American Psychological Association

Psychological Review

© 2018 American Psychological Association
0033-295X/19/$12.00

2019, Vol. 126, No. 1, 89–132
http://dx.doi.org/10.1037/rev0000128

Variable Precision in Visual Perception

Shan Shen
Baylor College of Medicine

Wei Ji Ma
New York University and Baylor College of Medicine

Given the same sensory stimuli in the same task, human observers do not always make the same response. Well-known sources of behavioral variability are sensory noise and guessing. Visual short-term memory (STM) studies have suggested that the precision of the sensory noise is itself variable. However, it is unknown whether precision is also variable in perceptual tasks without a memory component. We searched for evidence for variable precision in 11 visual perception tasks with a single relevant feature, orientation. We specifically examined the effect of distractor stimuli: distractors were absent, homogeneous and fixed across trials, homogeneous and variable, or heterogeneous and variable. We first considered 4 models: with and without guessing, and with and without variability in precision. We quantified the importance of both factors using 6 metrics: factor knock-in difference, factor knock-out difference, and log factor posterior ratio, each based on AIC or BIC. According to all 6 metrics, we found strong evidence for variable precision in 5 experiments. Next, we extended our model space to include potential confounding factors: the oblique effect and decision noise. This left strong evidence for variable precision in only 1 experiment, in which distractors were homogeneous but variable. Finally, when we considered suboptimal decision rules, the evidence also disappeared in this experiment. Our results provide little evidence for variable precision overall and only a hint when distractors are variable. Methodologically, the results underline the importance of including multiple factors in factorial model comparison: Testing for only 2 factors would have yielded an incorrect conclusion.

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

When presented with the same stimuli in the same perceptual task, human observers do not always make the same response. One source of such variability is noise in the encoding stage—the mapping from the stimulus to the internal representation. This mapping is 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 furthermore common to assume that such sensory or encoding noise follows a 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).

In recent years, the idea has been explored that encoding precision—roughly the inverse of the variance of the sensory noise—is itself a random variable. Such random variability is distinct from the systematic variation of precision with set size (Mazyar, van den Berg, & Ma, 2012; Mazyar, van den Berg, & Seilheimer, 2013; Palmer, 1990; Shaw, 1980; Wilken & Ma, 2004). Throughout this article, variability in precision will refer to variability at a given set size, but this variability could occur both across trials and within a trial across different stimuli. Variable-precision models have been used to model human (Donkin, Nosofsky, Gold, & Shiffrin, 2013; Fougnie, Suchow, & Alvarez, 2012; Keshvari, van den Berg, & Ma, 2012, 2013; Oberauer & Lin, 2017; Pratte, Park, Rademaker, & Tong, 2017; van den Berg, Awh, & Ma, 2014; van den Berg, Shin, Chou, George, & Ma, 2012) and monkey (D. T. Devkar, Wright, & Ma, 2015; D. Devkar, Wright, & Ma, 2017) behavior in visual short-term memory (STM) tasks as well as human behavior in visual search tasks (Bhardwaj, van den Berg, Ma, & Josić, 2016; Mazyar et al., 2012, 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 might have a parallel in the single (Bays, 2014) or double stochasticity of neural spike counts (Churchland et al., 2011; Goris, Movshon, & Simoncelli, 2014; van den Berg, Yoo, & Ma, 2017).

Variable precision could in principle be confounded with or partly explained by other factors. First, on some proportion of trials, encoding precision might be exactly zero, for example due to lapses in attention; this is typically modeled as a guessing rate

This article was published Online First October 18, 2018.

Shan Shen, Department of Neuroscience, Baylor College of Medicine; Wei Ji Ma, Center for Neural Science and Department of Psychology, New York University, and Department of Neuroscience, Baylor College of Medicine.

This work was funded by Grant R01EY020958 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. We thank Ronald van den Berg for advice and assistance during experimental design.

Correspondence concerning this article should be addressed to Wei Ji Ma, Center for Neural Science and Department of Psychology, New York University, 4 Washington Place, New York, NY 10003. E-mail: weijima@nyu.edu

(Harvey, 1986; Watson & Pelli, 1983; Wichmann & Hill, 2001). Moreover, in binary decisions, errors in the mapping between decision and motor output are mathematically equivalent to guesses. Because variable precision in a sense interpolates between zero precision and a fixed nonzero precision, it sometimes mimics guessing (Keshvari, van den Berg, & Ma, 2013; van den Berg et al., 2012, 2014). Second, variability in precision could be partly explained by systematic variations of precision across the stimulus space. For example, cardinal orientations (horizontal or vertical) are encoded with higher precision than oblique orientations. This phenomenon is called the "oblique effect" (Andrews, 1965, 1967; Appelle, 1972; Girshick, Landy, & Simoncelli, 2011; Pratte et al., 2017) and is an example of *heteroskedasticity*, whereby some measure of dispersion (skedasis) differs across subgroups. Heteroskedasticity has also been described in color perception and color visual STM (Bae, Olkkonen, Allred, Wilson, & Flombaum, 2014). Heteroskedasticity could be due to a nonuniform distribution of the preferred stimuli of visual cortical neurons (Li, Peterson, & Freeman, 2003; De Valois, Yund, & Hepler, 1982; Mansfield & Ronner, 1978). This distribution might in turn have adapted to stimulus statistics in natural environments (Attneave, 1954; Barlow, 1961; Girshick et al., 2011) and therefore might be explained by theories of efficient coding (Ganguli & Simoncelli, 2014; Wei & Stocker, 2015). Third, decision noise, or suboptimality in inference, might be confounded with sensory noise in general, and with variable precision in particular. Decision noise refers to any noise in the mapping from the internal representation to the decision (Mueller & Weidemann, 2008). Statistical inefficiency or inference noise (Burgess, Wagner, Jennings, & Barlow, 1981; Drugowitsch, Wyart, Devauchelle, & Koechlin, 2016; Liu, Knill, & Kersten, 1995), model mismatch (Beck, Ma, Pitkow, Latham, & Pouget, 2012; Orhan, Sims, Jacobs, & Knill, 2014), and other forms of systematic suboptimal inference (Gigerenzer & Goldstein, 1996; Shen & Ma, 2016; Simon, 1956) could mimic decision noise, because a model that treats the decision stage as optimal will attribute any systematic deviations from optimality to random variability, that is, decision noise.

In principle, it is possible that variability in precision found in previous work captures what is in reality guessing, heteroskedasticity, or decision noise. Only a few studies, all in the realm of visual STM, have attempted to disentangle some of these factors. Some studies have compared a variable-precision model with a fixed-precision model with a lapse rate (D. T. Devkar et al., 2015; Keshvari et al., 2012; van den Berg et al., 2012, 2014). Other studies have argued that the oblique effect accounts for most of what otherwise would be designated as variable precision (Pratte et al., 2017). Here, we attempt to distinguish guessing, the oblique effect, and decision noise from residual variable precision by including all factors simultaneously in our models.

Most previous studies that claim evidence for residual variable precision are in the realm of visual STM (Donkin et al., 2013; Fougnie et al., 2012; Keshvari et al., 2012, 2013; Oberauer & Lin, 2017; Pratte et al., 2017; van den Berg et al., 2012, 2014). In that domain, explanations for residual variable precision have included variability in spike counts for a given gain (Bays, 2014), fluctuations in attention (Cohen & Kohn, 2011; Cohen & Maunsell, 2009), shifts in attention (Lara & Wallis, 2012), and variable memory decay (Fougnie et al., 2012). However, only the last of these seems specific to memory; the other explanations would predict that residual variable precision also plays a role in perception without a memory component. Therefore, the present study, on residual variable precision in purely perceptual tasks, could help narrow down possible mechanisms of residual variable precision.

