly larger than knocking in factor O or factor V only (Fig. 11D). Knocking out both factors O and V from GDOV increases the AIC by 54 ± 18, much larger than knocking out either factor O or factor V only (Fig. 11E). The FPP of factors O and V is 0.85 ± 0.15, higher than the FPP of any of the factors only (Fig. 11F). These results reflect some trade-off between factors O and V, and a strong evidence of the combination of factors O and V.

Consistent with the knock-in analysis, models O, V, and OV fit the psychometric curve better than models Base, G and D. Consistent with the strong evidence we found for combination of factors O and V, the GD model fits worse than the full model GDOV (Fig. 11G).

In this experiment, in contrast to Experiments 1 to 7, the probability of reporting "right" no longer monotonically increases with the target orientation (Fig. 11G). This makes sense because the task is relative easy when the tilt of target stimulus was neither too large nor too small: Obviously, when the target tilt is close to zero, the performance is close to chance. But also, when the target tilt is large, subjects are likely to mistake the target for a distractor, thereby reducing performance.

In summary, in this experiment in which distractors are both heterogeneous within a stimulus display and variable across trials, a number of different combinations of factors could explain the data equally well. We found some evidence for factor O and strong evidence for the combination of factors O and V. Besides, we found higher evidence for D than in previous experiments, suggesting suboptimalities in decision-making in this more complex experiment (Beck et al., 2012).

Experiment 10: Discrimination of a single target with a variable number of heterogeneous distractors

The design of Experiment 10 was identical to Experiment 9, except that the set size was 1, 2, 4, or 8, drawn randomly on each trial. The stimulus placement was the same as in Experiment 4 (Fig. 12 A and B). This experiment combines distractors that are variable both within and across trials with multiple set sizes. Again, this experiment has ambiguity when the set size is greater than 1, and therefore does not allow for perfect performance. Subject accuracy was 72.6 ± 1.7%.

Models that contain factor O, or V, or both fit the data best, and have small or moderate differences in AICs amongst each other, with mean differences in AIC across subjects being less than 8.7. AICs of the models with neither factor O nor factor V (Base, G, D, GD) are higher than those of the best-fitting models, for example by 51 ± 12, 33.5 ± 9.4, 38.9 ± 8.4, and 23.3 ± 6.1, relative to GDOV, respectively (Fig. 12C).

Knocking in factors G, D, O, or V decreases the AIC of the Base model by 17.9 ± 4.3, 12.5 ± 5.1, 44 ± 10, and 51 ± 12, respectively (Fig. 12D). Knocking in factors O or V has a larger effect than knocking in factors G or D. Knocking out any factors G, D, O, or V barely affects the

AIC of the GDOV model, with an increase in AIC being less than 4.4, indicating that these factors are not necessary to explain the data, because a combination of different factors fit the data equally well (Fig. 12E). The FPPs of factors G, D, O and V are 0.392 ± 0.078, 0.47 ± 0.11, 0.69 ± 0.10, and 0.54 ± 0.11, respectively (Fig. 12F), indicating there are some evidence for factor O, and little or no evidence other factors. The strong evidence for factor O is consistent with Experiment 9 and might have the same explanation as in that experiment.

Knocking in both factors O and V decreases the AIC of the Base model by 52 ± 12, slightly larger than knocking in factor O or factor V only (Fig. 12D). Knocking out both factors O and V increases the AIC of the GDOV model by 23.3 ± 6.1, much larger than knocking out either factor O or factor V only (Fig. 12E). The FPP of factors O and V is 0.894 ± 0.089, higher than the FPP of any of the factors only (Fig. 12F). These results reflect some trade-off between factors O and V, and a strong evidence of the combination of factors O and V.

Consistent with the knock-in analysis, models O, V, and OV fit the psychometric curve better than models Base, G and D. Consistent with the strong evidence we found for combination of factors O and V, the GD model fits worse than the full model GDOV (Fig. 12G).

In summary, in this experiment with multiple set sizes and distractors being heterogeneous both within a stimulus display and across display, we found the results to be very similar to the previous experiment: a number of different combinations of factors could explain the data equally well; although none of the factors is necessary to explain the data, there is some evidence for factor O, and strong evidence for the combination of factors O and V.

Experiment 11: Detection of a single target with a variable number of heterogeneous distractors

In this last experiment, we examined the importance of different factors in a detection task with similar distractor context as the previous two experiments. We reanalyzed a target detection task published in a previous paper (Mazyar et al., 2013, Experiment 1). The basic paradigm was the same as in Experiment 8, except that the distractors were heterogeneous (Fig. 13A). Each distractor was independently drawn from a uniform distribution over the entire orientation space, which is the same as in Experiments 9 and 10 (Fig. 13B). This experiment was different from Experiment 9 and 10 not only in the type of task, but also in the absence of ambiguity: the stimulus statistics do not preclude perfect performance.

Models with factor D, factor V, or both fit the data best, and have small or moderate differences in AIC amongst each other, with mean differences across subjects being less than 8.0. The AICs of the remaining models – Base, G, O and GO – are much higher, for example by 67 ± 14, 37.9 ± 9.0, 66 ± 14, and 37.7 ± 8.3 relative to GDOV, respectively (Fig. 13C).

Knocking in factors G, D, O, or V decreases the AIC of the Base model by 28.8 ± 7.3, 60 ± 15, 1.0 ± 3.7, and 64 ± 14, respectively (Fig. 13D). Knocking in factors D or O improves the Base model a lot, having a larger effect than factor G. Knocking in factor O barely improves the Base model. Knocking out factor D increases the AIC of GDOV model by only 6.0 ± 1.8, while knocking out other factors causes even smaller increases in the AIC, indicating that none of the

factors is necessary to explain the data (Fig. 13E). Posterior probabilities of factors G, D, O and V are 0.536 ± 0.098, 0.75 ± 0.10, 0.427 ± 0.093, and 0.48 ± 0.14, respectively, indicating relatively strong evidence for factor D and little or no evidence for factors G, O, or V (Fig. 13F). Consistent with the results of the knock-in analysis, model D and V fit the psychometric curve much better than the Base model (Fig. 13G).

Knocking in (Fig. 13D) or knocking out (Fig. 13E) both factors O and V have similar effect as knocking in or knocking out factor V only, but the FPP of factors O and V is 0.65 ± 0.19, higher than that of factor O or factor V individually (Fig. 13F). These results indicate that although factor D by itself can explain the data well, a combination of factors O and V can explain the data equally well. Consistent with these results, there are no difference between model fits of GDOV and GD, and both of them fit the psychometric curves well (Fig. 13G).

In summary, in this experiment, we found that a number of different combinations of factors could explain the data equally well. Although none of the factors is necessary to explain the data, we found strong evidence for factor D, and some evidence for a combination of factors O and V. Strong evidence for factor D might indicate suboptimality in the decision rule (Beck et al., 2012). To our surprise, models with factor O but without D, V (O, GO) fit the data almost as bad as the Base model, which is very different from what we found in Experiment 9 and 10. One possible explanation is that distractors with large tilts are less relevant to the task in this experiment than in the previous two, so the informative stimuli are weighted less in the decision rule, either because of the nature of the task (optimal decision rule, Appendix 1, Experiment 11) or because subjects only pay attention to stimuli that are close to vertical and ignore the stimuli with large tilts. The latter reason is consistent with previous findings that the oblique effect is weaker when the stimulus is not attended (Kelly & Matthews, 2011; Takács, Sulykos, Czigler, Barkaszi, & Balázs, 2013). Both reasons make factor O hard to detect in this task.

Interim summary from Experiment 9-11

The importance of different factors becomes more complex in these three experiments. The data could be explained equally well by different combinations of parameters. This might be because the stimulus displays are very complex in these experiments, and different factors start to intermingle with each other. We found evidence for factor D in Experiment 9 and 11, indicating potential suboptimalities in these experiments (Beck et al., 2012). But we still found evidence for the combination of factors O and V, suggesting that with heterogeneous distractor contexts that are variable across trials, precision is likely to be variable.

Interpretation of factor importance metrics

We have used three metrics for quantifying the importance of a factor. We observed that they are usually, but not always consistent with each other. Before we discuss the overall results across experiments, we need to reflect on the interpretation of the metrics, especially when they are not consistent with each other.

Inconsistency between factor knock-in and knock-out

While the knock-in and knock-out metrics have a relatively straightforward interpretation by themselves, the question arises what inconsistency between them means.

Case 1: A factor is important in knock-in but not in knock-out. For example, in Experiments 9 to 11, factor V is important in the knock-in analysis, but not in the knock-out analysis. This could be an indication of a "trade-off" between factors: a change in one factor can be compensated by changes in other factors to yield an equally good fit. Such a "trade-off" between factors is an example of model mimicry (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004) and would go away in the limit of infinite data.

Case 2: A factor is important in knock-out but not in knock-in. The opposite is also possible: a factor is important in the knock-out analysis, but not in the knock-in analysis. This could be an indication of an "interaction" between factors: neither factor by itself is sufficient but their combination is, similar to finding an interaction without main effects in ANOVA. This case we have not encountered in our analyses. Factor posterior probabilities would be high for both factors.

Relation between factor posterior probabilities, knock-in, and knock-out

We expect in general that factor posterior probabilities are more closely related to knock-out than to knock-in. This is because FPPs average exponentiated negative half AICs within a model family (with factor or without factor). The exponentiation is similar to a max operation. Thus, the with-factor and without-factor family likelihoods will be dominated by their respective best family members. Starting from Eq. (7),

$$
\begin{aligned}
p\left(F=1\middle|\text{data}\right) &= \frac{\sum_{M:F=1} e^{-0.5\,\mathrm{AIC}(M)}}{\sum_{M:F=1} e^{-0.5\,\mathrm{AIC}(M)} + \sum_{M:F=-1} e^{-0.5\,\mathrm{AIC}(M)}} \\
&\approx \frac{\max_{M:F=1} e^{-0.5\,\mathrm{AIC}(M)}}{\max_{M:F=1} e^{-0.5\,\mathrm{AIC}(M)} + \max_{M:F=-1} e^{-0.5\,\mathrm{AIC}(M)}} \\
&= \frac{1}{1+\dfrac{\max_{M:F=-1} e^{-0.5\,\mathrm{AIC}(M)}}{\max_{M:F=1} e^{-0.5\,\mathrm{AIC}(M)}}} \\
&= \frac{1}{1+\exp\left[-0.5\left(\min_{M:F=-1}\mathrm{AIC}(M) - \min_{M:F=1}\mathrm{AIC}(M)\right)\right]}.
\end{aligned}
\tag{9}
$$

The lowest-AIC family member is usually the most highly parameterized member. When that is the case, the FPP becomes a monotonic function of the factor knock-out AIC difference.

Group effects in FPPs

We computed FPP for each individual subject, and reported the mean and s.e.m of these quantities. Alternatively, we could use sums of AICs across subjects in the computation of FPP. The advantage of this alternative is that we obtain stronger beliefs when there are more subjects. However, this method is based on an underlying assumption that all subjects follow the same model, which is not necessarily true. As a result, the presence of an outlier would dramatically change the FPP of a factor. Consider the FPP in an experiment with 5 subjects. If for all 5 subjects, the models with a factor fit exactly the same as the corresponding models without that factor, we would obtain a FPP equal to $\frac{1}{1+e^{5}} = 0.0067$, which would indicate a very low evidence for that factor. We know from Eq. (9) in the previous section that the FPP is dominated by the best member of the model family. If for one of the subjects, the best model with a factor fits a bit better than the best model without a factor, say by 7 in LL_{max}, then the FPP is approximately $\frac{1}{1+e^{5-7}} = 0.88$, indicating a strong evidence for that factor. A small difference in LL_{max} would potentially cause dramatic difference in the FPP, making it hard to interpret the results. We therefore decided to compute the FPPs for each individual subject and report the mean and s.e.m. With this method, the results are more stable and also comparable across different experiments.

An ideal way of analyzing the importance of a factor would take into account potential heterogeneity in the population: not every subject might be best-fitted by the same model. To account for this, a full hierarchical Bayesian model is available, which returns the probability that a model is the most common one (Stephan, Penny, Daunizeau, Moran, & Friston, 2009). However, this method usually requires large number of subjects, while our subject numbers range from 5 to 13.

Classification of factor posterior probabilities

Many of our conclusions are derived from FPPs. The number by itself already represents the strength of evidence for the factor, but usually an extra step is taken to verbally classify the number based on cut-offs, say 0.2 and 0.8. This would lead to a ternary division:

- FPP > 0.8: evidence for the *presence* of a factor;
- FPP between 0.2 and 0.8: inconclusive;
- FPP < 0.2: evidence for the *absence* of a factor.

When compared with the truth (factor absent or present), this division can lead to four types of "errors":

- *Type I*: factor absent, declared present, "false positive".
- *Type Ii*: factor absent, declared inconclusive;
- *Type II*: factor present, declared absent, "false negative".
- *Type IIi*: factor present, declared inconclusive.

Strictly speaking, our Type Ii and Type IIi errors are not errors, only failures to determine absence or presence, but calling them an error is in line with frequentist statistics; conventional

Type II errors are Type IIi in our classification. This classification of errors will help us explain several caveats in the interpretations of our FPPs.

First caveat: Lack of informative trials leads to inconclusiveness. The first caveat is a simple and general one. Assume for the moment that the calculation of FPPs is exactly correct and that the with-factor model is true. Then the value of the FPP will depend on the number of trials: with 10 trials, the FPP might be 0.53, whereas with 1000 trials, it might be 0.99. Taking this one step further, FPP depends on the number of *informative* trials. For example, easy trials on the ends of the psychometric curve tend to be very informative about the presence of guessing. Thus, the choice of stimuli, e.g. having too few "easy" trials if we are interested in guessing, can prevent us from "detecting" the factor. This would be a Type IIi error: there is an effect but the experiment is declared inconclusive. This problem can only be solved by running a larger number of informative trials. Therefore, if we find task differences in the classified FPP, they could be "real" or due to an inconclusiveness-problem. For example, although we found different evidence for guessing and orientation-dependent variable precision across different experiments, we would not expect them to be task-dependent based on previous lit to these factors. However, it is less clear about whether decision noise and orientation-independent variable precision are task-dependent or not.

Second caveat: trade-offs between factors lead to inconclusiveness or false negatives. A variant of the first caveat arises when considering multiple factors that "trade off" against each other, as discussed above. For example, a nonzero guessing rate could be mimicked by a zero guessing rate and a lower (mean) precision parameter. To illustrate this trade-off, we generate a synthetic data set with the G model for Experiment 4, with a precision of 0.08 deg^{-2} and a guessing rate of 0.02, and compute the log likelihood with different combinations of precision and guessing rate in the G model. Different combinations of precision and guessing rate fit the data equally well, including a precision with 0 guessing rate (Fig. 14A). In such a scenario, the LL_{max} of the with-factor model (G) could be identical to the LL_{max} of the without-factor model (Base) even though the factor is present. In another example, in Experiments 9-11, V might trade off against O and/or D, and the weaker evidence for factor V might be due to stronger evidence for factors O (9 and 10) or D (Experiment 11). To illustrate this scenario, we generate a synthetic data set with the V model for Experiment 9, with a scale parameter $\tau = 0.05$, and compute the log likelihood of different combinations of τ and β of the OV model. A combination of zero β and the true τ fit as well as different combinations of a non-zero β and a smaller τ (Fig. 14B). A smaller fitted τ indicates a weaker evidence for factor V, because the data is partly explained by the factor O. Trade-offs would lead to a Type IIi or a Type II error. This problem is present with any model comparison metric and finite data, but is exacerbated by our use of AIC, since AIC is insensitive to trade-offs between factors (Gelman, Hwang, & Vehtari, 2014). Like the first problem, the second problem would be resolved with infinite data.