Experimental Methods

Experimental Design

We conducted eight new target categorization experiments and analyzed the results of three previously published experiments (Table 1, Figure 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.
- Subjects fixated and all stimuli were presented at the same eccentricity (5° of visual angle).
- Stimuli were presented for a short duration (50 ms or 83 ms).
- There was no substantial visual STM component.
- Subjects made binary choices.
- There were no intertrial dependencies.

We designed our experiments to search for variability in precision that cannot be accounted for by set size, guessing, the oblique effect, or decision noise. In the realm of visual STM, it has been suggested that precision is variable due to stochasticity in the rate of decay of memory (Fougnie et al., 2012); however, our study is not a memory study. Another idea has been that stimulus context has a large effect on the quality of any one stimulus (Brady & Alvarez, 2015). We take inspiration from this suggestion and examine whether evidence for residual variable precision is stronger in experiments where the stimulus context is more variable. Concretely, we considered a variety of visual perception tasks that differed in the complexity of the distractor context (Table 1): in Experiments 1 to 4, there were no distractors; in Experiments 5 and 6, distractors were homogeneous and their value remained unchanged over trials; in Experiments 7 and 8, distractors were homogeneous but varied across trials; and finally, in Experiments 9–11, distractors were both heterogeneous and variable across trials. We hypothesized that the evidence for the "residual" variable precision would be higher when the distractor context is more complex.

In addition to distractor context, our experiments differed in other aspects (Table 1), including task type (Experiments 8 and 11 are detection tasks; others are categorization tasks), orientation range (narrow range in Experiments 1, and 3–8; full range in Experiments 2, 9, and 11), number of targets (multiple targets in Experiment 2; one target in all others), set size (set size equal to 1 on all trials in Experiment 1 and some trials in Experiments 4, 6, 8, 10, and 11; set size is greater than 1 otherwise), set size context (single set size in Experiments 1, 2, 3, 5, 7, 9; multiple set sizes in other experiments), and ambiguity (Experiments 9 and 10 contain ambiguity, others not). To the best of our ability, we examined the effects of these factors, to ensure that our conclusions are robust.

Table 1
Overview of Experiments

Experiment	Number of subjects	Number of stimuli	Number of targets	Task	Distractors	Ambiguity in the task
1	6	1	1	Target categorization relative to vertical	None	No
2	5	2	1	Target categorization relative to reference (stimulus on the right)	None	No
3	6	4	all	Target categorization relative to vertical	None	No
4	6	1, 2, 4, 8	all	Target categorization relative to vertical	None	No
5	6	4	1	Target categorization relative to vertical	Vertical	No
6	6	1, 2, 3, 4	1	Target categorization relative to vertical	Vertical	No
7	10	4	1	Target categorization relative to vertical	Homogeneous, variable	No
8	13	1, 2, 4, 8	0, 1	Target detection, vertical target	Homogeneous, variable	No
9	6	4	1	Target categorization relative to vertical	Heterogeneous, variable	Yes
10	11	1, 2, 4, 8	1	Target categorization relative to vertical	Heterogeneous, variable	Yes
11	6	1, 2, 4, 8	0, 1	Target detection, vertical target	Heterogeneous, variable	No

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

Apparatus and Stimuli

Subjects were seated at a viewing distance of approximately 60 cm. All stimuli were displayed on a 21-in. LCD monitor with a refresh rate of 60 Hz and a resolution of 1,280 × 1,024 pixels. The stimulus displays were composed of Gabor patches shown on a gray background. In Experiments 1–7, 9, and 10, background luminance was 29.3 cd/m^2 and the Gabors had a peak luminance of 35.2 cd/m^2, a spatial frequency of 3.1 cycles per degree of visual angle, a standard deviation of the Gaussian envelope of 0.25° of visual angle, and a phase of 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^2 and the Gabors had a peak luminance of 122 cd/m^2, a spatial frequency of 1.6 cycles per degree of visual angle, a standard deviation of the Gaussian envelope of 0.29° of visual angle, and a phase of 0 for the cosine pattern.

Experimental Procedures

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 responded by pressing a button. Response time was not limited. After the response, feedback regarding correctness was given by changing the color of the fixation dot (green for correct, red for incorrect; 500 ms; Figure 1). Experiments 1–7, 9, and 10 were visual categorization tasks and Experiment 8 and 11 were visual detection tasks.

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 = 3,000 trials per subject. All blocks were statistically identically to each other.

Experiment 8 (Experiment 2 in Mazyar et al., 2013) consisted of four sessions; here, we analyze only the two sessions with homogeneous distractors (for details, see Mazyar et al., 2013). Each session consisted of four blocks, and each block contained 175 trials, for a total of 2 × 4 × 175 = 1,400 trials per subject analyzed here.

Experiment 11 (Experiment 1 in Mazyar et al., 2013) consisted of six sessions; here, we analyze only the two sessions with “high” heterogeneity (for details, see Mazyar et al., 2013). Each session consisted of four blocks, and each block contained 175 trials, for a total of 2 × 4 × 175 = 1,400 trials per subject analyzed here.

Stimulus Displays and Tasks

We now describe the stimulus display in each of the 11 experiments (Figure 1). We will use the phrase “drawn randomly” as shorthand for “drawn randomly from a uniform distribution over the values specified.” The radial positions of all stimuli in all experiments were 5° of visual angle relative to fixation. For the angular positions of the stimuli, we use the standard convention of polar coordinates: 0° corresponds to the positive horizontal axis, and positive values correspond to positions counterclockwise with respect to that axis. For stimulus orientations, we use the convention that is most natural given our orientation distributions: 0° is vertical and positive values are clockwise.

Experiment 1. The stimulus display consisted of a single stimulus in one of four angular positions: −135°, −45°, 45°, and 135°. Stimulus orientation was drawn randomly from 19 values equally spaced between −15° and 15°. The subject reported the tilt with respect to vertical of a single oriented stimulus.

Experiment 2. The stimulus display consisted of two stimuli, placed on the horizontal axis to the left and right of the fixation. The stimulus on the right was the reference stimulus, whose

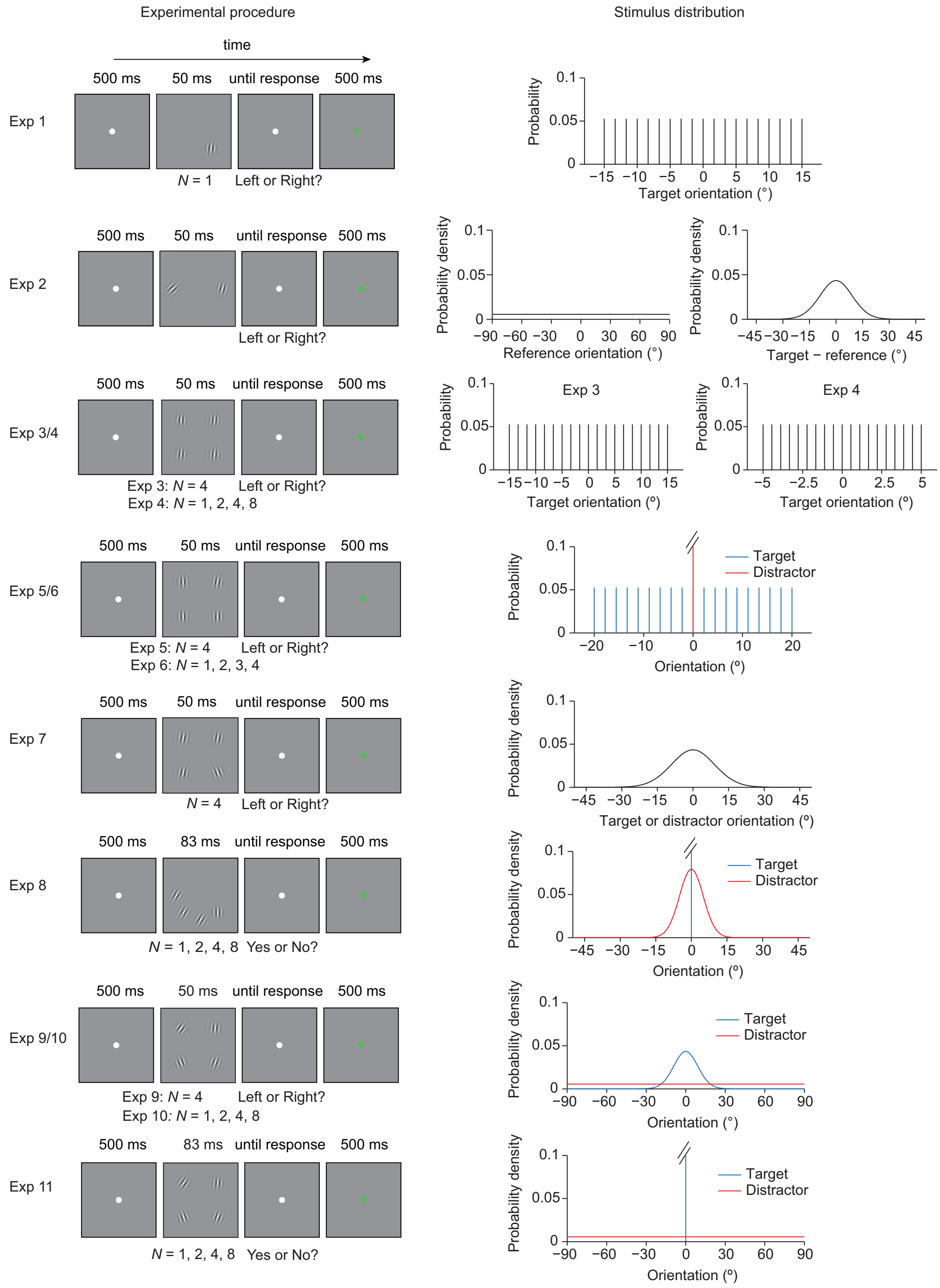


Figure 1. Experimental designs. The left column shows the trial procedure and the right column shows the orientation distribution of the stimuli. See the online article for the color version of this figure.

orientation s_{ref} was drawn randomly (from a uniform distribution over the entire orientation space). The stimulus on the left was the target stimulus, whose orientation was drawn randomly from a Von Mises distribution centered at s_{ref} with a concentration parameter of 10. The subject reported whether the target was oriented clockwise or counterclockwise with respect to s_{ref}.

Experiments 3 and 4. All stimuli were targets, and the subject reported the tilt of their common orientation. In Experiment 3, set size was 4, and the angular positions were as in Experiment 1. The common orientation was drawn randomly from 19 values equally spaced between $-15°$ and $15°$. In Experiment 4, set size was 1, 2, 4, or 8, drawn randomly. Angular positions were chosen as follows, in order to maximize spacing. At set size 8, we used all eight angular positions: 0°, 45°, 90°, 135°, 180°, $-45°$, $-90°$, and $-135°$. At set sizes 1, 2, and 4, we placed the first stimulus at a random angular position. At set size 2, we then placed the second stimulus diametrically opposite to the first, while at set size 4, we placed the remaining three stimuli at every other position. Stimulus orientation was drawn randomly from 19 values equally spaced between $-5°$ and 5°.

Experiment 5 and 6. Experiment 5 and 6 were target classification tasks. Subjects reported the tilt of the target relative to vertical. In Experiment 5, 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 from the four positions, was the target. Target orientation was drawn randomly from 19 values equally spaced between $-20°$ and 20°. The design of Experiment 6 was identical to that of Experiment 5, except that set size was 1, 2, 3, or 4, drawn randomly. Angular positions were drawn randomly.

Experiment 7. We reanalyzed data from a previously published target classification experiment (Shen & Ma, 2016). Set size was 4 and the angular positions were the same as in Experiment 1. Each display contained one target and three distractors; target position was drawn randomly from the four positions. The target orientation and the common distractor orientation were drawn independently from the same Von Mises distribution centered at vertical, with a concentration parameter of 10 (similar to a Gaussian distribution with a standard deviation of 9.06°). Subjects reported the tilt of the target (the unique stimulus).

Experiment 8. Experiment 8 was an orientation detection task. Subjects reported whether or not a target was present. Set size was 1, 2, 4, or 8, drawn pseudorandomly. 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 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 from the available positions. The common orientation of the distractors was drawn from a Von Mises distribution centered at vertical, with a concentration parameter of 32 (similar to a Gaussian distribution with a standard deviation of 5.06°).

Experiment 9 and 10. Experiments 9 and 10 were target classification tasks. Subjects reported the tilt of the target relative to vertical. In Experiment 9, 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. Target orientation was drawn from a Von Mises distribution with a mean of 0 and a concentration parameter of 10 (similar to a Gaussian distribution with a standard deviation of 9.06°). Each of the distractor orientations was drawn independently from a uniform distribution over the entire orientation space. The tasks in these experiments contained ambiguity, in the sense that on some trials, either answer could be correct even in the absence of sensory noise, because of the overlap between the target and distractor distributions; as experimenters, we set the tilt of the generated target as the correct answer. To help subjects learn the task, we provided 10 static example Gabor patches from the target and distractor distributions and verbally explained that the distractors were more likely to have large tilts than the targets. The subjects performed well above chance (71.7 ± 1.6%, student's t test: $t(5) = 13.9$, $p < 10^{-4}$). 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. This experiment combined distractors that were variable both within and across trials with multiple set sizes. Again, this experiment had ambiguity when set size was greater than 1 and therefore did not allow for perfect performance. Nevertheless, subject accuracy was 72.6 ± 1.7%, well above chance (student's t test: $t(10) = 13.1$, $p < 10^{-6}$).

Experiment 11. We reanalyzed data from a previously published target detection task (Experiment 1 in Mazyar et al., 2013). The basic paradigm was the same as in Experiment 8, except that the distractors were heterogeneous. Each distractor was independently drawn from a uniform distribution over the entire orientation space, which was the same as in Experiments 9 and 10. This experiment differed from those not only in the type of task (detection instead of categorization), but also in the absence of ambiguity: the stimulus statistics did not preclude perfect performance.

Theory

For each experiment, we build process models in which the stimuli give rise to measurements and the observer applies a decision rule to the measurements to produce a category estimate (Figure 2A). Our models consist of three steps:

1. The generative model: a statistical description of the noisy internal measurements of the stimuli and of the observer's beliefs about how the stimuli are generated (which may or may not be how they are actually generated). In this step, we consider two kinds of variability in the precision of the measurement noise: the oblique effect (O) and "residual" variable precision (V).