Third caveat: an idiosyncracy of AIC. Even if we had infinite data so that the first two problems would not apply, another issue arises from the nature of the penalty term in AIC. We approximated FPP as normalized exp(-AIC/2), where -AIC/2 is the maximum of the parameter

log likelihood, LL_{max}, minus a penalty term equal to the number of parameters. Since each factor is associated with an extra parameter, the relative penalty is 1. A with-factor model cannot have a lower LL_{max} than a without-factor model, because the former has extra flexibility. Therefore, their difference in -AIC/2 can never be smaller than -1, and what we called the factor posterior probability can – apart from simulation noise in LL_{max} – never be smaller than exp(-0.5)=0.27 (Eq. (8)), even if the factor is not present and we have infinitely many trials. Thus, there is a fundamental asymmetry between evidence for factor absence and evidence for factor presence. This puts us in a situation similar to the likelihood ratio test (Casella & Berger, 2002): we cannot convincingly reject the with-factor model. This would lead to a Type Ii error if the cut-off were lower than 0.2: the factor is absent but the experiment is declared inconclusive. This problem would be solved by using log marginal likelihood or sampling-based metrics (such as DIC, WAIC, or LOO-CV) instead of AIC. The associated posterior probability can then be arbitrarily low for the with-factor model (see e.g. Eq. 28.9 in MacKay, 2005 for marginal likelihoods). However, all those models are computationally much more expensive, and marginal likelihood has the additional disadvantage that one has to make assumptions about the priors over parameters.

Relationship between task features and importance of factors

Armed with these caveats on the factor importance metrics, we can now review the importance of the four factors across the 11 experiments.

The experiments differed in the following design features that might affect the importance of different factors in behavioral variability (Table 3): set size greater than 1 (divided attention), set size variability, number of targets greater than 1, task type (discrimination or detection), the distribution of the target orientation, the distribution of the orientation of the reference (Experiment 2) or the distractors (all other experiments), distractor variability across displays, distractor variability within displays, and the presence of ambiguity (in the form of overlapping category distributions).

Experimental design										Factor posterior probability			
Experiment number	Multiple set sizes?	Max set size > 1?	Multi targets?	Task	Target distribution or range	Distractor distribution (Exp 2: reference distribution)	Distractor variability across displays	Distractor variability within display	Ambiguity	Guessing (G)	Decision noise (D)	Orientation-dependent variable precision (O)	Orientation-independent variable precision (V)
1	0	0	0	Dis	[-15, 15]	-	None	-	0	0.66 ± 0.12	0.243 ± 0.011	0.455 ± 0.085	0.49 ± 0.10
2	0	1	0	Dis	$N(s_{ref}, 9.1)$	$U(-90, 90)$	High	Low	0	0.671 ± 0.095	0.269 ± 0.019	0.86 ± 0.14	0.392 ± 0.064
3	0	1	1	Dis	[-15, 15]	-	None	Low	0	0.64 ± 0.12	0.273 ± 0.009	0.523 ± 0.092	0.511 ± 0.070
4	1	1	1	Dis	[-5, 5]	-	None	Low	0	0.61 ± 0.13	0.239 ± 0.041	0.370 ± 0.085	0.379 ± 0.081

5	0	1	0	Dis	[-20, 20]	$\delta(0)$	None	Low	0	0.844 ± 0.074	0.316 ± 0.035	0.311 ± 0.022	0.374 ± 0.046
6	1	1	0	Dis	[-20, 20]	$\delta(0)$	None	Low	0	0.910 ± 0.070	0.388 ± 0.065	0.347 ± 0.044	0.398 ± 0.096
7	0	1	0	Dis	$N(0, 9.1)$	$N(0, 9.1)$	Low	Low	0	0.59 ± 0.11	0.44 ± 0.10	0.327 ± 0.075	0.870 ± 0.063
8	1	1	0	Det	0	$N(0, 5.1)$	Low	Low	0	0.378 ± 0.068	0.367 ± 0.067	0.466 ± 0.064	0.786 ± 0.074
9	0	1	0	Dis	$N(0, 9.1)$	$U(-90, 90)$	High	High	1	0.260 ± 0.028	0.54 ± 0.12	0.64 ± 0.16	0.44 ± 0.15
10	1	1	0	Dis	$N(0, 9.1)$	$U(-90, 90)$	High	High	1	0.392 ± 0.078	0.47 ± 0.11	0.69 ± 0.10	0.54 ± 0.11
11	1	1	0	Det	0	$U(-90, 90)$	High	High	0	0.536 ± 0.098	0.75 ± 0.10	0.427 ±0.093	0.48 ±0.14

Table 3: Features of experiments. Here are the abbreviations and notations. Dis: Discrimation; Det: Detection; ref: reference; s_{ref} : reference orientation; $U(x_1, x_2)$ denotes a continuous uniform distribution on the interval $[x_1, x_2]$; $N(s_0, \sigma)$ denotes Gaussian distribution with a mean of s_0 and a standard deviation of σ; δ(x) denotes Dirac's delta function. The units of all orientations are degrees, and the Von Mises distributions are "converted" to Gaussian distributions to make the comparison across experiments easier.

By examining the importance of factors in all these 11 experiments, we found that some factors are important when certain features are present. We now summarize the importance of each factor we tested and attempt to link to the features of the experiments (Table 3).

Guessing (G)

Guessing, representing stimulus-independent lapses of attention or motor errors, is a factor that has been widely accepted to be present in psychophysical tasks, and it is routinely included in psychometric curve fits (Wichmann & Hill, 2001). Consistent with this, we found in many of our experiments (Experiments 1, 3, 5, 6, 7) that knocking in factor G decreases the AIC of the Base model by more than 50 (Fig. 15A) and an obvious improvement in model fits to the psychometric curves (Figs. 3, 5, 7, 8, 9, panel G). Among these experiments, in Experiments 6, factor G is necessary to explain the data (Fig. 15B) and has a mean FPP of greater than 0.9 (Fig. 15C). This experiment has a relative large number of easy trials, because the target orientation is drawn randomly from 19 values equally spaced between -20º and 20º (Fig. 8B), which is the largest range of the target orientation among all experiments. Also, Experiment 6 contains four set sizes, 1, 2, 4, 8. When the set size is 1 and target orientation close to ±20°, the trials are very easy, where a mistake is only explainable with factor G. In Experiment 5, in which the stimulus range is the same but only with set size equal to 4, factor G is no longer necessary to explain the data, but evidence for factor G is greater than 0.8. For other experiments, it seems that the larger the proportion of "easy" trials, the higher the evidence for factor G. With fewer "easy" trials, models without factor G fit the data equally well as models with factor G, by estimating a lower encoding precision (Fig. 14A). For example, in Experiment 4, where the target orientation range is very narrow (between -5° and 5°), the Base model fits as well as the G model, but the

estimated precision is lower. This comparison might be a "false negative" or "inclusiveness" because of both the lack of informative trials and trade-off between parameters.

Decision noise (D)

Decision noise might reflect random variability or systematic suboptimality (Beck et al., 2012). In most of our experiments (Experiments 1-10), we found little or no evidence for factor D (Fig. 15C), suggesting that human subjects are close to optimal. This is consistent with the conclusion of our previous paper (Shen & Ma, 2016), where we compared many suboptimal decision rules with the optimal rule in an orientation discrimination task (Experiment 7 in this paper) and found that the more similar a suboptimal rule to the optimal rule, the better it fits the data. However, we find more evidence for factor D in Experiment 11 (Fig. 15C), suggesting more random variability or greater suboptimality in this experiment. It is not clear why, but the combination of a detection task and the heterogeneity of distractors might provide a more complex stimulus contexture and therefore induce some subtoptimality. One possible explanation is that in this experiment where the large tilted stimuli are less relevant to the task, the subject pay less attention to those large tilted stimuli.

Orientation-dependent variable precision (O)

Orientation-dependent variable precision seems to be an intrinsic property of neural populations in early visual areas. Physiological studies have shown that there are more neurons in primary visual cortex that are tuned to cardinal orientations than oblique orientations in cats (Bauer & Jordan, 1993; Kalia & Whitteridge, 1973; Li et al., 2003; Payne & Berman, 1983) and monkeys (De Valois, William Yund, & Hepler, 1982; Mansfield & Ronner, 1978). This distribution matches the stimulus statistics of natural environments (Attneave, 1954; Barlow, 1961; Girshick et al., 2011) and supports the theory of efficient coding (Ganguli & Simoncelli, 2014; Wei & Stocker, 2015). Therefore, O should be a factor that is commonly present in perception, but it is easier to be detected when the stimulus distribution covers a larger orientation range. Indeed, we found strong evidence for factor O in Experiments 2, 9 and 10 (Fig. 15C), where the stimulus distribution covers the entire orientation space (Figs. 4, 11, 12, panel B). The low evidence found in experiments with narrow orientation range (Experiments 1, 3-8) is another case of "false negative" or "inconclusiveness" because of the lack of informative trials. In Experiment 11, however, although the distractor stimulus also covers the entire space, the evidence for factor O is weak, which might be explained by two reasons. First, stimuli with large tilts are informative of factor O, but these stimuli are weighted less even in the optimal decision rule (Appendix 1, Experiment 11), therefore making factor O harder to detect. Second, because the stimuli with large tilts and their precisions are less relevant to the task, subject pay less attention to the large tilted stimuli. This is a kind of suboptimality and is reflected in the high evidence for factor D we found in this experiment. The weak evidence for factor O is consistent with previous findings that the "oblique effect" is weaker when the visual stimuli are unattended (Kelly & Matthews, 2011; Takács et al., 2013). Both scenarios lead to a case of "false negative" or

"inconclusiveness" resulting from the lack of informative stimuli, because the informative large tilts are not used in the decision rule.

Orientation-independent variable precision (V)

As noted before (van den Berg et al., 2012), sources of orientation-independent variable precision could include fluctuations in attention (Adam et al., 2015; Cohen & Maunsell, 2009; Luck et al., 1997) and stochastic memory decay (Fougnie et al., 2012). Therefore, in a perceptual task without memory component, factor V is more likely to be detected when the task is more likely to induce attentional fluctuations. To induce allocation of attentional resource, we introduced multiple targets (Experiment 3), multiple targets with multiple set sizes (Experiment 4), vertical distractors (Experiments 5 and 6), homogeneous distractors variable across trials (Experiments 7 and 8), and heterogeneous distractors variable across trials (Experiments 9, 10, and 11). Only in Experiments 7 and 8 did we find strong evidence for factor V (Fig. 15C, Table 3). These results suggest that variability of distractors across the trials might be necessary to induce detectable orientation-independent variable precision, regardless of task type. We expected that heterogeneous variable distractors (Experiments 9, 10, and 11) would induce more attentional fluctuations than homogeneous variable distractors (Experiments 7 and 8), but we found weaker evidence for factor V. One possible explanation is that the factor V is harder to detect in the presence of high orientation-dependent precision variability (Experiments 9 and 10) or high decision noise (Experiment 11). This is a case of "inconclusiveness" because of the trade-off between factors V and O (Experiments 9 and 10), or between factors V and D (Experiment 11).

Relationship between mean precision and set size

Experiments 4, 6, 8, 10, and 11 used multiple set sizes, allowing us to explore the effects of task on the relationship between mean precision and set size. Mean precision decreases strongly with set size in Experiments 8, 10 and 11 (significant effect of set size: repeated-measures ANOVA, $p < 0.05$), where the distractors were variable across trials. There was no significant effect of set size in Experiment 6 (repeated-measures ANOVA, $F(3, 6) = 1.1$, $p = 0.38$), where the distractors were fixed at vertical (Fig. 16). There are no obvious differences between detection (Experiments 8 and 11) and discrimination (Experiment 10). In Experiment 4, all stimuli were targets but with an orientation that was unpredictable across trials. Here, we found that mean precision also decreases with set size (Fig. 16A, significant effect of set size: repeated-measures ANOVA, $F(3, 6) = 4.18$, $p = 0.013$).

Experiments 8 and 11 were from Mazyar et al. (2013) (Experiment 2 and Experiment 1, respectively), and even though there were minor differences between the models, the relationship between mean precision and set size was very similar as in the original paper. An earlier paper (Mazyar et al., 2012) considered one more visual search condition. When the distractors were fixed at 5°, mean precision was constant across different set sizes. Based on the results of both studies, the latter paper hypothesized that mean precision decreases with set size if the

distractors are unpredictable across trials. The results from Experiments 6 and 10 are broadly consistent with this conclusion. However, the design of Experiment 4 was not covered by this hypothesis: there were no distractors but yet we found a significant effect of set size. A unifying hypothesis could be that the less predictable the *entire stimulus display* is across trials, the stronger the decrease of mean precision with set size.

Ultimately, it would be more satisfactory to have a normative explanation: *why* does mean precision decrease with set size to different extents for different stimulus statistics? One recent proposal is that set size effects are due to an optimal trade-off between behavioral performance and the neural costs associated with stimulus encoding (van den Berg & Ma, 2017). Greater predictability might allow for more efficient neural coding, which would lead to savings in neural cost, and that in turn to a weaker set size effect.

Suboptimal decision rules

We performed factorial analysis on four factors and ended up with 16 models in total. However, there are still a large number of models we did not cover. For example, a subject may have used a suboptimal decision rule in performing some of the tasks. An example of a suboptimal rule would be the max rule (Baldassi & Verghese, 2002; Eckstein, 1998; Green & Swets, 1966; Nolte, 1967; Palmer, 1990), in which the subject performs the task only based on the item with the largest tilt. Another case arises when subjects have incorrectly or incompletely learned the class-conditioned stimulus distributions in the experiment, $p(s|C=-1)$ and $p(s|C=1)$, yet perform Bayesian inference under those wrong beliefs. In the paper where Experiment 7 was originally presented (Shen & Ma, 2016), we systematically tested the optimal decision and 24 suboptimal decision rules for one experiment; however, of the four variability factors (G, D, O, V), we only included factor G in that paper. In the present paper, we tested all variability factors, but assumed an optimal decision rule. We found that a combination of variability factors including V fits better than the G model. This slightly but not majorly changes the conclusion of the previous paper; however, we did not test for combinations of suboptimal decision rule with all factors G, D, O, and V. If a model of that type would fit substantially better than our best model found here, it would imply a major change of the conclusion of the previous paper. To start exploring this, we crossed the suboptimal rules from Shen & Ma (2016) with the factor models in this current study, to the extent that the rules themselves did not change (this can in principle be extended). This led to 132 extra models (for a more detailed description, see Appendix 2). This analysis confirms the conclusions from Shen & Ma (2016) and the present paper. a) Simple rules (Class I and Class II) fit the data worse than the optimal decision rules, regardless of the factor model they are crossed with, with mean AIC differences being more than 50. This result confirms the conclusion from Shen & Ma (2016): human behaviors are closer to optimality than to simplicity in this task. b) Among all models tested, the best-fitting models are the combinations of the optimal rule and factors models containing factor V, confirming the strong evidence we found for factor V in the present paper (Fig. 17).

DISCUSSION

We studied the contributions of four factors to behavioral variability: guessing, decision noise, orientation-dependent variable precision, and orientation-independent variable precision. We analyzed data from 11 visual experiments (8 new and 3 previously published) that used very similar oriented stimuli, and performed factorial model comparison and three factor importance metrics. We found that the importance of different factors in explaining the data depends on specific features of the experimental design. We found stronger evidence for guessing in experiments with more easy trials. We found little or no evidence for decision noise in most experiments. We found stronger evidence for orientation-dependent precision when the range of stimulus orientations was wider. Finally, we found little evidence for orientation-independent variable precision, except when distractors were variable across trials. We identified several caveats associated with the limited number of trials, trade-offs between parameters, and an idiosyncracy of AIC, all of which could produce inconclusiveness or false negatives.

Relation to previous work

Relation to work on visual short-term memory (VSTM)

Recent studies that used the variable-precision model in VSTM (Devkar & Wright, 2015; Fougnie et al., 2012; Keshvari et al., 2012, 2013; Salahub & Emrich, 2016; van den Berg et al., 2012) listed, but did not empirically distinguish, different sources of variability in precision; instead, the variability was considered random. One possible source is stimulus-dependent variable precision, as has been found for orientation (Pratte et al., 2017) and color (Bae et al., 2015, 2014). This leaves the question of how much stimulus-independent variability is present.