2. The observer's decision model. In each experiment, we assume that the observer applies an optimal decision rule, which maximizes the posterior distribution under the generative model of Step 1. In some experiments, we also consider alternative suboptimal decision rules. Note that even the optimal decision rule is not optimal in an absolute sense, because (a) measurement noise is present, (b) the generative model assumed by the observer is not necessarily identical to the true generative model, and

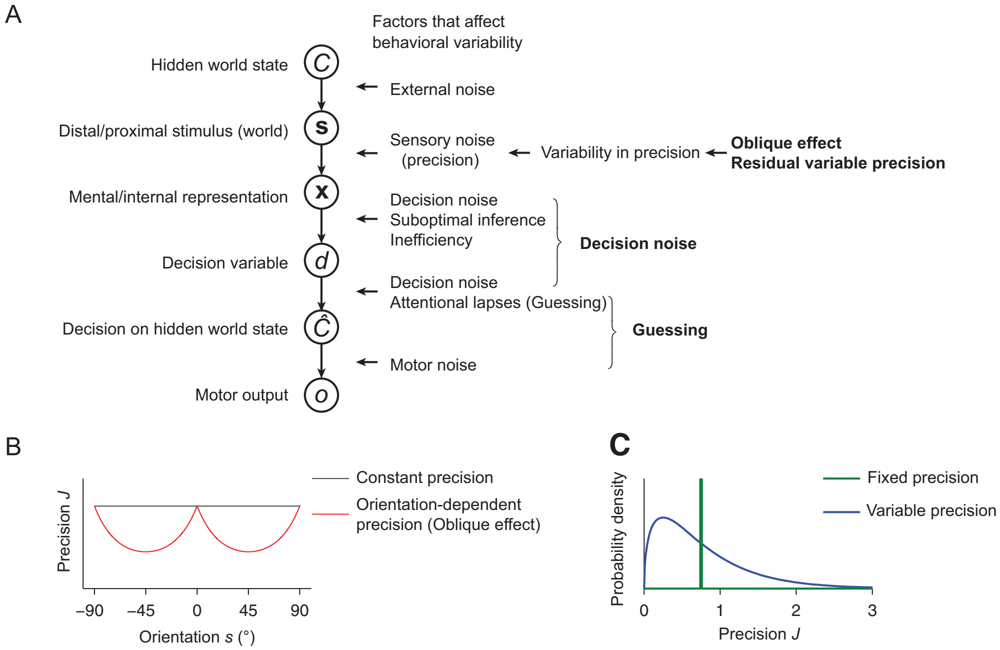


Figure 2. Generative model and factors that might affect behavioral variability. (A) The diagram shows the generic generative model of our tasks. Each node represents a variable and each arrow between two nodes represents a conditional dependence. Factors that might affect behavioral variability are listed to the right of the diagram. Here, we test the bold-faced ones: oblique effect, residual variable precision, decision noise and guessing. (B) We model the dependence of precision *J* on orientation *s* (the oblique effect) as (red dashed). The black (solid) line represents constant precision ($\beta = 0$). (C) In variable-precision models, we model the probability distribution over precision as a gamma distribution; an example with a mean of 0.75 and a scale parameter τ of 0.5 is shown in blue (dashed line). The green (solid) line represents a delta function over precision, corresponding to fixed precision ($\tau = 0$). See the online article for the color version of this figure.

(c) we allow for the decision to be corrupted by decision noise (D).

3. Prediction for the probabilities of the possible subject responses on a given trial, (i.e., given the stimulus values on that trial). This step combines the stimulus-conditioned measurement distributions from Step 1 with the decision rule from Step 2. We also incorporate guessing (G) in this step.

We now describe each of these steps in greater detail.

Step 1a: Generative Model: Noisy Measurements

We assume that the observer makes a noisy measurement x_i of each physical orientation s_i, where $i = 1, \ldots, N$ 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 assume that the measurements are independent given the stimuli,

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

Because our stimulus feature is orientation, its space is periodic. Therefore, the most principled choice for the noise distribution $p(x_i|s_i)$ is a circular distribution. Specifically, following other work (van den Berg et al., 2012; Wilken & Ma, 2004), we choose a Von Mises distribution,

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

where κ_i is the concentration parameter, and I_0 is the modified Bessel function of the first kind of order 0.

However, when the stimulus range is narrow relative to the entire 2π radians of the circle, the Von Mises distribution is well approximated by a Gaussian distribution, which is both analytically and numerically more tractable. Therefore, when the stimulus range is narrow (such as in Experiments 1 and 3–8), or the orientation difference is narrow (Experiment 2), we assume that the distribution of x_i given s_i is Gaussian:

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

Noise level or precision is controlled by the concentration parameter κ_i (Von Mises) or by the standard deviation σ_i (Gaussian). The factor 2 in the exponent of the Von Mises distribution appears because orientation space is $[0, \pi)$ instead of $[0, 2\pi)$. In the limit of large κ_i, the Von Mises distribution converges to the Gaussian distribution, with $\kappa_i = \frac{1}{4\sigma_i^2}$.

A general definition of precision based on $p(x_i|s_i)$ is as Fisher information (Cover & Thomas, 2006). Fisher information, denoted by J, is related to the parameters above through

$$J = \frac{1}{\sigma^2} \text{ (Gaussian)} \qquad J = \frac{4\kappa I_1(\kappa)}{I_0(\kappa)} \text{ (Von Mises)}, \tag{3}$$

where I_1 is the modified Bessel function of the first kind of order 1. In previous work (Keshvari et al., 2012; Mazyar et al., 2012, 2013; van den Berg et al., 2012), we did not include the factor of 2 in Equation (1) and the factor of 4 in Equation (3), but instead rescaled orientations from $[0, \pi)$ to $[0, 2\pi)$ 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.

Step 1b: Generative Model: Variability in Precision

Next, we consider variability in the precision of the measurement noise. This variability can be due to multiple sources.

Oblique effect (O). To model the oblique effect, we introduce stimulus dependence in the dispersion parameter of the measurement distribution. For Gaussian noise, we take the standard deviation of the noise to be a rectified sine function of the stimulus orientation (Girshick et al., 2011):

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

where σ_0 is the baseline noise level and β is the amplitude parameter of orientation dependence. When $\beta = 0$, there is no oblique effect. For precision J_i, we obtain (Figure 2B):

$$J_i = \frac{1}{\sigma_0^2(1 + \beta|\sin(2s_i)|)^2} = \frac{J_0}{(1 + \beta|\sin(2s_i)|)^2}, \tag{4}$$

where $J_0 = \frac{1}{\sigma_0^2}$ is the baseline precision. We use the latter equation also for Von Mises noise.

"Residual" variable precision (V). Besides the oblique effect, precision might vary for other reasons; we will consider all other sources collectively and call them "residual" variable precision, denoted by V. Variable-precision models have been successful in describing behavior in many visual STM tasks, including delayed estimation (Fougnie et al., 2012; van den Berg et al., 2014, 2012), change detection (Keshvari et al., 2012, 2013), and change localization (D. T. Devkar et al., 2015; Keshvari et al., 2012). Most of these articles have formalized variability in precision by a assuming a gamma distribution over J_i:

$$p(J_i) = \text{Gamma}(J_i; \frac{\bar{J}}{\tau}, \tau), \tag{5}$$

where $\bar{J}$ is the mean precision, and τ is called the scale parameter (Figure 2C). We will follow this formalism here.

Combining factors O and V. In all experiments, we tested all four combinations of the two forms of precision variability: a base model with fixed precision (base), a model with only the oblique effect (O), a model with only residual variable precision (V), and a model with both (OV). In the base model, J_i is the same across stimulus i and across trials. In the O model, J_i is computed from Equation (4). In the V model, J_i is drawn independently across i and across trials from a gamma distribution with mean $\bar{J}$ and scale parameter τ (Equation 5). In the OV model, we first compute $\bar{J}$ from Equation (4), then draw J_i from a gamma distribution with mean $\bar{J}$ and scale parameter τ (Equation 5).

In experiments with multiple set sizes, we allowed J (models with fixed precision), J_0 (models with the oblique effect), or $\bar{J}$ (models with residual variable precision) to vary with set size; we did not impose a parametric form but fitted the parameter independently at each set size (Mazyar et al., 2012, 2013).

Step 1c: Generative Model: Experimental Statistics

The generative model consists not only of the distribution $p(\mathbf{x}|\mathbf{s})$, but also of the observer's beliefs about the experimental statistics. The variables relevant to those beliefs are category C (target tilted left or right in the categorization experiments, target present or absent in the detection experiments), and the individual stimulus orientations $\mathbf{s}$. We assume that the observer's beliefs about the category distribution $p(C)$ and the category-conditioned stimulus distributions $p(\mathbf{s}|C)$ are identical to the true ones, that is, the ones set 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 characterize the observer's prior probability that the stimulus was tilted right (p_{right}) in the categorization experiments, or that the stimulus was present (p_{present}) in the detection experiments.
- In Experiments 1, 3, 4, 5, and 6 we used discrete stimulus values, for example, 19 values spaced linearly between −15° and 15° (Experiments 1 and 3), between −5° and 5° (Experiment 4), and 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 observer believed that this distribution was Gaussian with the same mean and standard deviation as the actual distribution:

$$p(s_T|C) = 2 \cdot N(s_T; 0, \sigma_s^2)H(C \cdot s_T), \tag{6}$$

where σ_s denotes the standard deviation of the actual distribution, s_T denotes the target orientation, and $H(x)$ denotes the Heaviside step function. In the Results section, we examine whether this assumption affects our results.

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 ratio of the probabilities of $C = 1$ and $C = -1$ given $\mathbf{x}$:

$$d = \log\frac{p(C = 1|\mathbf{x})}{p(C = -1|\mathbf{x})}. \tag{7}$$

The Bayesian observer without decision noise uses the decision rule to report "$C = 1$" when d is positive, or in other words,

$$\hat{C} = \mathrm{sgn}(d). \tag{8}$$

The Bayesian observer is not strictly optimal, because we made two modifications in Step 1c; for a detailed distinction between the terms Bayesian and optimal, see Ma (2012). Other than those, we assume that Bayesian observer has the full knowledge of the noise distribution in the measurements (Equations 1, 2, 4, and 5).

Starting with Equation (7), we first apply Bayes rule to both the numerator and denominator:

$$d = \log\frac{p(C=1)}{p(C=-1)} + \log\frac{p(\mathbf{x}|C=1)}{p(\mathbf{x}|C=-1)}. \tag{9}$$

Now we evaluate the likelihood of each C by averaging over the stimulus vector $\mathbf{s}$:

$$p(\mathbf{x}|C) = \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|C)d\mathbf{s}. \tag{10}$$

This is an example of Bayesian marginalization, in which each possible value of the unknown variable $\mathbf{s}$ is considered.

The derivation so far applies to all experiments. $p(\mathbf{x}|\mathbf{s})$ has been evaluated in Equation (1) or (2). We now evaluate $p(\mathbf{s}|C)$ in Equation (10) separately for different experiments and for different values of C.

Case 1: All Stimuli Are Targets

This case applies to all trials ($C = \pm 1$) in three target categorization tasks (Experiments 1, 3, and 4). We write $p(\mathbf{s}|C)$ as a marginalization over the scalar target orientation s_{T}:

$$p(\mathbf{s}|C) = \int p(\mathbf{s}|s_{\mathrm{T}})p(s_{\mathrm{T}}|C)ds_{\mathrm{T}}. \tag{11}$$

Here, $p(s_{\mathrm{T}}|C)$ is the category-conditioned distribution of s_{T} (a truncated Gaussian). $p(\mathbf{s}|s_{\mathrm{T}})$ represents the distribution of the N-dimensional stimulus vector $\mathbf{s}$ given a particular s_{T}. Since all stimuli are targets, we can write $p(\mathbf{s}|s_{\mathrm{T}})$ as:

$$p(\mathbf{s}|s_{\mathrm{T}}) = \prod_{L=1}^{N} \delta(s_L - s_{\mathrm{T}}), \tag{12}$$

where s_L is the stimulus orientation at the Lth location, and δ denotes the Dirac delta function. Equation (12) illustrates that each stimulus in the display only takes the value s_{T}.

Case 2: All Stimuli Are Distractors

This case applies to the target-absent trials ($C = -1$) in the target detection tasks (Experiments 8 and 11). We write $p(\mathbf{s}|C)$ as a marginalization over the N-dimensional vector of distractor orientations $\mathbf{s}_{\mathrm{D}}$:

$$p(\mathbf{s}|C=-1) = \int p(\mathbf{s}|\mathbf{s}_{\mathrm{D}})p(\mathbf{s}_{\mathrm{D}})d\mathbf{s}_{\mathrm{D}}.$$

Because all stimuli are distractors, we can evaluate $p(\mathbf{s}|\mathbf{s}_{\mathrm{D}})$ as:

$$p(\mathbf{s}|\mathbf{s}_{\mathrm{D}}) = \prod_{L=1}^{N} \delta(s_L - s_{\mathrm{D}L}),$$

where $s_{\mathrm{D}L}$ is the distractor orientation at the Lth location. The distractor distribution $p(\mathbf{s}_{\mathrm{D}})$ is different between Experiment 8 (homogeneous distractors) and 11 (heterogeneous distractors).

Case 3: Exactly One Stimulus Is the Target

This case applies to all trials ($C = \pm 1$) in five target categorization tasks (Experiments 5, 6, 7, 9, and 10) and target-present trials ($C = 1$) in the target detection tasks (Experiments 8 and 11). We write $p(\mathbf{s}|C)$ as a marginalization over both the scalar target orientation s_{T} and the (N-1)-dimensional vector of distractor orientations $\mathbf{s}_{\mathrm{D}}$:

$$p(\mathbf{s}|C) = \iint p(\mathbf{s}|s_{\mathrm{T}}, \mathbf{s}_{\mathrm{D}})p(s_{\mathrm{T}}|C)p(\mathbf{s}_{\mathrm{D}})ds_{\mathrm{T}}d\mathbf{s}_{\mathrm{D}}. \tag{13}$$

Here, $p(\mathbf{s}|s_{\mathrm{T}}, \mathbf{s}_{\mathrm{D}})$ represents the distribution of the stimulus vector $\mathbf{s}$ given s_{T} and $\mathbf{s}_{\mathrm{D}}$. Given that exactly one stimulus is the target, we can write this distribution as:

$$p(\mathbf{s}|s_{\mathrm{T}}, \mathbf{s}_{\mathrm{D}}) = \frac{1}{N}\sum_{L=1}^{N} \delta(s_L - s_{\mathrm{T}})\delta(\mathbf{s}_{\backslash L} - \mathbf{s}_{\mathrm{D}}), \tag{14}$$

where $\mathbf{s}_{\backslash L}$ denotes the vector of all stimuli except the Lth one. In Equation (14), the average over L represents another example of Bayesian marginalization: the observer does not know which stimulus is the target and therefore has to consider all possibilities.

The target distribution $p(s_{\mathrm{T}}|C)$ and the distractor distribution $p(\mathbf{s}_{\mathrm{D}})$ differ between experiments.

Case 4: The Stimulus on the Left Is the Target, While the Stimulus on the Right Is the Reference

This case applies to all trials ($C = \pm 1$) in the categorization task of Experiment 2. We write $p(\mathbf{s}|C)$ as a marginalization over both the scalar target orientation s_{T} and the scalar reference orientation s_{ref}:

$$p(\mathbf{s}|C) = \iint p(\mathbf{s}|s_{\mathrm{T}}, s_{\mathrm{ref}})p(s_{\mathrm{T}}|s_{\mathrm{ref}}, C)p(s_{\mathrm{ref}})ds_{\mathrm{T}}ds_{\mathrm{ref}}, \tag{15}$$

The stimuli $\mathbf{s} = (s_{\mathrm{left}}, s_{\mathrm{right}})$ are completely determined by the values of s_{T} and s_{ref}, and therefore $p(\mathbf{s}|s_{\mathrm{T}}, s_{\mathrm{ref}})$ is

$$p(\mathbf{s}|s_{\mathrm{T}}, s_{\mathrm{ref}}) = \delta(s_{\mathrm{left}} - s_{\mathrm{T}})\delta(s_{\mathrm{right}} - s_{\mathrm{ref}}). \tag{16}$$

Finally, $p(s_{\mathrm{ref}})$ is a uniform distribution and $p(s_{\mathrm{T}}|s_{\mathrm{ref}}, C)$ is a Gaussian distribution centered at s_{ref}, truncated to either half depending on C.