Here, we separated orientation-dependent variable precision from orientation-independent variable precision. We did not find strong evidence for the latter, except in Experiments 7 and 8. In Experiments 1-6, there are either no distractors (Experiments 1-4) or vertical distractors (Experiments 5-6), which are different from the experimental paradigms in previous VSTM studies. The weak evidence for factor V in Experiments 9, 10, and 11 is probably a case of inconclusiveness due to a trade-off between factors. The fact that we did find evidence for orientation-independent precision variability in Experiments 7 and 8 suggests that memory is not necessary to induce orientation-independent variable precision.

Relation to work on discriminating noise in different stages

Previous work has characterized different kinds of noise in human behavior with various approaches. In contrast detection studies, varying external noise allows one to estimate internal noise (Burgess et al., 1981; Liu et al., 1995; Pelli & Farell, 1999). In Burgess et al., 1981, Pelli & Farewell, 1999, this method is based on a linear relationship between threshold signal energy and noise energy. They then define the intercept to be the “internal noise” and the slope to be the “sampling efficiency”. The “internal noise” roughly corresponds to sensory noise in our

framework, although the noise in the decision stage would also contribute to this measure. "Sampling efficiency" characterizes how close to optimal the decoder is, e.g. how well matched Gabor filters are to the stimulus[1]. Like other forms of suboptimality, low sampling efficiency could cause more variability in human behaviors, and in our study it would be absorbed into decision noise (Beck et al., 2012).

More recently in the Bayesian modeling framework, Drugowitsch et al. (2016) distinguishes sources of suboptimality in an evidence accumulation task. The factors they test include noise in the encoding, inference, and decision stages, as well as deterministic biases. They compared models with noise in each stage with the base model without noise, similar to our knock-in analysis. They found that the model with noise in the inference stage explains the data best. We did not specifically test noise in the inference stage, because in our experiment, it is equivalent to noise in the decision stage, because we only have one inference step. The reason that Drugowitsch et al. (2016) could separate noise in inference and decision because their task requires accumulation of information. Inference noise would be introduced in every step of accumulation, while the decision noise only applies in the last step.

Model proliferation and model identifiability

In van den Berg et al. (2014), we brought up the problem of model proliferation when doing factorial model comparison: with k factors, we would have at least 2^k models (more if a factor has more levels than just absent or present). This problem is also present here, as we tested 16 models. However, it would be exacerbated if we were to cross these 16 models with alternative decision rules. In "suboptimal decision rules" above, we made a start with addressing this issue, but a complete solution is far away: in Experiment 7, testing all combinations of variability models introduced here and decision rules from Shen & Ma (2016) would result in a total of the order of 25×16=400 models (not exactly because the set of suboptimal rules to consider might depend on variability factors). Testing all these models in all experiments would be computationally prohibitive. Moreover, many of these models would in practice be difficult to identify (Acerbi, 2014; Lehmann & Casella, 1998, Definition 1.5.2), a problem that we also encountered in van den Berg et al., 2014, Shen & Ma (2016).

The computational demands and the unidentifiability go beyond these specific examples. Our current view on how to deal with these issues is to keep in mind that very often, one is interested in the evidence for a factor rather than for a specific model. Therefore, a solution could be to further develop techniques for summarizing a factorial model comparison into evidence for factors, as van den Berg et al. (2014) did using the metric of model family comparison and we have done here using knock-in, knock-out, and factor posterior probabilities. It is often much easier to draw conclusions about the importance of a model factor than about the evidence for a specific model. We expect that this toolkit will be expanded and refined as model comparison in psychology becomes more sophisticated.

[1] Confusingly, Liu et al., 1995 also measure the efficiency by varying the external noise, but the efficiency they defined is purely a result from the internal noise.

ACKNOWLEDGMENTS

This work was funded by grant R01 EY020958 from the National Institutes of Health. We thank Luigi Acerbi for sharing his Bayesian Adaptive Direct Search algorithm at an early stage, and for discussions on quantifying the importance of factors. We thank Ronald van den Berg for advice and assistance during experimental design.

REFERENCES

Acerbi, L. (2014). A Framework for Testing Identifiability of Bayesian Models of Perception. In *Advances in Neural Information Processing Systems* (pp. 1026–1034).

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. *bioRxiv*. doi:https://doi.org/10.1101/150052

Adam, K. C. S., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The Contribution of Attentional Lapses to Individual Differences in Visual Working Memory Capacity. *Journal of Cognitive Neuroscience*, *27*(8), 1601–1616. doi:10.1162/jocn_a_00811

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactionson Automatic Control*, *19*(6), 716–723. doi:10.1109/TAC.1974.1100705

Andrews, D. P. (1965). Perception of Contours in the Central Fovea. *Nature*. doi:10.1038/2051218a0

Andrews, D. P. (1967). Perception of contour orientation in the central fovea. II. Spatial integration. *Vision Research*, *7*, 999–1013. doi:10.1016/0042-6989(67)90015-6

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. doi:10.1037/h0033117

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193. doi:10.1037/h0054663

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744–763. doi:10.1037/xge0000076

Bae, G.-Y., Olkkonen, M., Allred, S. R., Wilson, C., & Flombaum, J. I. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision*, *14*(4), 1–23. doi:10.1167/14.4.7.doi

Baldassi, S., & Verghese, P. (2002). Comparing integration rules in visual search. *Journal of Vision*, *2*, 559–570. doi:10.1167/2.8.3

Barlow, H. B. H. (1961). Possible principles underlying the transformation of sensory messages. In *Sensory Communication* (pp. 217–234). doi:10.1080/15459620490885644

Bauer, R., & Jordan, W. (1993). Different anisotropies for texture and grating stimuli in the visual map of cat striate cortex. *Vision Research*, *33*(11), 1447–1450. doi:10.1016/0042-6989(93)90138-M

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron*. doi:10.1016/j.neuron.2012.03.016

Bhardwaj, M., Van Den Berg, R., Ma, W. J., & Josic, K. (2016). Do people take stimulus

correlations into account in visual search? *PLoS ONE*, *11*(3), 1–16. doi:10.1371/journal.pone.0149402

Brady, T. F., & Alvarez, G. A. (2016). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, *15*(2015), 1–69. doi:10.1167/15.15.6.doi

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. (1981). Efficiency of human visual signal discrimination. *Science*, *214*(4516), 93–94. doi:10.1126/science.7280685

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. *Book*. doi:10.1007/978-3-319-02868-2_3

Casella, G., & Berger, R. L. (2002). *Statistical Inference*. *BOOK*. doi:10.1057/pt.2010.23

Churchland, A. K., Kiani, R., Chaudhuri, R., Wang, X.-J., Pouget, A., & Shadlen, M. N. (2011). Variance as a signature of neural computations during decision making. *Neuron*, *69*(4), 818–31. doi:10.1016/j.neuron.2010.12.037

Cohen, M. R., & Maunsell, J. H. R. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, *12*(12), 1594–1600. doi:10.1038/nn.2439

Cover, T. M., & Thomas, J. A. (2005). *Elements of Information Theory*. *Elements of Information Theory*. doi:10.1002/047174882X

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. doi:10.1038/nature04766

De Valois, R. L., William Yund, E., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 531–544. doi:10.1016/0042-6989(82)90112-2

Devkar, D. T., & Wright, A. A. (2015). The same type of visual working memory limitations in humans and monkeys. *Journal of Vision*, *13*(2015), 1–18. doi:10.1167/15.16.13.doi

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, *92*(6), 1–14. doi:10.1016/j.neuron.2016.11.005

Eckstein, M. (1998). The Lower Visual Search Efficiency for Conjunctions Is Due to Noise and not Serial Attentional Processing. *Psychological Science*, *9*(2), 111–118. doi:10.1111/1467-9280.00020

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(april), 292–303. doi:10.1038/nrn2258

Fechner, G. T. (1860). Elemente der Psychophysik. *Elemente Dur Psychophysik*, 572.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229. doi:10.1038/ncomms2237

Ganguli, D., & Simoncelli, E. P. (2014). Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations. *Neural Computation*, *26*(10), 2103–2134. doi:10.1162/NECO_a_00638

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. doi:10.1007/s11222-013-9416-2

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. doi:10.1038/nn.2831

Goris, R. L. T., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability.

Nature Neuroscience, *17*(6), 858–865. doi:10.1038/nn.3711

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. *Society* (Vol. 1). doi:10.1901/jeab.1969.12-475

Kalia, M., & Whitteridge, D. (1973). The visual areas in the splenial sulcus of the cat. *Journal of Physiology*, *232*(2), 275–83. doi:10.1113/jphysiol.1973.sp010269

Kelly, J. G., & Matthews, N. (2011). Attentional oblique effect when judging simultaneity. *Journal of Vision*, *11*(6), 1–15. doi:10.1167/11.6.10

Keshvari, S., van den Berg, R., & Ma, W. J. (2012). Probabilistic computation in human perception under variability in encoding precision. *PLoS ONE*, *7*(6). doi:10.1371/journal.pone.0040216

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No Evidence for an Item Limit in Change Detection. *PLoS Computational Biology*, *9*(2), 15–16. doi:10.1371/journal.pcbi.1002927

Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation , Second Edition Springer Texts in Statistics. Design* (Vol. 41). doi:10.2307/1270597

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique Effect□: A Neural Basis in the Visual Cortex Oblique Effect□: A Neural Basis in the Visual Cortex. *Journal of Neurophysiology*, *90*(February 2003), 204–217. doi:10.1152/jn.00954.2002

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, *35*(4), 549–568. doi:10.1016/0042-6989(94)00150-K

London, M., Roth, A., Beeren, L., Häusser, M., & Latham, P. E. (2010). Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature*, *466*(7302), 123–7. doi:10.1038/nature09086

Luck, S., Chelazzi, L., Hillyard, S., & Desimone, R. (1997). Neural Mechanisms of Spatial Selective Attention in Areas V1 , V2, and V4 of Macaque Visual Cortex. *Journal of Neurophysiology*, *77*(1), 24.

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in Cognitive Sciences*, *16*(10), 511–518. doi:10.1016/j.tics.2012.08.010

MacKay, D. J. C. (2005). *Information Theory, Inference, and Learning Algorithms David J.C. MacKay. Learning* (Vol. 100). doi:10.1198/jasa.2005.s54

Mansfield, R. J. W., & Ronner, S. F. (1978). Orientation anisotropy in monkey visual cortex. *Brain Research*, *149*(1), 229–234. doi:10.1016/0006-8993(78)90603-0

Mardia, K. (1975). *Statistics of Directional Data*. *Journal Of The Royal Statistical Society Series B-Methodological* (Vol. 37). doi:doi: 10.2307/2984782

Mazyar, H., Berg, R. Van Den, & Seilheimer, R. L. (2013). Independence is elusive□: Set size effects on encoding precision in visual search. *Journal of Vision*, *13*(2013), 1–14. doi:10.1167/13.5.8.doi

Mazyar, H., van den Berg, R., & Ma, W. J. (2012). Does precision decrease with set size? *Journal of Vision*, *12*(6), 10–10. doi:10.1167/12.6.10

Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, *86*(1), 82–85. doi:10.1109/JPROC.1998.658762

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: an explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*(3), 465–494. doi:10.3758/PBR.15.3.465

Nolte, L. W. (1967). More on the Detection of One of M Orthogonal Signals. *The Journal of the Acoustical Society of America*, *41*(2), 497. doi:10.1121/1.1910360

Orhan, A. E., & Jacobs, R. R. a. (2013). Are Performance Limitations in Visual Short-Term

Memory Tasks Due to Capacity Limitations or Model Mismatch? *Psychological Review*, *120*(2), 1–62. doi:10.1037/a0031541

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology. Human Perception and Performance*, *16*(2), 332–350. doi:10.1037/0096-1523.16.2.332

Payne, B. R., & Berman, N. (1983). Functional organization of neurons in cat striate cortex: variations in preferred orientation and orientation selectivity with receptive-field type, ocular dominance, and location in visual-field map. *Journal of Neurophysiology*, *49*(4), 1051–1072.

Pelli, D. G., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America a-Optics Image Science and Vision*, *16*(3), 647–653. doi:10.1364/JOSAA.16.000647

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for Stimulus-Specific Variation in Precision Reveals a Discrete Capacity Limit in Visual Working Memory, *43*(1), 6–17. doi:10.1037/xhp0000302

Salahub, C. M., & Emrich, S. M. (2016). Tuning perception: Visual working memory biases the quality of visual awareness. *Psychonomic Bulletin & Review*, *23*(6), 1854–1859. doi:10.3758/s13423-016-1064-z

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123. doi:10.1016/j.visres.2016.02.002

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Shen, S., & Ma, W. J. (2016). A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological Review*, *123*(4), 452–480. doi:10.1037/rev0000028

Soltani, A. (2006). A Biophysically Based Neural Model of Matching Law Behavior: Melioration by Stochastic Synapses. *Journal of Neuroscience*, *26*(14), 3731–3744. doi:10.1523/JNEUROSCI.5159-05.2006

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. doi:10.1016/j.neuroimage.2009.03.025

Takács, E., Sulykos, I., Czigler, I., Barkaszi, I., & Balázs, L. (2013). Oblique effect in visual mismatch negativity. *Frontiers in Human Neuroscience*, *7*(September), 1–13. doi:10.3389/fnhum.2013.00591

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. doi:10.1037/h0070288

Tolhurst, D. J., Movshon, J. A., & Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research*. doi:10.1016/0042-6989(83)90200-6

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003a). Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *20*(7), 1419–1433. doi:10.1364/JOSAA.20.001419

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003b). Statistical decision theory and trade-offs in the control of motor response. *Spat.Vis.*, *16*(3–4), 255–275. doi:10.1163/156856803322467527

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124–149. doi:10.1037/a0035234

van den Berg, R., & Ma, W. J. (2017). A rational theory of the limitations of working memory and attention. *bioRxiv*. doi:http://dx.doi.org/10.1101/151365

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. doi:10.1073/pnas.1117465109

Van Horn, K. S. (2003). Constructing a logic of plausible inference: A guide to Cox's theorem. *International Journal of Approximate Reasoning*. doi:10.1016/S0888-613X(03)00051-3

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*(1), 28–50. doi:10.1016/j.jmp.2003.11.004

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience*, *18*(10), 1509–17. doi:10.1038/nn.4105

Wichmann, F. a, & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. doi:10.3758/BF03194544

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11. doi:10.1167/4.12.11

Wolpert, D. M., & Landy, M. S. (2012). Motor control is decision-making. *Current Opinion in Neurobiology*, *22*(6), 996–1003. doi:10.1016/j.conb.2012.05.003

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233-U13. doi:Doi 10.1038/Nature06860

Appendix 1: Decision rules

Notations:
erf: error function; x_i: internal measurement of the i^{th} item; J_i: encoding precision of the i^{th} item; J_s: precision to generate stimulus orientations; κ_i: concentration parameter of Von Mises distribution; κ_s: concentration parameter of the von Mise distribution to generate stimulus orientations; p_{right}, $p_{\text{clockwise}}$, or p_{present}: prior probability of reporting "right", "clockwise" or "present"; N: set size.

Experiment 1:
The observer reports "right" when

$$d = \log \frac{1+\text{erf}\dfrac{xJ}{\sqrt{2\left(J+J_s\right)}}}{1-\text{erf}\dfrac{xJ}{\sqrt{2\left(J+J_s\right)}}} + \log \frac{p_{\text{right}}}{1-p_{\text{right}}} > 0.$$

Experiment 2:
The observer reports "clockwise" when

$$d = \log \frac{1+\text{erf}\dfrac{\Delta x J_c}{\sqrt{2\left(J_c+J_s\right)}}}{1-\text{erf}\dfrac{\Delta x J_c}{\sqrt{2\left(J_c+J_s\right)}}} + \log \frac{p_{\text{clockwise}}}{1-p_{\text{clockwise}}} > 0,$$

where $J_c = \dfrac{1}{\dfrac{1}{J}+\dfrac{1}{J_R}}$. J_R denotes the encoding precision of the reference, and J denotes the encoding precision of the target. Δx denotes the internal measurement of the target relative to that of the reference.