The worked-out Bayesian decision rules for all experiments are given in Appendix A. In models with the oblique effect (O), we assume that the observer knows the noise level σ when evaluating the decision variable d (Equation 7), but does not "realize" that there is a relationship between σ and orientation, s. Therefore, the observer does not infer s from σ, for example by marginalizing over σ. For a more principled examination of the implications of heteroskedasticity for Bayesian observer models, see Wei and Stocker (2015).

Decision noise (D). Decision noise (D) has been modeled using a softmax function (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Soltani & Wang, 2006), as Gaussian noise on the decision criterion (Mueller & Weidemann, 2008), or as Gaussian noise on the log posterior ratio (Drugowitsch et al., 2016; Keshvari et al., 2012, 2013). Here, we use the last approach: The decision variable $\tilde{d}$ follows a Gaussian distribution with a mean of d (Equation 7) and a standard deviation of σ_{d}:

$$(\tilde{d}|d) = N(\tilde{d};\, d,\, \sigma_{\mathrm{d}}^2). \tag{17}$$

The observer reports $C = 1$ if $\tilde{d}$ is positive.

Step 3: Predictions: Sampling of Measurements and Guessing Rate

Step 2 produces a mapping from a set of measurements, **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 **s** are the physical stimuli on that trial. This distribution is obtained as an average (marginalization) over measurement vectors **x**:

$$p(\hat{C}|\mathbf{s}) = \int p(\hat{C}|\mathbf{x})p(\mathbf{x}|\mathbf{s})d\mathbf{x}. \tag{18}$$

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 **x** based on the physical stimuli **s** on that trial. For each **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 2,000 samples.

Guessing (G). 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

$$p_{\text{with guessing}}(\hat{C}|\mathbf{s}) = 0.5\lambda + (1 - \lambda)p(\hat{C}|\mathbf{s}). \tag{19}$$

0.5 comes from the assumption that guesses are equally distributed across the responses.

Factorial model comparison. We will denote the factors by G, O, D, and V (see Table 2). We tested these factors in a factorial manner (Acerbi, Vijayakumar, & Wolpert, 2014; van den Berg et al., 2014), and got 16 models including all combinations of factor presence and absence. We will denote each model by the combinations of factors in the model. For example, GDV has all factors except for O. In some experiments, we will combine these combinations with both optimal versus suboptimal decision rules, but in most experiments, we will only consider the optimal decision rule.

Table 2
Models, Model Factors, and Parameters

Model or model factor	Corresponding parameter(s)
Model: Base	Category prior: p_{prior} Precision: J
Factor: Guessing, G	Guessing rate: λ
Factor: Oblique effect, O	Amplitude parameter of orientation dependence: β
Factor: Decision noise, D	Decision noise: σ_{d}
Factor: Residual variable precision, V	Scale parameter: τ
Model: Full or GODV	All parameters above

Note. The base model has parameters J and $\mathrm{p}_{\text{Prior}}$. G, O, D, and V denote the model factors that can be added to the base model, each with an associated parameter. The full or GODV model is obtained by adding all four factors. In each model, we fitted all parameters on an individual-subject basis.

Modeling Methods

Model Fitting

We fitted the free parameters in each model (Table 2) 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 jth trial, $\hat{C}_j$ is the subject's response on the jth trial, and we have assumed that there are no sequential dependencies between trials. The probability of the subject response, $\log p(\hat{C}_j\,|\,\mathbf{s}_j, M, \text{parameters})$, is obtained from Equations (18) or (19). To find the values of parameters that maximize log L_M(parameters), we used Bayesian Adaptive Direct Search (BADS; Acerbi & Ma, 2017), initialized with random values for all parameters. 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, $\mathrm{LL}_{\max}(M)$. As a sanity check, we found that those parameter combinations that gave the $\mathrm{LL}_{\max}(\mathrm{M})$ were in a reasonable range (Appendix D).

Model Comparison Metrics

We use the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) as metrics of badness of fit. These metrics penalize a model for having more free parameters:

$$\begin{aligned}\mathrm{AIC}(M) &= 2k_M - 2\mathrm{LL}_{\max}(M),\\ \mathrm{BIC}(M) &= \log(N_{\text{trials}})k_M - 2\mathrm{LL}_{\max}(M),\end{aligned}$$

where k_M is the number of parameters of model M and N_{trials} is the number of trials. Both AIC and BIC have their own advantages and disadvantages (Burnham & Anderson, 2002, pp. 293–305, for a pro-AIC account; Kass & Raftery, 1995, for a pro-BIC account). 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 only draw conclusions when both metrics provide strong evidence (see section "Jeffreys' scale"). We use AIC/BIC and the corresponding factor importance metrics (below) for all formal conclusions. For some models in some experiments, we also show fits to the psychometric curves, but these are only meant as qualitative visual checks of the relative and absolute goodness of fit of the models.

Factor Importance Metrics

We consider four model factors: guessing (G), the oblique effect (O), decision noise (D), and "residual" variable precision (V). Each can have two levels (absent and present), for a total of 16 models. We would like to draw conclusions about the importance of each factor regardless of model. We are also interested in the combination of O and V, because both are forms of variable precision; in the context of factor importance metrics, we will for brevity also refer to this combination as a factor.

In van den Berg, Awh, and Ma (2014), we quantified factor importance by calculating the proportion of subjects for whom all models in a given "model family" (e.g., all models in which G is absent) are rejected (according to 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; and (b) it outputs a curve (function) rather than a number. Therefore, we introduce three new factor importance metrics here (Figure 3): knock-in difference (representing evidence for factor usefulness), knock-out difference (representing evidence for factor necessity), and log factor likelihood ratio (representing evidence for factor presence). The terms "useful" and "necessary" only refer to goodness of fit, not to usefulness or necessity to the observer.

Factor Usefulness: Knock-In Difference (KID)

We measure the evidence that a factor is useful as the amount by which the goodness of fit improves relative to the base model by adding, or "knocking in," that factor (Figure 3A). We define the *knock-in difference based on AIC* (KID_{AIC}) of a factor F(which takes values G, O, D, V, or OV) as the AIC difference between the Base model and the knock-in model with F, denoted by "Base + F":

$$\text{KID}_{\text{AIC}}(F) = \text{AIC}(\text{Base}) - \text{AIC}(\text{Base} + F). \tag{20}$$

A positive KID_{AIC} means that the knock-in model fits better than the base model, and represents evidence that the factor is useful. Drugowitsch, Wyart, Devauchelle, and Koechlin (2016) applied a similar analysis.

We call KID a measure of the "usefulness" of a factor and not of its "sufficiency," because we take "insufficiency" of a model to refer to the deviation between the model and the true distribution (as estimated using deviance, Wichmann & Hill, 2001; or Kullback-Leibler divergence, Shen & Ma, 2016), which is not what we quantify here.

Factor Necessity: Knock-Out Difference (KOD)

We measure the evidence that a factor is necessary as the amount by which the goodness of fit of the full model (GODV) decreases by lesioning, or "knocking out," that factor (Figure 3B). We define the *knock-out difference based on AIC* (KOD_{AIC}) of a factor F as the AIC difference between the corresponding knock-out model (ODV, GOV, GDV, GOD, or GD, denoted by "Full-F") and the full model:

$$\text{KOD}_{\text{AIC}}(F) = \text{AIC}(\text{Full-}F) - \text{AIC}(\text{Full}). \tag{21}$$

A positive KOD_{AIC} means that the knock-out model fits worse than the full model, and represents evidence that the factor is necessary.

Factor Presence: Log Factor Likelihood Ratio (LFLR)

Finally, we estimate the evidence that a factor is present in the true model underlying a subject's behavior as the log likelihood ratio of a factor being present versus absent, which we will refer to as the log factor likelihood ratio (LFLR; Figure 3C). Although this quantity reflects most objectively the degree of belief in a factor (Van Horn, 2003), additional assumptions are needed to estimate it. To find the marginal likelihood that a factor F is present, we marginalize over all models M in the model space:

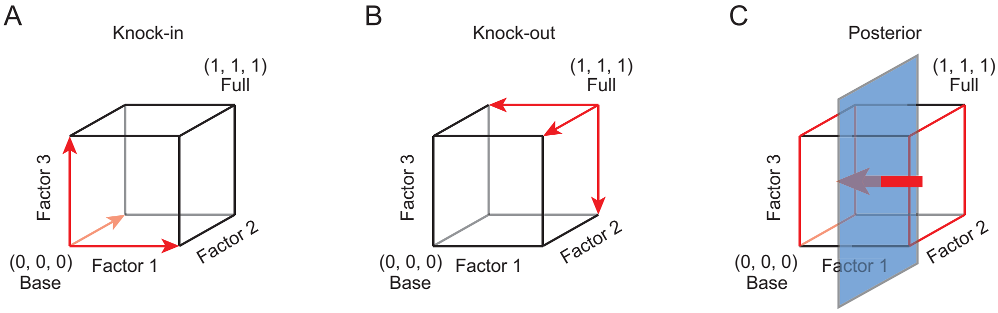


Figure 3. Factor importance metrics. In each diagram, each dimension represents a binary factor and each vertex a model; we show an example with 3 factors and thus a total of 8 models. The Base model, with none of the factors, is (0, 0, 0) and the Full model, with all factors, is (1, 1, 1). (A) Knock-in difference (KID, red arrows): the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) difference between the Base model (0, 0, 0) and the knock-in model with each single factor. (B) Knock-out difference (KOD, red arrows): the AIC or BIC difference between the corresponding knock-out model and the Full model (1, 1, 1). (C) The log factor likelihood ratio (LFLR). We compute the log likelihood ratio of a factor being present versus absent by marginalizing over all models with or without that factor, respectively. See the online article for the color version of this figure.

$$L(F\text{ present}) \equiv p(\text{data}|F\text{ present}) = \sum_{\text{all models } M} p(\text{data}|M)p(M|F\text{ present}).$$

Next, we assume that the models we tested are representative, so that the sample average is a good approximation of the theoretical average:

$$L(F\text{ present}) \approx \sum_{M\text{ tested}} p(\text{data}|M)p(M|F\text{ present}).$$

Next, we assume that all models containing *f* are a priori equally probable, so that

$$L(F\text{ present}) \approx \frac{1}{\text{number of tested models that contain } F} \sum_{M\text{ tested containing } F} p(\text{data}|M).$$

Finally, we approximate the log marginal likelihood of a given model by −0.5 times the AIC of that model (Akaike, 1978, 1979; Burnham & Anderson, 2002, Chapter 2.9):

$$L(F\text{ present}) \approx \frac{1}{\text{number of tested models that contain } F} \sum_{M\text{ tested containing } F} e^{-0.5\,AIC(M)}. \tag{22}$$

We analogously define the marginal likelihood of factor absence, and then the *log factor likelihood ratio based on AIC* (LFLR_{AIC}) as

$$\text{LFLR}_{\text{AIC}}(F) \equiv \log\frac{p(\text{data}|F\text{ present})}{p(\text{data}|F\text{ absent})} \approx \log\frac{\sum_{M\text{ tested containing } F} e^{-0.5\,AIC(M)}}{\sum_{M\text{ tested containing } F} e^{-0.5\,AIC(M)}}, \tag{23}$$

where we have assumed that equal numbers of tested models contain and do not contain F, as is the case throughout this paper. LFLR_{AIC} is similar to the log evidence ratio of AIC weights (E.-J. Wagenmakers & Farrell, 2004), except that the latter is an estimate of the log likelihood ratio between two models, instead of between factor presence and absence.

We now discuss an important special case. If adding a factor does not improve the unpenalized goodness of fit, which means that the model containing the factor has the exact same LL_{max} as the corresponding model without that factor, then its LFLR_{AIC} is:

$$\text{LFLR}_{\text{AIC}}(F) \approx \log\frac{\sum_{M:F=1} e^{\text{LL}_{\text{max}}(M)-k_M}}{\sum_{M:F=0} e^{\text{LL}_{\text{max}}(M)-k_M}} = -1. \tag{24}$$

Therefore, the LFLR_{AIC} of a factor should always be higher than −1, but in practice, it is possible to be slightly lower because of the simulation noise.

Similarly, we could also use −0.5 BIC as an approximation of log marginal likelihood. Then, Equations (20) to (23) would be analogous, and the lower bound of LFLR_{BIC} in Equation (24) would be −4.0 (Experiments 1–7, 9, and 10), or −3.6 (Experiments 8 and 11), depending on the number of trials. In practice, BIC penalizes extra parameters by more than AIC. Therefore, LFLR_{BIC} is generally lower than LFLR_{AIC} for the same factor.

To facilitate comparison with KID and KOD, we will report the value of 2 · LFLR rather than LFLR itself, because KID and KOD are computed with AIC or BIC while LFLR is computed with 0.5 AIC or 0.5 BIC. For each metric, a higher positive value means more evidence that the factor is important.

Relation Between KID and KOD

While the KID and KOD metrics each have a relatively straightforward interpretation, the question arises what inconsistency between them could mean. Finding that a factor is important in KID but not in KOD could indicate a "trade-off" between factors, or a logical OR operation: The effect of knocking out one factor is compensated for by other factors, to yield an equally good fit. Such a "trade-off" between factors is an example of model mimicry (Townsend, 1972; E. J. Wagenmakers, Ratcliff, Gomez, & Iverson, 2004) and would go away in the limit of infinite data. The opposite is also possible: A factor is important in KOD, but not in the KID. This could indicate an "interaction" between factors or a logical AND operation: Neither factor is useful by itself, but their combination is, similar to finding an interaction without main effects in ANOVA.

Relation Between LFLR and KID/KOD

In general, we expect LFLR to be more closely related to KOD than to KID. This is because the log-sum-exponent operation in the calculation of LFLR, Equation (23), is similar to a max operation (Ma, Shen, Dziugaite, & van den Berg, 2015). Thus, the marginal likelihoods of factor presence and absence will often be dominated by the best models with and without the factor, respectively. Take LFLR_{AIC} as an example. Starting from Equation (23),

$$\text{LFLR}_{\text{AIC}}(F) \approx \log\frac{\sum_{M:F=1} e^{-0.5\,AIC\,(M)}}{\sum_{M:F=0} e^{-0.5\,AIC\,(M)}} \approx \log\frac{\max_{M:F=1} e^{-0.5\,AIC\,(M)}}{\max_{M:F=0} e^{-0.5\,AIC\,(M)}}.$$

If furthermore, the lowest-AIC model is the most highly parameterized model, then $2 \cdot \text{LFLR}_{\text{AIC}}$ becomes identical to KOD_{AIC}.