Experiments 3 and 4
The observer reports "right" when

$$d = \log \frac{1 + \operatorname{erf} \dfrac{\sum_{i=1}^{N} x_i J_i}{\sqrt{2\left(\left(\sum_{i=1}^{N} J_i\right) + J_s\right)}}}{1 - \operatorname{erf} \dfrac{\sum_{i=1}^{N} x_i J_i}{\sqrt{2\left(\left(\sum_{i=1}^{N} J_i\right) + J_s\right)}}} + \log \frac{p_{\text{right}}}{1 - p_{\text{right}}} > 0.$$

Experiments 5 and 6

The observer reports “right” when

$$d = \log \frac{\sum_{i=1}^{N} \left(1 + \operatorname{erf} \dfrac{x_i J_i}{\sqrt{2\left(J_i + J_s\right)}}\right) \exp\left(\dfrac{x_i^2 J_i^2}{2\left(J_i + J_s\right)}\right) \sqrt{\dfrac{J_s}{J_i + J_s}}}{\sum_{i=1}^{N} \left(1 - \operatorname{erf} \dfrac{x_i J_i}{\sqrt{2\left(J_i + J_s\right)}}\right) \exp\left(\dfrac{x_i^2 J_i^2}{2\left(J_i + J_s\right)}\right) \sqrt{\dfrac{J_s}{J_i + J_s}}} + \log \frac{p_{\text{right}}}{1 - p_{\text{right}}} > 0.$$

Experiment 7

The observer reports “right” when

$$d = \log \frac{\sum_{i=1}^{N} \sqrt{\dfrac{1}{J_i + J_s}} \sqrt{\dfrac{1}{\left(\sum_{j \neq i} J_j\right) + J_s}} \left(1 + \operatorname{erf} \dfrac{x_i J_i}{\sqrt{2\left(J_i + J_s\right)}}\right) \exp \dfrac{-x_i^2}{2\left(\dfrac{1}{J_i} + \dfrac{1}{J_s}\right)} \exp\left(\dfrac{1}{2} \dfrac{\left(\sum_{j \neq i} x_j J_j\right)^2}{\left(\left(\sum_{j \neq i} J_j\right) + J_s\right)} - \dfrac{1}{2} \sum_{j \neq i} x_j^2 J_j\right)}{\sum_{i=1}^{N} \sqrt{\dfrac{1}{J_i + J_s}} \sqrt{\dfrac{1}{\left(\sum_{j \neq i} J_j\right) + J_s}} \left(1 - \operatorname{erf} \dfrac{x_i J_i}{\sqrt{2\left(J_i + J_s\right)}}\right) \exp \dfrac{-x_i^2}{2\left(\dfrac{1}{J_i} + \dfrac{1}{J_s}\right)} \exp\left(\dfrac{1}{2} \dfrac{\left(\sum_{j \neq i} x_j J_j\right)^2}{\left(\left(\sum_{j \neq i} J_j\right) + J_s\right)} - \dfrac{1}{2} \sum_{j \neq i} x_j^2 J_j\right)} + \log \frac{p_{\text{right}}}{1 - p_{\text{right}}} > 0.$$

Experiment 8

The subject reports “present” when

$$d = \log \frac{\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{\left(\sum_{j\neq i} J_j\right) + J_s}} \exp\left(\frac{1}{2}\left(\sum_{j\neq i} x_j J_j\right)^2 \frac{1}{\left(\sum_{j\neq i} J_j\right) + J_s}\right)}{\frac{1}{\sqrt{\left(\sum_{i=1}^{N} J_i\right) + J_s}} \exp\left(\frac{1}{2}\left(\sum_{i=1}^{N} x_j J_j\right)^2 \frac{1}{\sqrt{\left(\sum_{i=1}^{N} J_i\right) + J_s}}\right)} + \log \frac{p_{\text{present}}}{1 - p_{\text{present}}} > 0.$$

Experiments 9 and 10

The subject reports "right" when

$$d = \log \frac{\sum_{i=1}^{N} \int_{0}^{\frac{\pi}{2}} \text{VM}\left(2x_i; 2s_{\text{T}}, \kappa_i\right) \text{VM}\left(2s_{\text{T}}; 0, \kappa_{\text{T}}\right) ds_{\text{T}}}{\sum_{i=1}^{N} \int_{-\frac{\pi}{2}}^{0} \text{VM}\left(2x_i; 2s_{\text{T}}, \kappa_i\right) \text{VM}\left(2s_{\text{T}}; 0, \kappa_{\text{T}}\right) ds_{\text{T}}} + \log \frac{p_{\text{right}}}{1 - p_{\text{right}}} > 0,$$

where VM(x; s, κ) denotes a Von Mises distribution with a mean of s and concentration parameter of κ, s_{T} denotes the target orientation, and κ_{T} denotes the concentration parameter of the von Mise distribution to generate target orientations.

Experiment 11

The subject reports "present" when

$$d = \log\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\exp\left(\kappa_i \cos 2x_i\right)}{I_0\left(\kappa_i\right)}\right) + \log \frac{p_{\text{present}}}{1 - p_{\text{present}}} > 0.$$

Appendix 2: Hybrid models of suboptimal rules and factor combinations in Experiment 7

We tested combinations of suboptimal rules and different factors. Among these combinations, Class I suboptimal rules are combined with the total 16 factor models, Class II and Class III models are only combined with models Base, G, D, GD.

Class I suboptimal models do not contain precision in their decision rules, so the decision rule does not change with the addition of factors O or/and V. For models containing factors O, V

or both, we just generate the noisy measurements with variable precision and use the same rule to compute the decision variable. By contrast, Class II and Class III models contain precision in their decision rules, so adding factor O or/and V change the decision rules. Therefore, we did not test the combinations with factors O or/and V.

For all suboptimal decision rules other than the Sign rule in Class I, the decision rule takes the form $d > 0$, where d is the decision variable (Shen & Ma, 2016, Appendix 1). To combine with factor D in the models, a Gaussian noise with standard deviation σ_d is added to d, and is treated as a free parameter in the model fitting.

Adding the prior p_{right} would change the form of the suboptimal rules, so we did not include the prior as a parameter in these hybrid models.

FIGURES

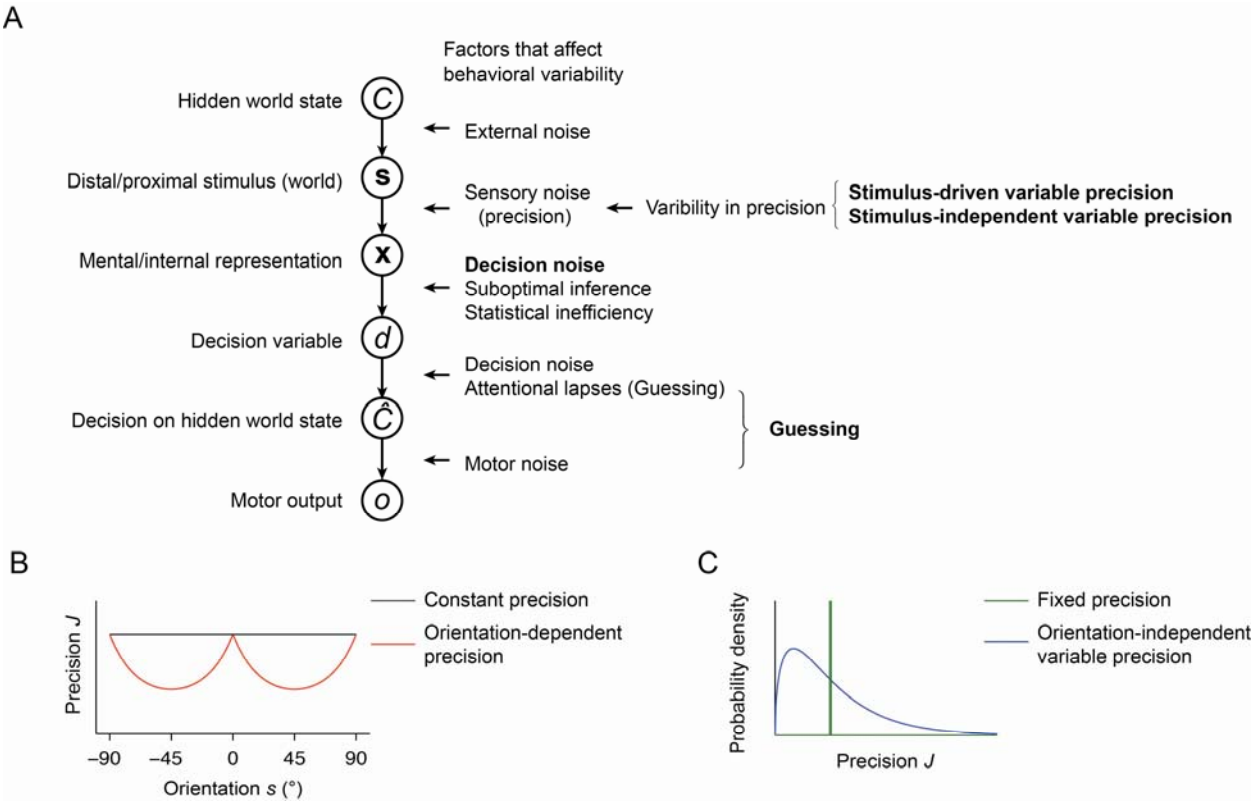

Figure 1. Factors that affect behavioral variability. (A) The generative model of a typical perceptual task is shown on the left. Each node represents a variable and each arrow between two nodes represents a distribution. The right side to generative model shows the possible factors that might affect behavioral variability identified in the field so far. Among these factors, we test the following (marked as bold-faced in the plot) in this study: stimulus (orientation)-dependent variable precision, stimulus (orientation)-independent variable precision, decision noise and guessing. (B) Orientation-dependent precision J is modeled as $J = \frac{J_0}{\left(1+\beta\left|\sin\left(2s\right)\right|\right)^2}$ at orientation s (red line), where J_0 denotes the baseline precision and β denotes the amplitude parameter of the orientation dependence. The black line represents the constant orientation when β equals zero. (C) Probabilistic distribution of orientation-independent precision is modeled as a Gamma distribution $p(J) = \text{Gamma}\left(J; \frac{\overline{J}}{\tau}, \tau\right)$, where $\overline{J}$ denotes the mean precision, and τ denotes the scale parameter that characterizes the variability in the precision. The black line represents the distribution of a fixed precision $\overline{J}$, when τ equals to zero.

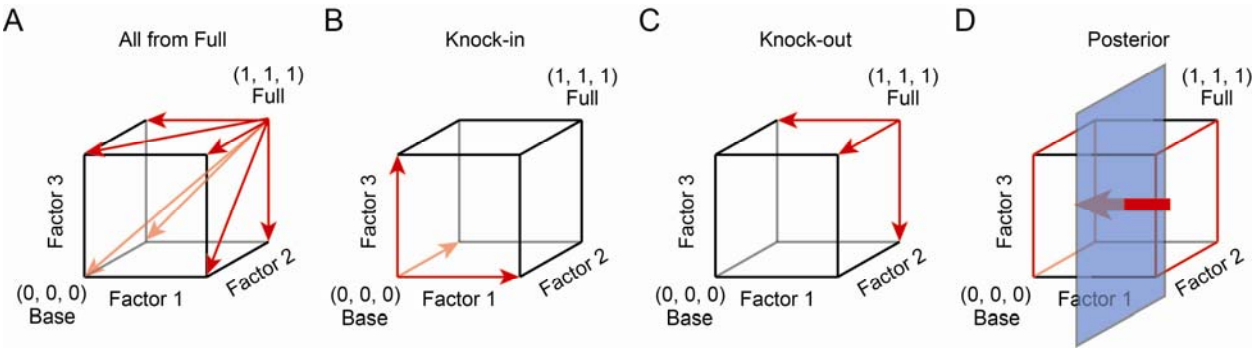

Figure 2. Factor metrics. In each graph, each dimension represents a factor and each vertex represents a model. Here we show an example with 3 factors and get a total of 8 models. The model with none of the factors is (0, 0, 0) and the full model with all factors is (1, 1, 1). (A) All from full (complete model comparison). Goodness-of-fits of all models are compared with that of the full model (1, 1, 1). (B) Factor knock-in. We compute how much improvement in the goodness-of-fit when adding each single factor (red arrows) to the Base model (0, 0, 0). (C) Factor knock-out. We compute how much worse in the goodness-of fit when removing a factor (red arrows) from the full model (1, 1, 1). (D) Factor posterior probability. We compute the posterior probability of the existence of the factor by marginalizing all models containing that factor.

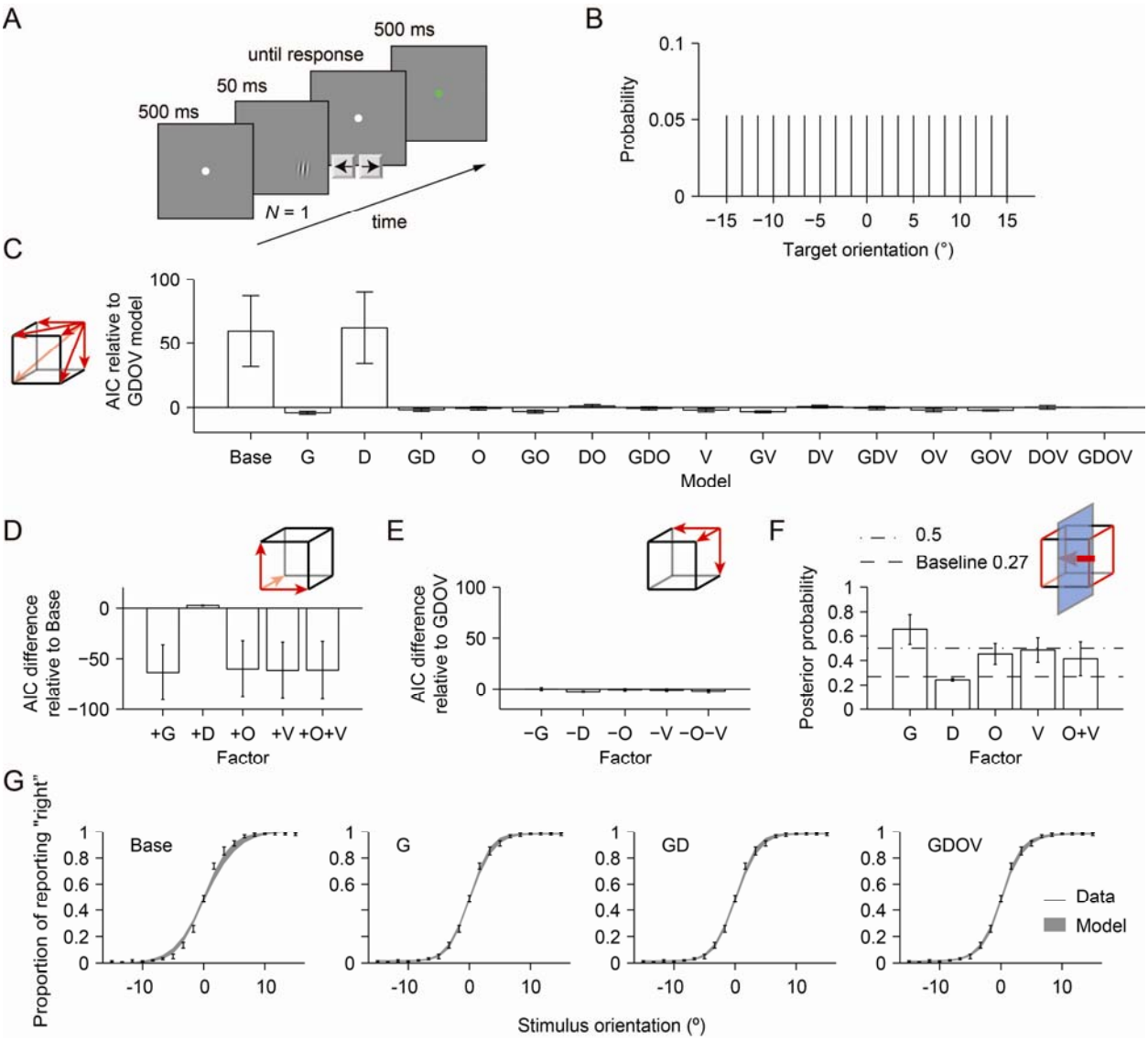

Figure 3. Experiment 1: Single stimulus, four possible locations. (A) Trial procedure. Each trial starts with a fixation dot for 500 ms, then a single stimulus appears for 50 ms in one of the four angular positions: -135º, -45º, 45º, and 135º. The subject reports the tilt of the stimulus with respect to vertical (0°), and a feedback for correctness will be given after the subject response. (B) Stimulus distribution. The stimulus orientation is randomly drawn from 19 values equally spaced between -15º and 15º on each trial. (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting “right” as a function of the stimulus orientation. Solid lines and error bars: data; grey areas: model fits.

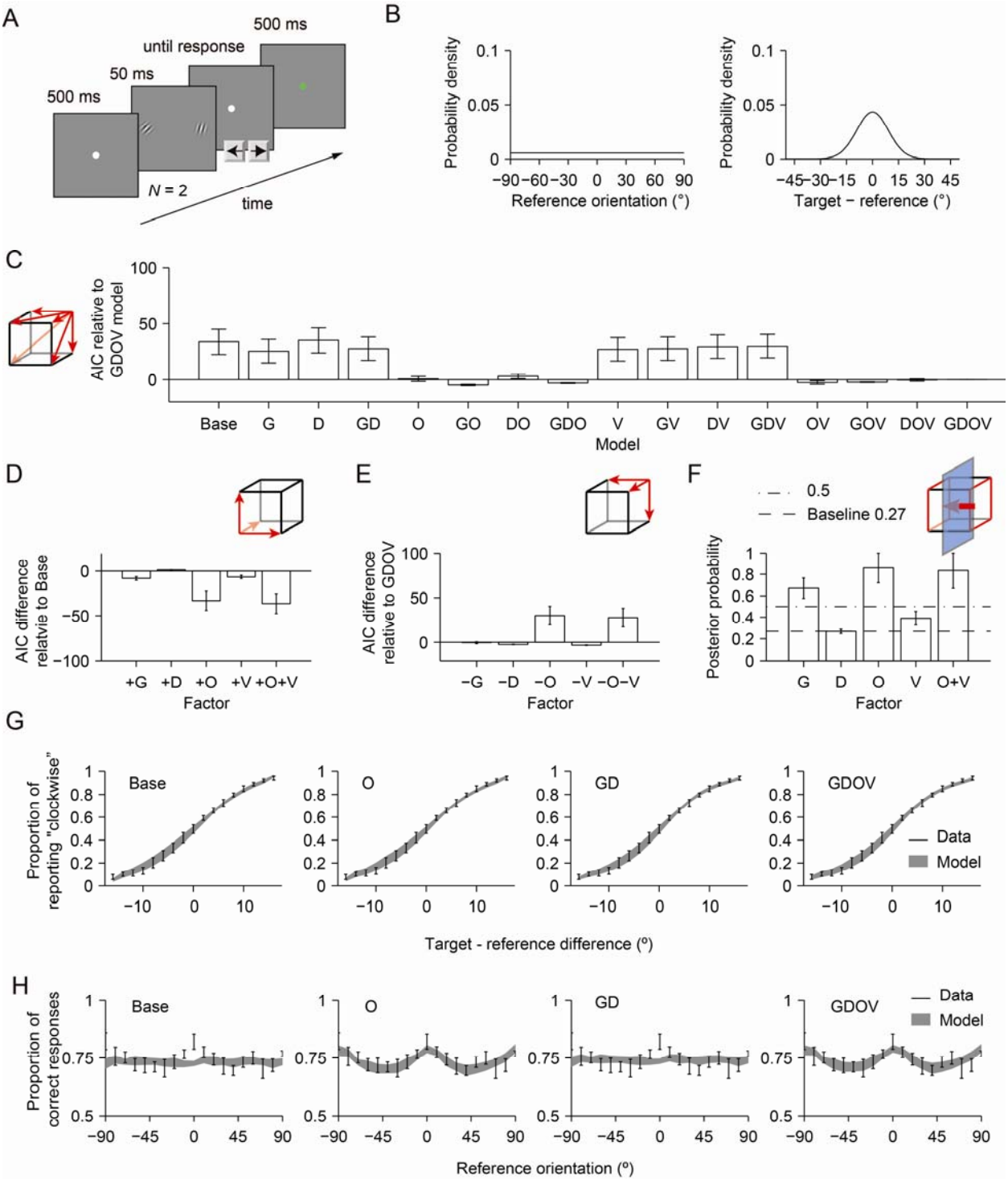

Figure 4. Experiment 2: Discrimination of a single target with respect to a variable reference orientation. (A) Trial procedure. The trial procedure is similar to Experiment 1, but the stimulus display consists of two stimuli, placed on the horizontal axis left and right of fixation. The

stimulus on the right is the reference stimulus, and the stimulus on the left is the target stimulus. The subject reports whether target was oriented clockwise or counterclockwise with respect to the reference orientation. (B) Stimulus distribution. Left panel: the reference orientation is randomly drawn from a uniform distribution over the entire orientation space. Right panel: the target orientation is randomly drawn from a von Mise distribution centered at the reference orientation with a concentration parameter of 10. (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting "clockwise" as a function of the orientation difference between the target and the reference. Solid lines and error bars: data; grey areas: model fits. (H) Proportion of correct responses as a function of the reference orientation. Solid lines and error bars: data; grey areas: model fits.

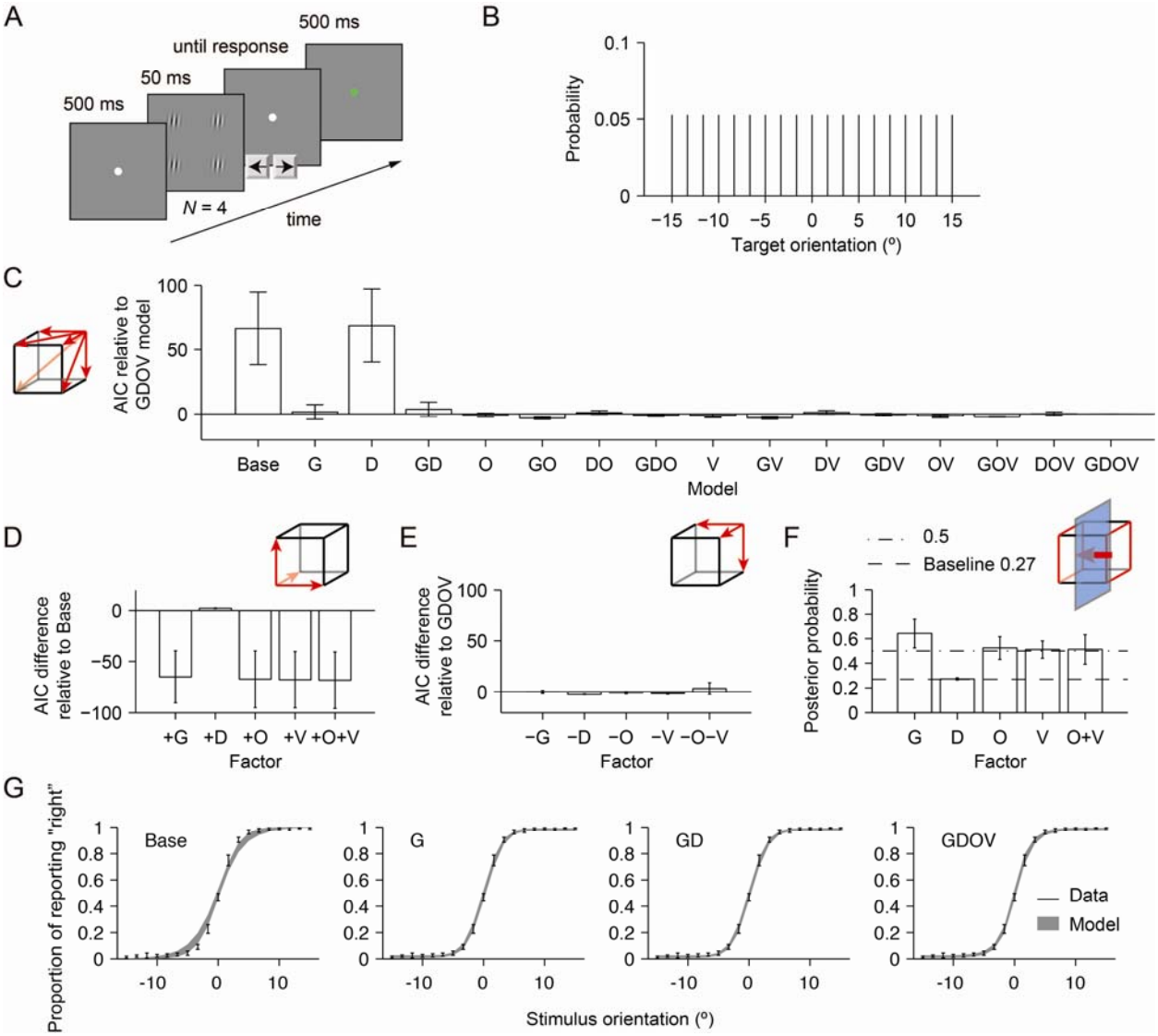

Figure 5. Experiment 3: Discrimination with all stimuli being targets. (A) Trial procedure. The trial procedure is similar to Experiment 1, but the stimulus display consists of four stimuli, all of which are targets. The subject reports the tilt of the common orientation. (B) Stimulus distribution. The target orientation is randomly drawn from 19 values equally spaced between -15° and 15° on each trial. (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting “right” as a function of the target orientation. Solid lines and error bars: data; grey areas: model fits.

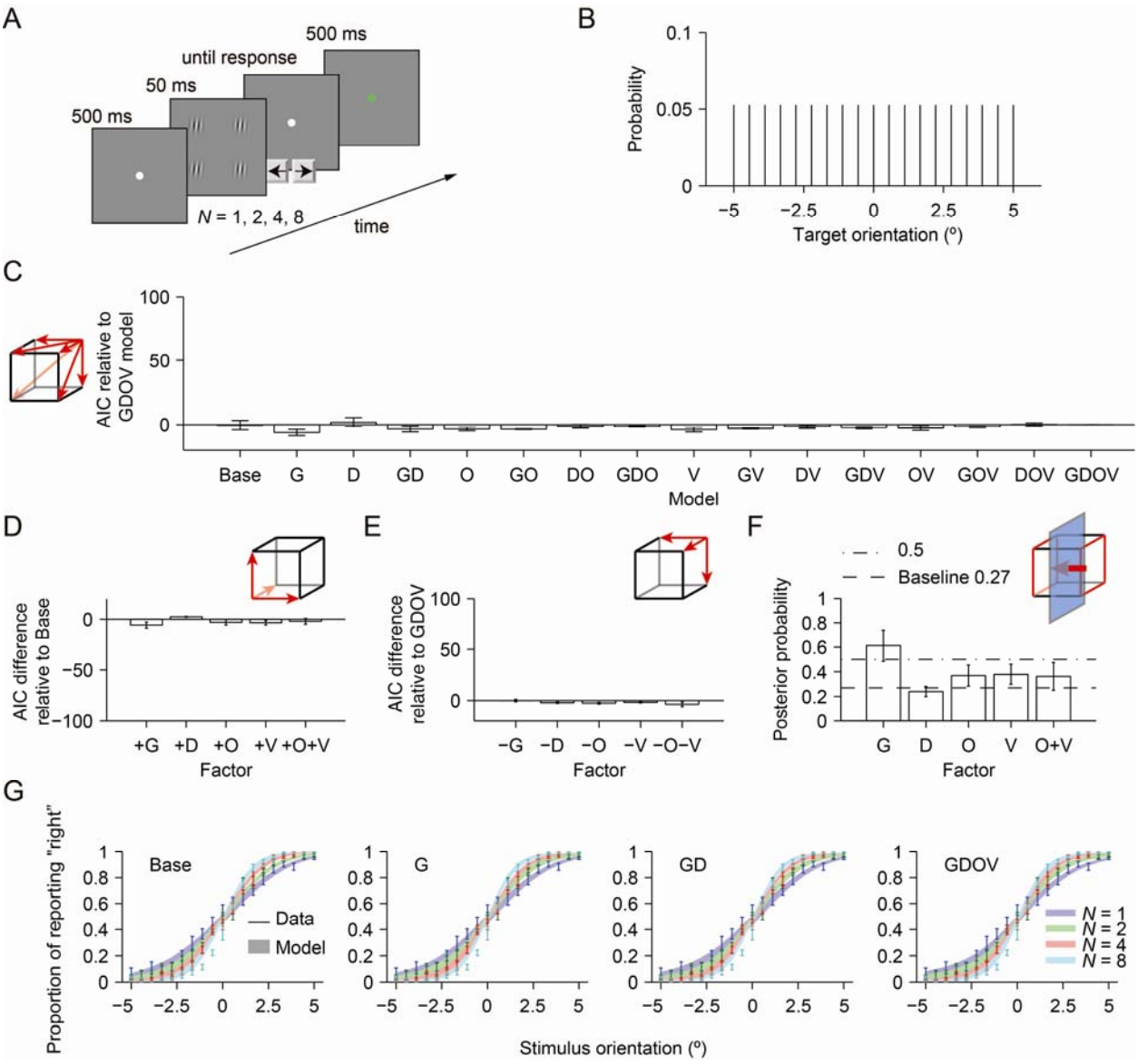

Figure 6. Experiment 4: Discrimination with all stimuli being targets and multiple set sizes. (A) Trial procedure. The trial procedure is similar to Experiment 3, but the set size is 1, 2, 4, or 8, drawn randomly on each trial. (B) Stimulus distribution. The target orientation is randomly drawn from 19 values equally spaced between -5º and 5º on each trial. (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting “right” as a function of the target orientation. Solid lines and error bars: data; shaded areas: model fits. Different colors represent different set sizes.

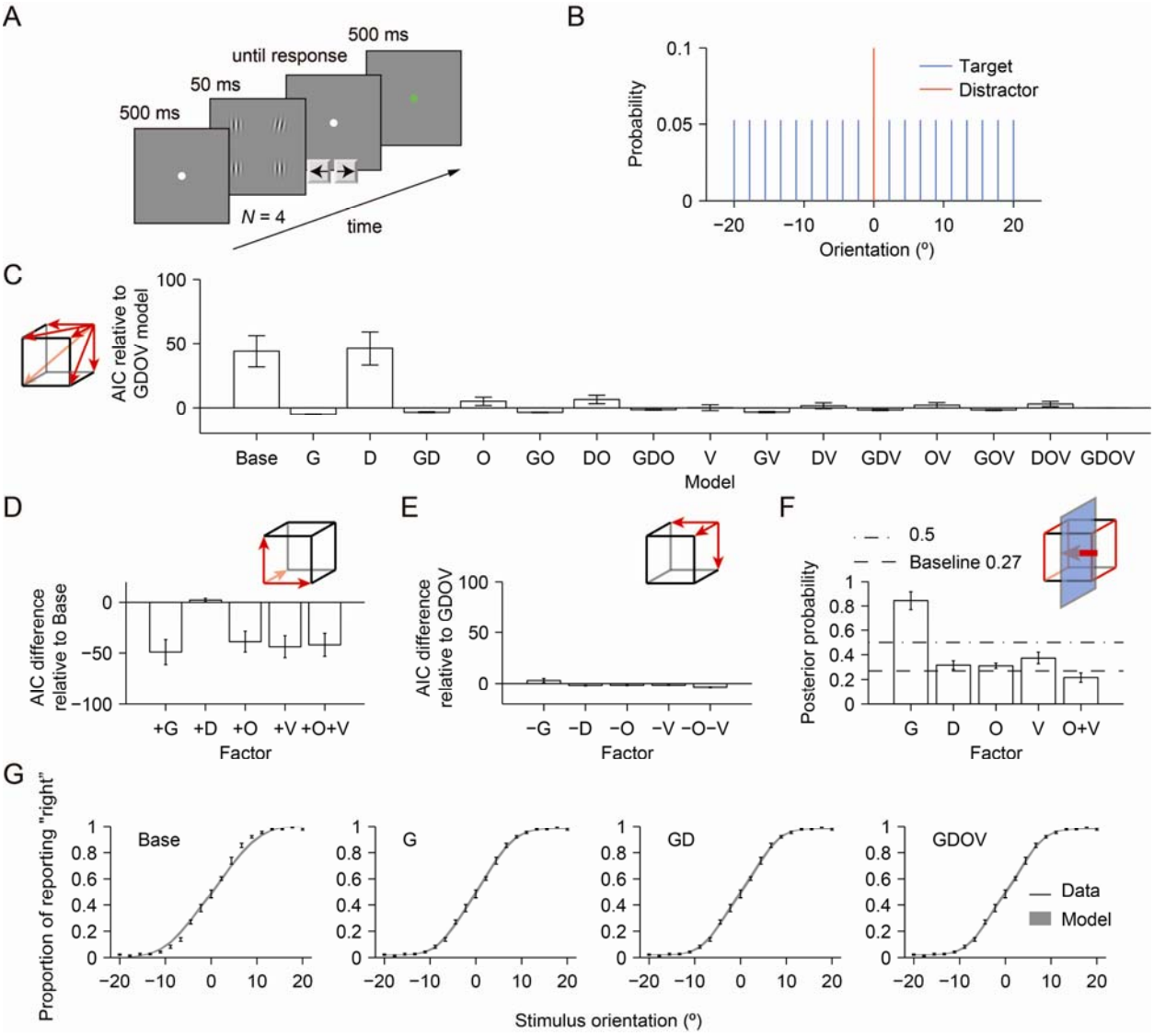

Figure 7. Experiment 5: Discrimination of a single target with a fixed number of vertical distractors. (A) Trial procedure. The trial procedure is similar to Experiment 3, but the stimulus display consists of four stimuli. Three of them are vertical distractors, and the other is the target. The subject reports the tilt of the target with respect to vertical (0°). (B) Stimulus distribution. The target orientation is randomly drawn from 19 values equally spaced between -20º and 20º on each trial (blue line), and the distractors are always vertical (red line). (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting “right” as a function of the target orientation. Solid lines and error bars: data; grey areas: model fits.

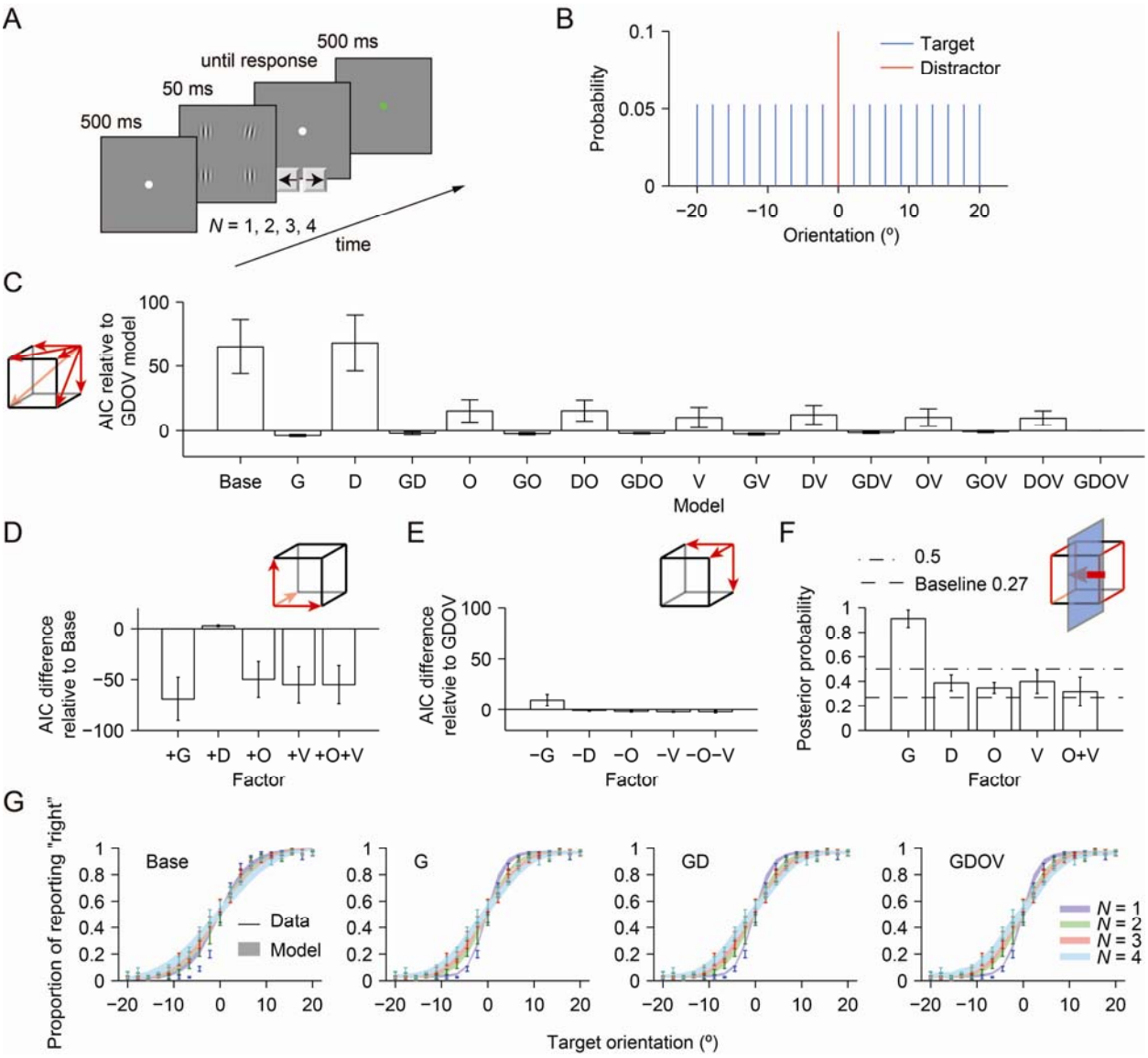

Figure 8. Experiment 6: Discrimination of a single target with a variable number of vertical distractors. (A) Trial procedure. The trial procedure is similar to Experiment 5, but the set size is 1, 2, 3, or 4, drawn randomly on each trial. (B) Stimulus distribution. The target orientation is randomly drawn from 19 values equally spaced between -20º and 20º on each trial (blue light) and the distractor orientation is always vertical (red line). (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting “right” as a function of the target orientation. Solid lines and error bars: data; shaded areas: model fits. Different colors represent different set sizes.

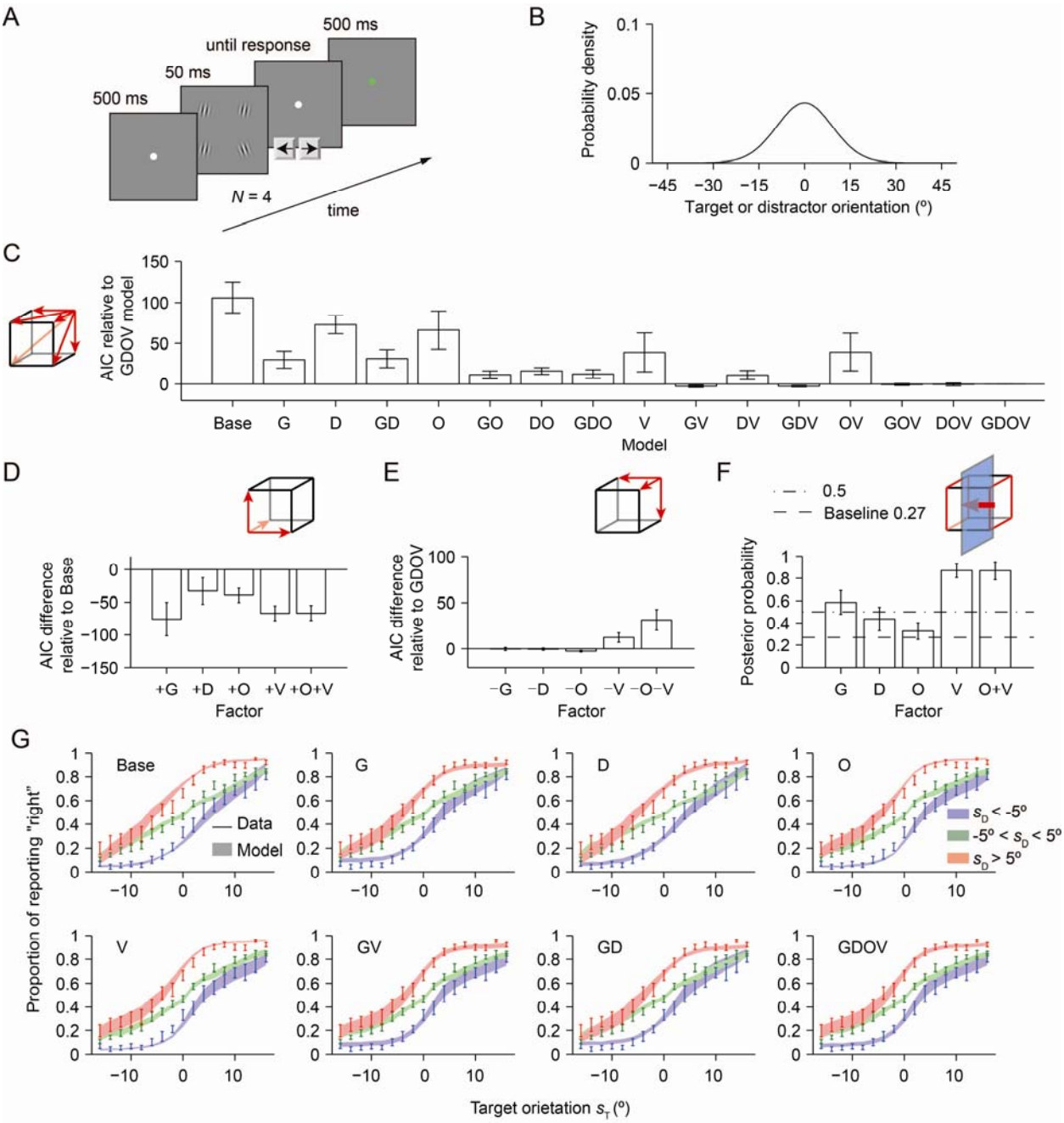

Figure 9. Experiment 7: Discrimination of a single target with a fixed number of homogeneous distractors. (A) Trial procedure. The trial procedure is similar to Experiment 5, but the stimulus display is different, although still consists of four stimuli. Three of the stimuli have identical orientations, and they are the distractors. The fourth stimulus is the target. The subject reports the tilt of the target with respect to vertical (0°). (B) Stimulus distribution. On each trial, the target orientation and the common distractor orientation are drawn independently from the same Gaussian distribution, which has a mean of 0° and a standard deviation of 9.06°. (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC

between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting “right” as a function of the target orientation. Solid lines and error bars: data; shaded areas: model fits. Different colors represent different distractor orientation ranges.

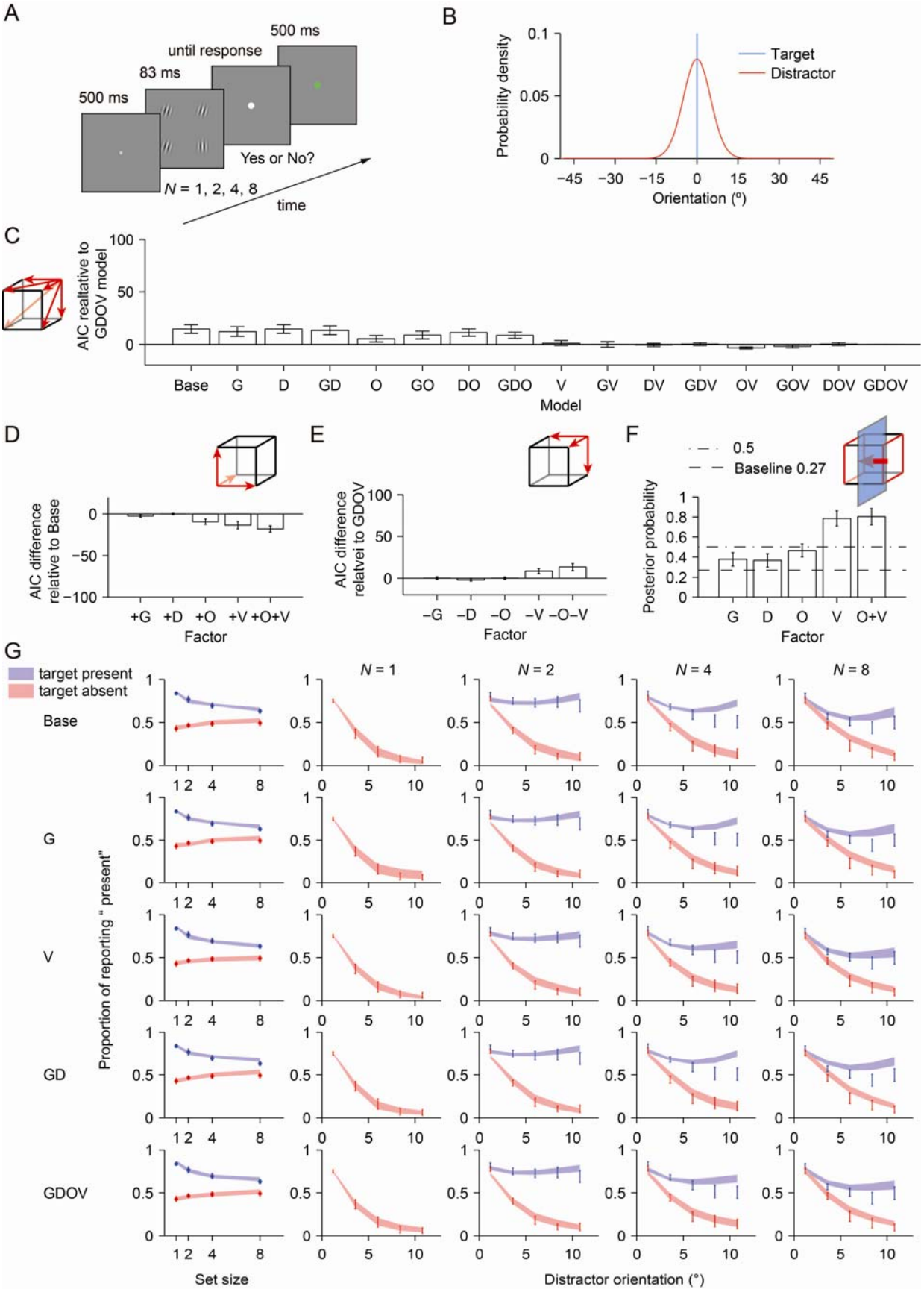

Figure 10. Experiment 8: Detection of a single target with a variable number of homogeneous distractors. (A) Trial procedure. The trial procedure is similar to Experiment 7, but the task is different. The subject reports whether the target is present or not, which is drawn randomly on each trial. On a target-absent trial, all stimuli have identical orientations and they are the distractors. On a target-present trial, one stimulus is the vertical target, and the remaining stimuli are distractors with identical orientations. The set size of each trial is 1, 2, 4, or 8, drawn randomly. (B) Stimulus distribution. The target orientation is always vertical (blue line), and the distractor orientation is drawn from a Von Mises distribution centered at vertical, with a concentration parameter of 32 (red line). (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting "present" as a function of set size or distractor orientation. Solid lines and error bars: data; shaded areas: model fits. Different colors represent target-present trials or target absent trials.

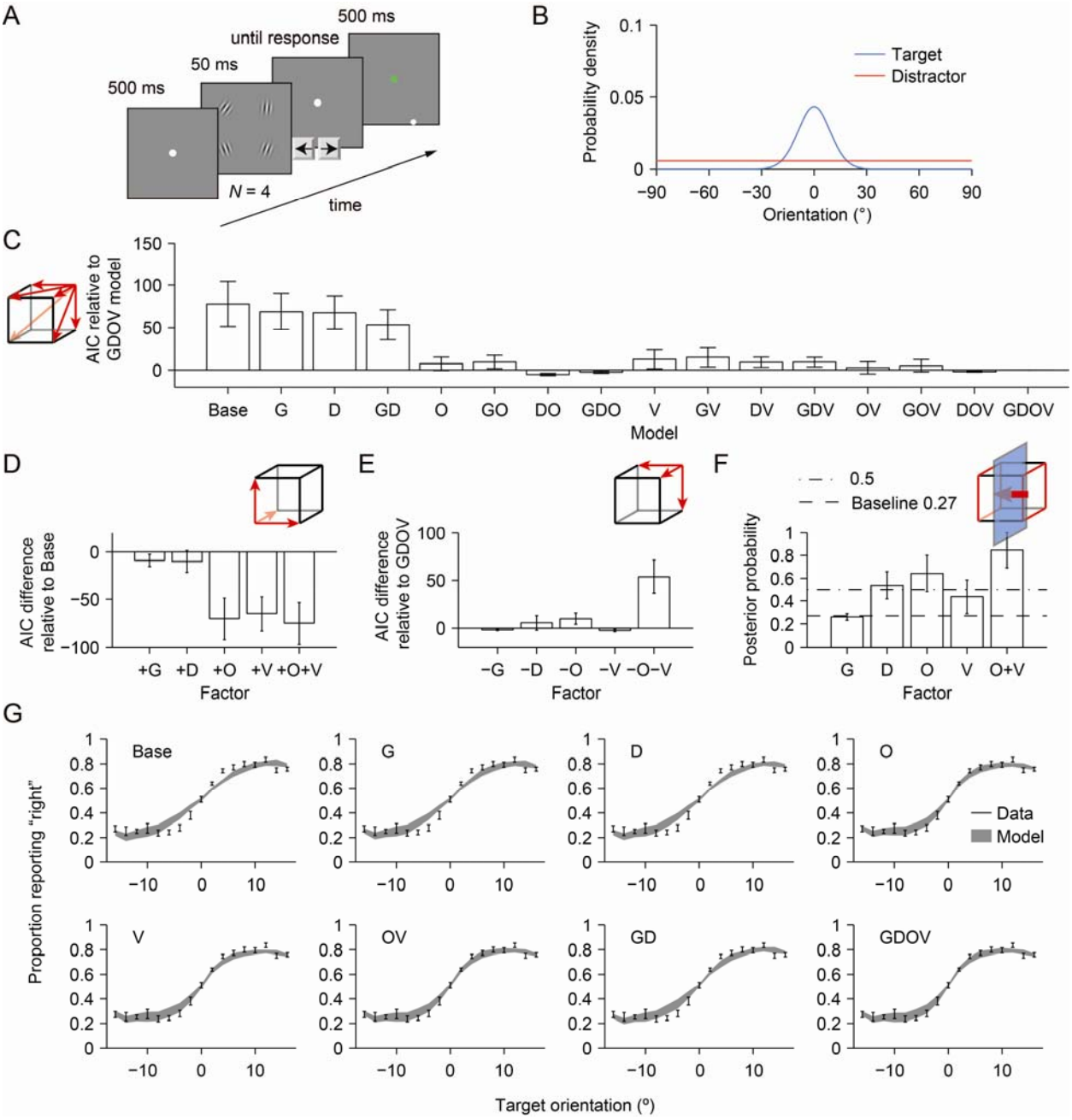

Figure 11. Experiment 9: Discrimination of a single target with a fixed number of heterogeneous distractors. (A) Trial procedure. The trial procedure is similar to Experiment 7, but the stimulus display is different, although still consists of four stimuli. Three of the stimuli are distractors, whose orientations are randomly drawn from a uniform distribution over the entire orientation space. The fourth stimulus is the target, and is randomly drawn from a narrow Von Mises distribution. The subject reports the tilt of the target with respect to vertical (0°). (B) Stimulus distribution. On each trial, the target orientation randomly drawn from a Von Mises distribution centered at vertical and with a concentration parameter of 10 (blue line). The orientation of each

distractor is independently drawn from a uniform distribution over the entire orientation space (red line). (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting "right" as a function of the target orientation. Solid lines and error bars: data; grey areas: model fits.

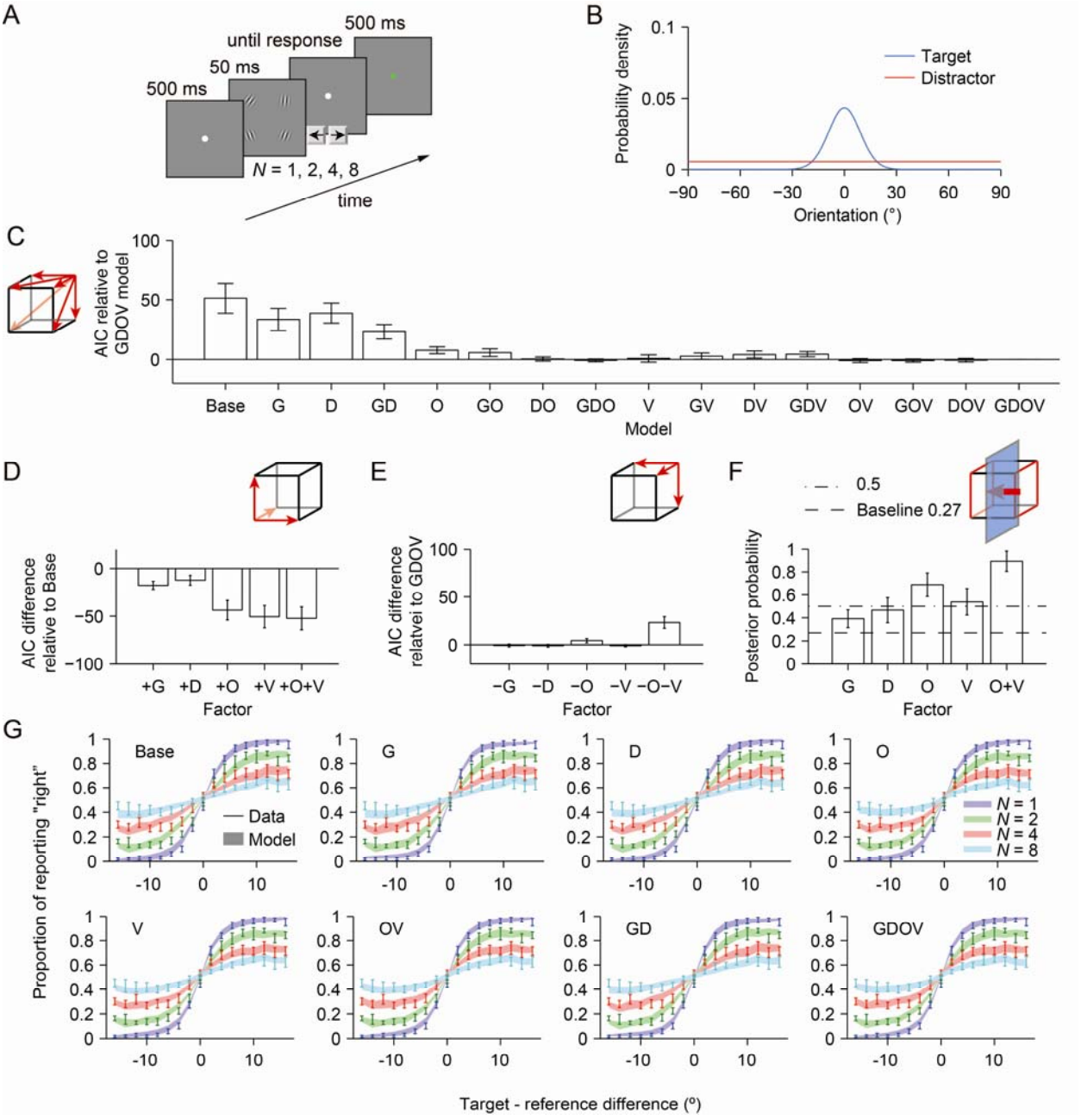

Figure 12. Experiment 10: Discrimination of a single target with a variable number of heterogeneous distractors. (A) Trial procedure. The trial procedure is similar to Experiment 9, but the set size is 1, 2, 4, or 8, drawn randomly on each trial. (B) Stimulus distribution. On each trial, the target orientation randomly drawn from a Von Mises distribution centered at vertical and with a concentration parameter of 10 (blue line). The orientation of each distractor is independently drawn from a uniform distribution over the entire orientation space (red line). (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base

model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting “right” as a function of the target orientation. Solid lines and error bars: data; grey areas: model fits. Different colors represent different set sizes.

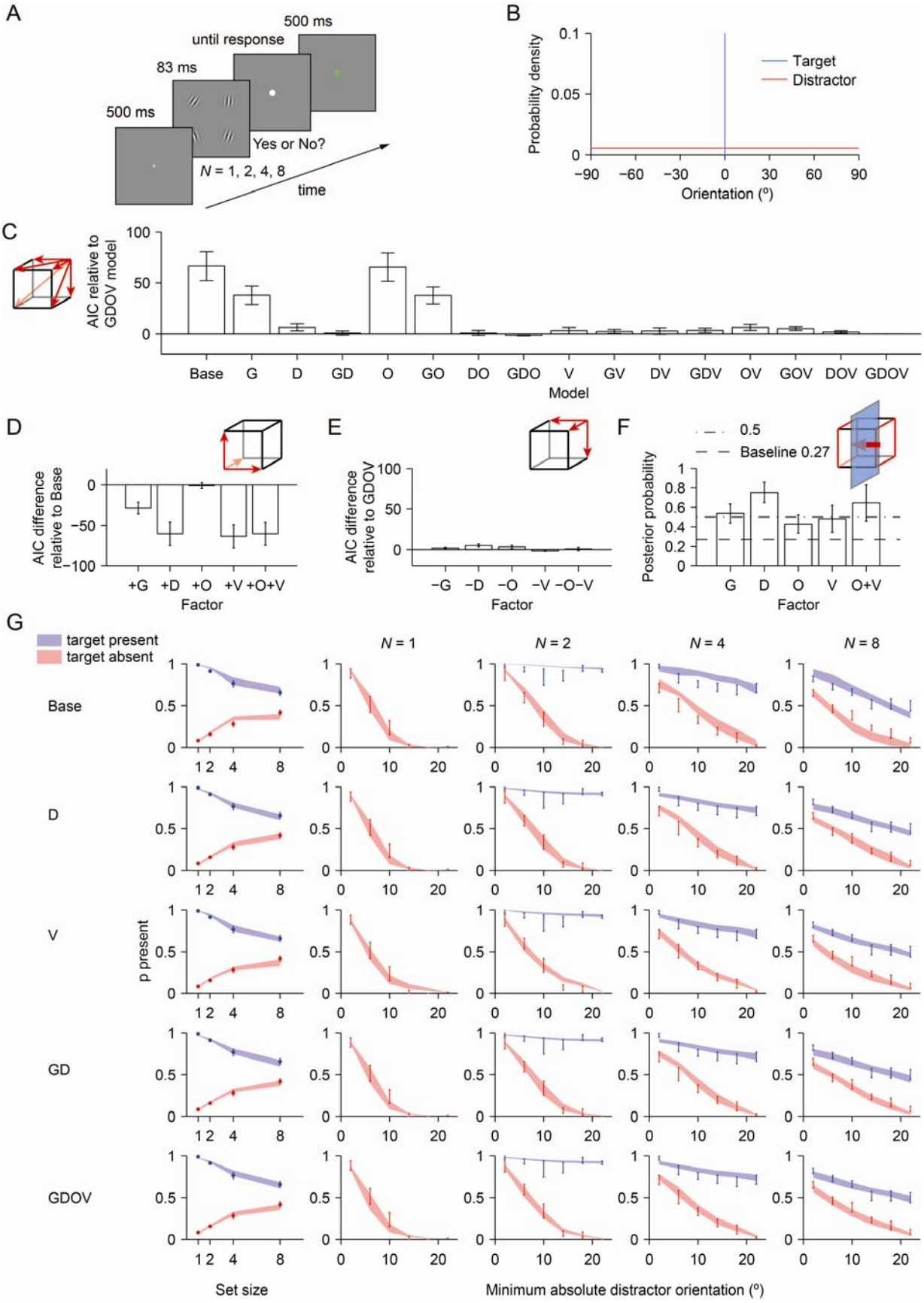

Figure 13. Experiment 11: Detection of a single target with a variable number of heterogeneous distractors. (A) Trial procedure. The trial procedure is similar to Experiment 8, but the stimulus display is different. On a target-absent trial, orientations of all stimuli are independently drawn from a uniform distribution over the entire orientation space. On a target-present trial, one stimulus is the vertical target, and the remaining stimuli are distractors with orientations independently drawn from the uniform distribution. The set size of each trial is 1, 2, 4, or 8, drawn randomly. (B) Stimulus distribution. The target orientation is always vertical (blue line), and the distractor orientations are independently drawn from a uniform distribution over the entire orientation space (red line). (C) All from full (complete model comparison). Mean and s.e.m. of the difference in AIC between each model and the full model GDOV. (D) Factor knock-in. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (E) Factor knock-out. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (F) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V. (G) Proportion of reporting "present" as a function of set size or minimum absolute distractor orientation. Solid lines and error bars: data; shaded areas: model fits. Different colors represent target-present trials or target absent trials.

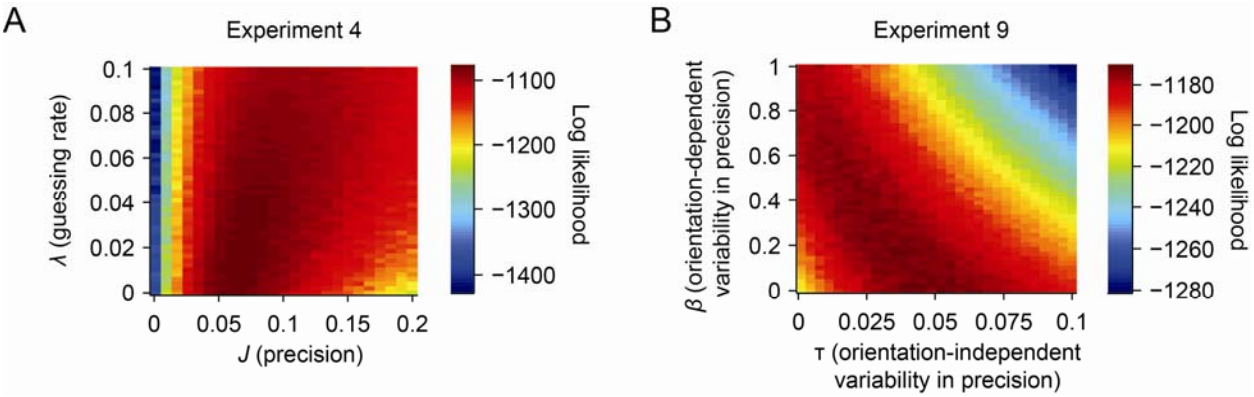

Figure 14. Trade-off. (A) Trade-off between precision J and guessing rate λ. We generate a synthetic data set with the G model in Experiment 4, with a J of 0.08 deg^{-2} and a λ of 0.02, and fit the data with the G model. The color plot shows the log likelihood of each combination of J and λ. Different combinations of J and λ have similar log likelihood given the synthetic data, including a zero guessing rate and lower precision than the true precision. (B) Trade-off between factors O and V. We generate a synthetic data set with model V for Experiment 9, with an orientation-independent variable precision parameter τ = 0.05, and fit the data with the OV model. The color plot shows the marginal log likelihood of each combination of orientation-dependent parameter β and τ. Different combinations of β and τ have similar log likelihoods given the synthetic data. For example, a combination of zero β and the true τ could be mimicked by a combination of non-zero β and a smaller τ.

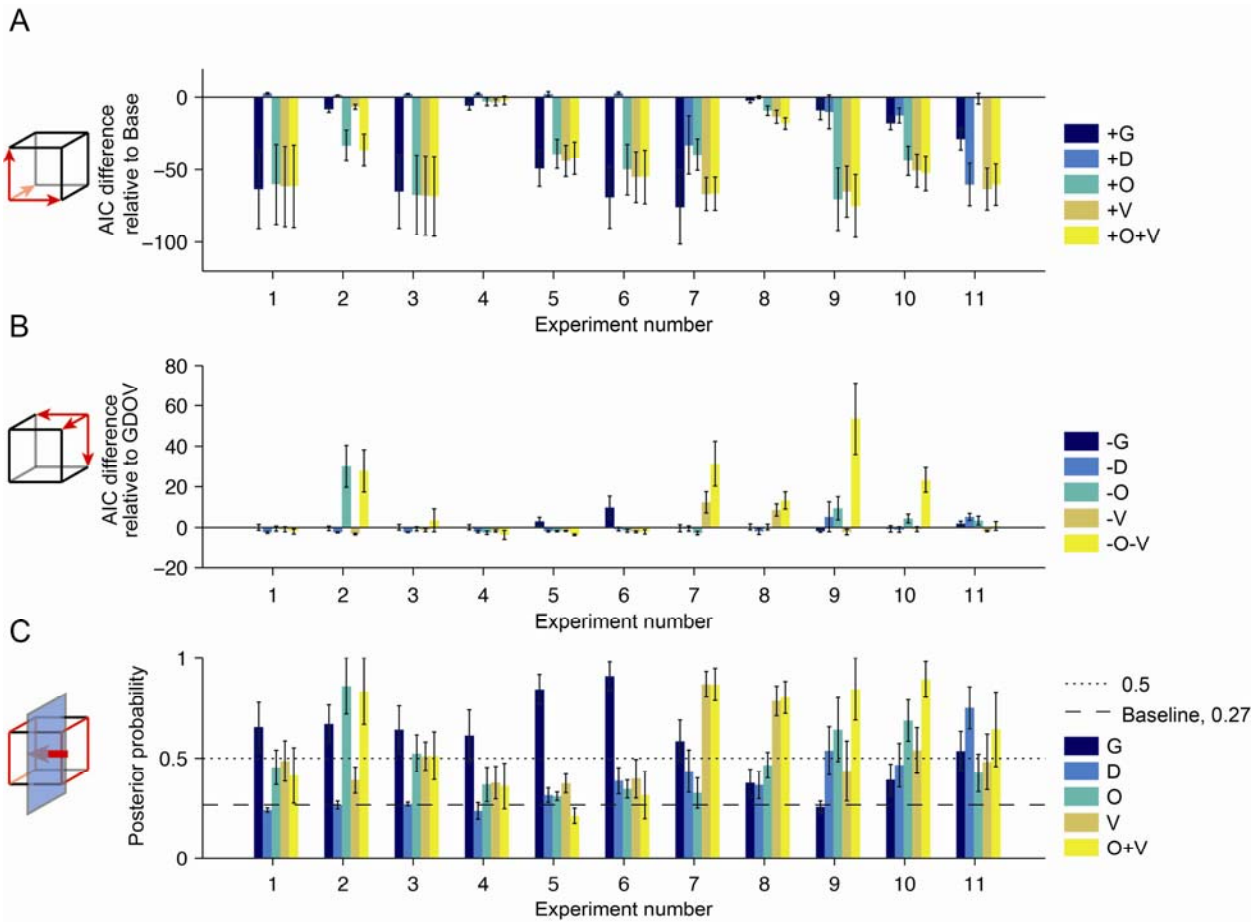

Figure 15. Summary results. (A) Factor knock-in results of all experiments. Mean and s.e.m. of the difference in AIC between models with each single factor (or a combination of factors O and V) and the Base model. (B) Factor knock-out results of all experiments. Mean and s.e.m. of the difference in AIC between models without each single factor (or a combination of factors O and V) and the full model GDOV. (C) Factor posterior probabilities (mean and s.e.m.) of each factor and of the combination of factors O and V.

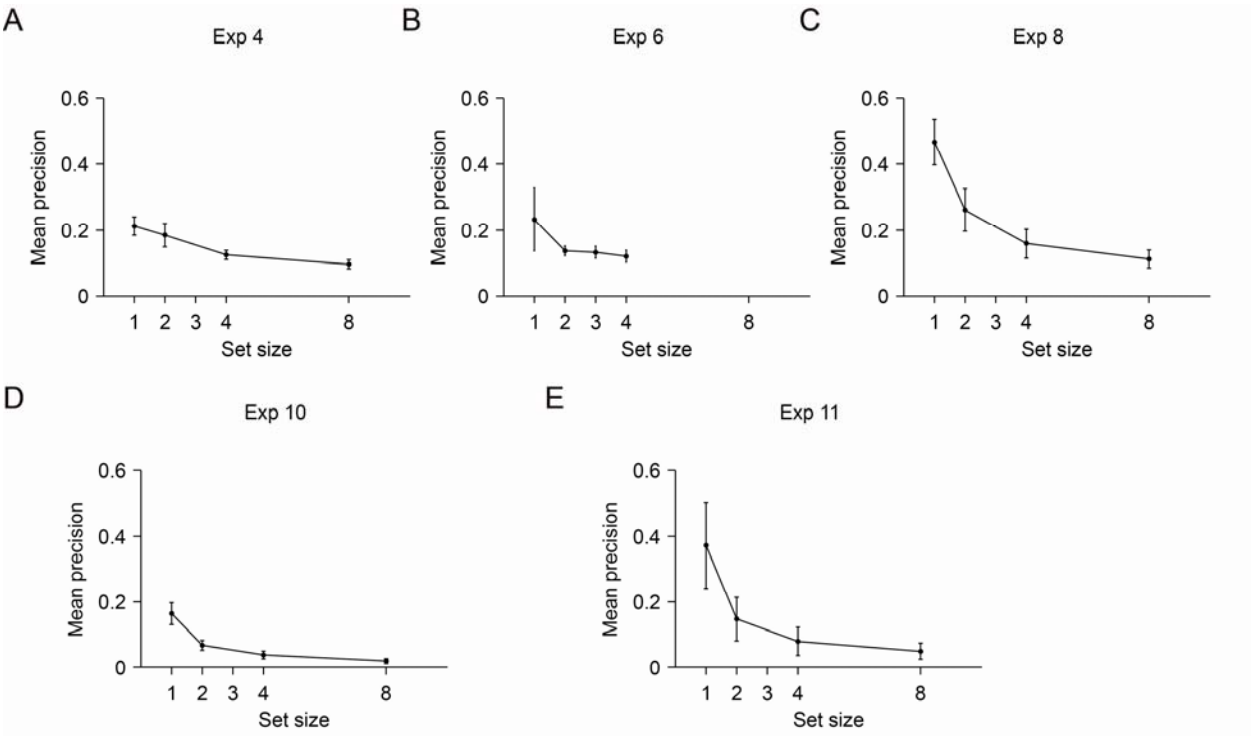

Figure 16. Relationship between mean precision and set size. The figure shows the mean precision as function of set size, estimated with the full model GDOV. Dots and error bars are the means and the s.e.ms. (A) Estimated mean precisions versus set size for Experiment 4. There is significant effect on set size: repeated-measures ANOVA, $F(3, 6) = 4.18$, $p = 0.013$. (B) Estimated mean precision versus set size for Experiment 6. There is no significant effect on set size: repeated-measures ANOVA, $F(3, 6) = 1.1$, $p = 0.38$. (C) Estimated mean precision versus set size for Experiment 8. There is significant effect on set size: repeated-measures ANOVA, $F(3,15) = 8.5$, $p < 10^{-6}$. (D) Estimated mean precision versus set size for Experiment 10. There is significant effect on set size: repeated-measures ANOVA, $F(3, 11) = 10.8$, $p < 10^{-4}$. (E) Estimated mean precision versus set size for Experiment 10. There is significant effect on set size: repeated-measures ANOVA, $F(3, 6) = 3.52$, $p = 0.034$.

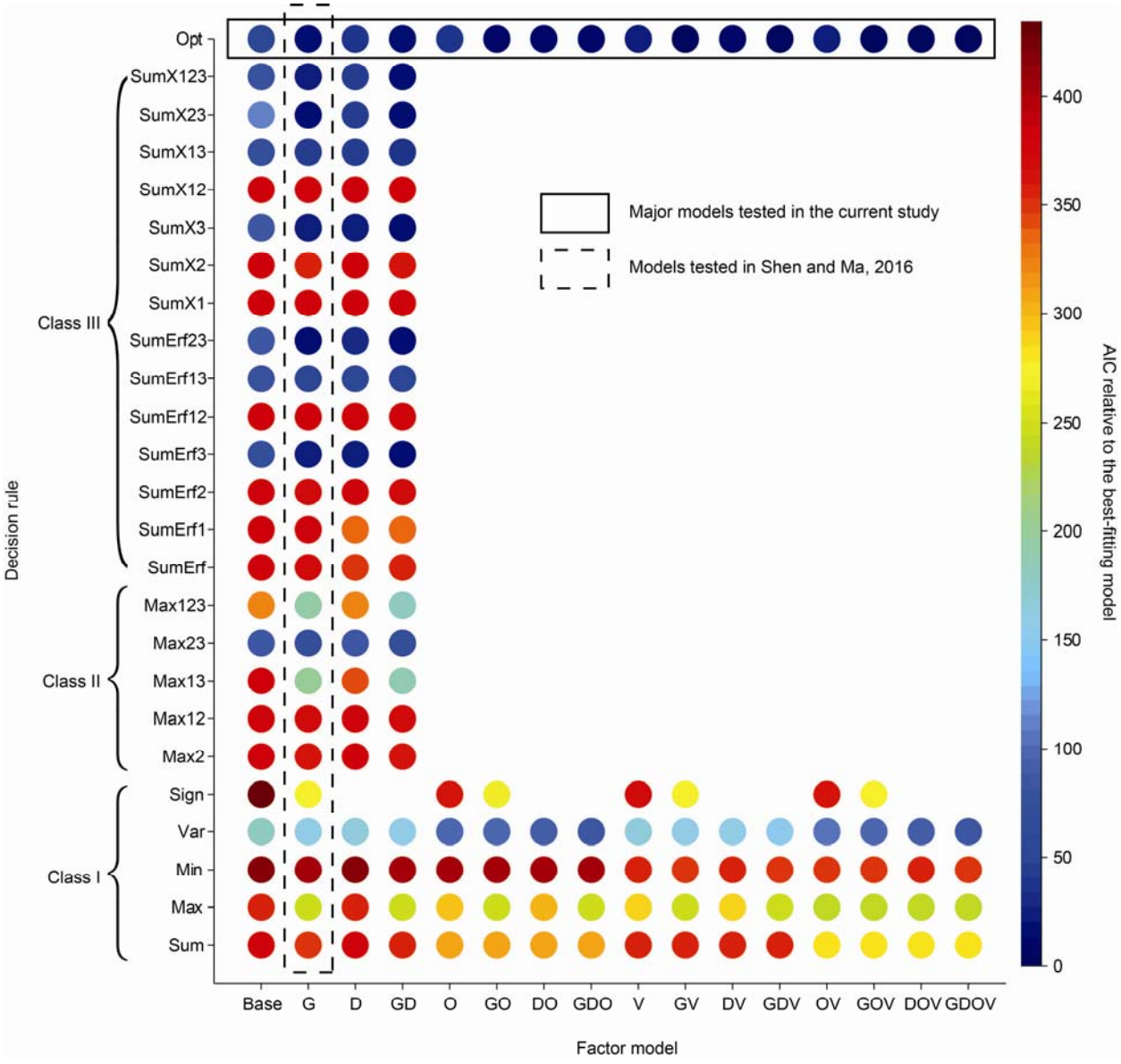

Figure 17. Cross the suboptimal decision rules with the factor models. The x-axis lists factor models, and the y-axis lists different decision rules in Shen & Ma (2016). The color of the dot represents the AIC of a hybrid model with a certain decision rule and factor model. Because the generation of Class II and Class III models in Shen and Ma, 2016 is based on the assumption of fixed precision across items, we only tested Base, G, D, GD models for Class II and Class III rules. Also, it is not clear how factor D is combined with the Sign rule in Class I, so the combination between the Sign rule and factor models with D are also missing.

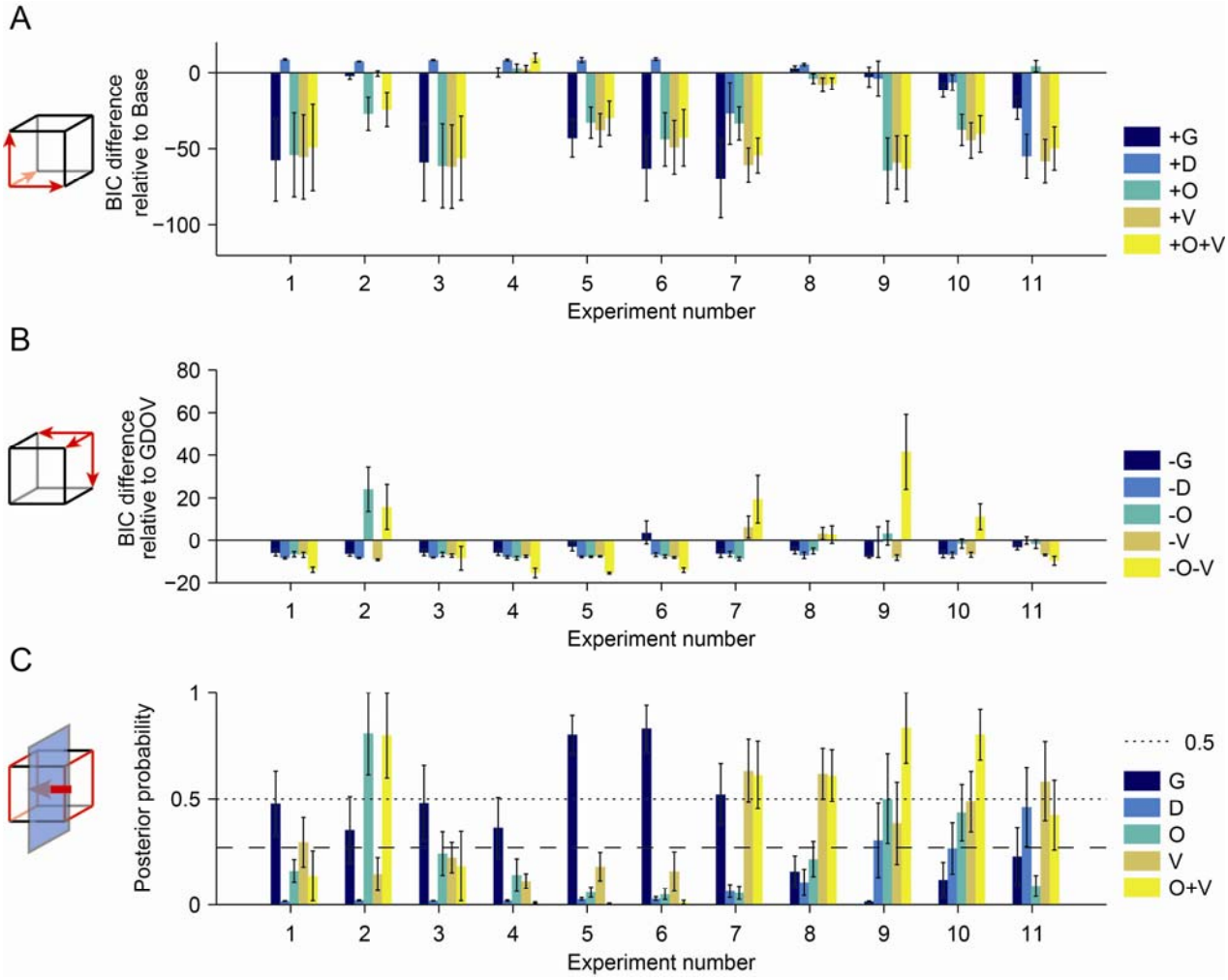

Figure A1. Summary results with BIC the measure of goodness-of-fit. Similar to Figure 15, but all quantities are computed with BIC. Results are similar to those with AIC, but the importance of all factors are lower because of more severe penalties to extra parameters.