Jeffreys' Scale

To interpret the numerical values of the factor importance metrics, we use Jeffreys' scale (Jeffreys, 1961; Table 3), which is

Table 3
Jeffreys' Scale for Bayes' Factors (Jeffreys, 1961)

Bayes factor (BF)	2 · log(BF)	Interpretation
>100	>9.2	Extreme evidence for H1
30 to 100	6.8 to 9.2	Very strong evidence for H1
10 to 30	4.6 to 6.8	Strong evidence for H1
3 to 10	2.2 to 4.6	Moderate evidence for H1
1 to 3	0 to 2.2	Anecdotal evidence for H1
1	0	No evidence
1/3 to 1	−2.2 to 0	Anecdotal evidence for H0
1/10 to 1/3	−4.6 to −2.2	Moderate evidence for H0
1/30 to 1/10	−6.8 to −4.6	Strong evidence for H0
1/100 to 1/10	−9.2 to −6.8	Very strong evidence for H0
<1/100	<−9.2	Extreme evidence for H0

commonly used to interpret Bayes factors. (One could make the case that such categorization is unnecessary, but people are easily seduced by categories.) We make a few modifications (Table 4): (a) to be conservative, we are more careful with our adjectives than Jeffreys; (b) also to be conservative, we base our interpretation on the lowest of the *F*actor *I*mportance *M*etrics $\mathrm{FIM}_{\mathrm{AIC}}$ and $\mathrm{FIM}_{\mathrm{BIC}}$; and (c) our scale is not symmetric between positive and negative values, because FIM has a lower bound due to being based on AIC or BIC (see Equation 24).

Results

Motivated by the visual STM literature, we searched for evidence for variable precision in 11 experiments, most of which used visual search tasks. In doing so, we tested for three factors that could be confounded with variability in precision, namely guessing (G), the oblique effect (O), decision noise (D), and in some cases, suboptimal decision rules. Our approach relies on quantitative model comparison, the results of which we summarize through three novel "factor importance metrics," each crossed with AIC and BIC.

The Importance of Variable Precision (V) When Taking Into Account 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). In the current study, we started searching for evidence for variable precision by only considering G. We get four models: base model with no factors (base), variable precision model (V), fixed precision with guessing (G), and variable precision with guessing (GV). We applied all six factor importance metrics to this model set.

In most experiments (Experiments 1, 3, 5, 6, 7, 9, 10, and 11), mean $\mathrm{KID}_{\mathrm{AIC}}(\mathrm{V})$ and mean $\mathrm{KID}_{\mathrm{BIC}}(\mathrm{V})$ were both greater than 9.2 (Figure 4A), indicating very strong evidence that factor V is useful to explain the data. This is consistent with the model fits to the psychometric curves. (Figure 4D, Figures B1, B3, B5, B6, B7, B9, B10, panel B in Appendix B, compare the base and V models. The numbers of the Figures 1 to 11 in Appendix B correspond to the experiment numbers). In Experiments 2 and 4, $\mathrm{KID}_{\mathrm{AIC}}(\mathrm{V})$ and $\mathrm{KID}_{\mathrm{BIC}}(\mathrm{V})$ were much smaller, indicating little or no evidence that factor V is useful.

In Experiments 7, 9, 10, and 11, mean $\mathrm{KOD}_{\mathrm{AIC}}(\mathrm{V})$ and mean $\mathrm{KOD}_{\mathrm{BIC}}(\mathrm{V})$ were both greater than 9.2, indicating very strong evidence that factor V is necessary to explain the data (Figure 4B). This is consistent with the model fits to the psychometric curves (Figure 4E, Figures B7, B9, B10, and panel B in Appendix B, compare the G and GV models). In Experiment 8, KOD(V) was 12.1 ± 4.3 (AIC) and 6.9 ± 4.3 (BIC), indicating strong evidence that factor V is necessary (Figure 4B). In Experiments 1–6, however, mean $\mathrm{KOD}_{\mathrm{AIC}}(\mathrm{V})$ and mean $\mathrm{KOD}_{\mathrm{BIC}}(\mathrm{V})$ were both lower than 4.6, indicating little or no evidence that factor V is necessary. The large difference between KID(V) and KOD(V) in Experiments 1, 3, 5, and 6 arose because factor G could also explain the data well, as indicated by a high KID(G; Figure 4A) and illustrated by the model fits (Figures 4D, A1, A3, A5, A6).

Using LFLR, we found very strong evidence for the presence of factor V in Experiments 7, 9, 10, and 11, with mean $2 \cdot \mathrm{LFLP}_{\mathrm{AIC}}(\mathrm{V})$ and mean $2 \cdot \mathrm{LFLP}_{\mathrm{BIC}}(\mathrm{V})$ both greater than 9.2 (Figure 4C). We also found strong evidence for the presence of factor V in Experiment 8, with a $2 \cdot \mathrm{LFLP}(\mathrm{V})$ equal to 12.7 ± 4.4 (AIC) and 7.8 ± 4.5 (BIC). The common feature of these five experiments was that distractors were variable across trials. In Experiments 7 and 8, the distractors were homogenous within a trial, while in Experiments 9–11, distractors were heterogeneous within a trial. We found little or no evidence for factor V in Experiments 1–6, with mean $2 \cdot \mathrm{LFLP}_{\mathrm{AIC}}(\mathrm{V})$ and mean $2 \cdot \mathrm{LFLP}_{\mathrm{BIC}}(\mathrm{V})$ both smaller than 4.6. The common feature of these experiments was that there were either no distractors (Experiments 1–4), or fixed distractors (Experiments 5–6).

Overall, with consideration of guessing, we found evidence for the presence of variable precision to be very little in Experiments 1–6, strong in Experiment 8, and very strong in Experiment 7 and 9–11.

How do these results compare to the visual STM literature? The analog of a comparison between the G and V models has been made in several visual STM experiments (D. T. Devkar et al., 2015; Keshvari et al., 2012, 2013; van den Berg et al., 2012), with V fitting better. In one article, GV was compared with G, with GV fitting better (Fougnie et al., 2012; van den Berg et al., 2014). Finally, in one article, a form of GV (there called VP-F) was compared with both V (VP-A) and a form of G (EP-F), with GV fitting best; however, the guessing was set size-dependent in a specific way (dictated by an item limit; Fougnie et al., 2012; van den Berg et al., 2014). All these results were obtained in experiments with multiple set sizes and "heterogeneous distractors," most similar to our Experiments 10 and 11. Indeed, the results are consistent with ours in those experiments.

The Importance of Variable Precision (V) When Taking Into Account Guessing (G) and the Oblique Effect (O)

Although we found evidence for variable precision (V) in Experiments 7–11 when considering G, some of the variability could be explained by other confounding factors. We first considered the oblique effect (O), the phenomenon that oblique orientations are encoded with lower precision than cardinal ones (Appelle, 1972; Girshick et al., 2011; Pratte et al., 2017). We implemented O using a rectified sine function (Figure 2B and Equation 4).

Inclusion of factor O did not qualitatively change the importance of factor V in Experiments 1–8 and 11 (Figure 5A–C). However, it greatly reduced the importance of factor V in Experiments 9 and 10. KOD(V) was 4.4 ± 3.3 (AIC) and −1.6

Table 4
Jeffreys' Scale Adapted for Our Three Factor Importance Metrics (FIMs)

$\min(\mathrm{FIM}_{\mathrm{AIC}}(F), \mathrm{FIM}_{\mathrm{BIC}}(F))$	Interpretation
>9.2	Very strong evidence that *F* is important
6.8 to 9.2	Strong evidence that *F* is important
4.6 to 6.8	Moderate evidence that *F* is important
<4.6 (including < 0)	Little or no evidence that *F* is important

Note. "Important" can mean "useful" (KID), "necessary" (KOD), or "present" (LFLR).

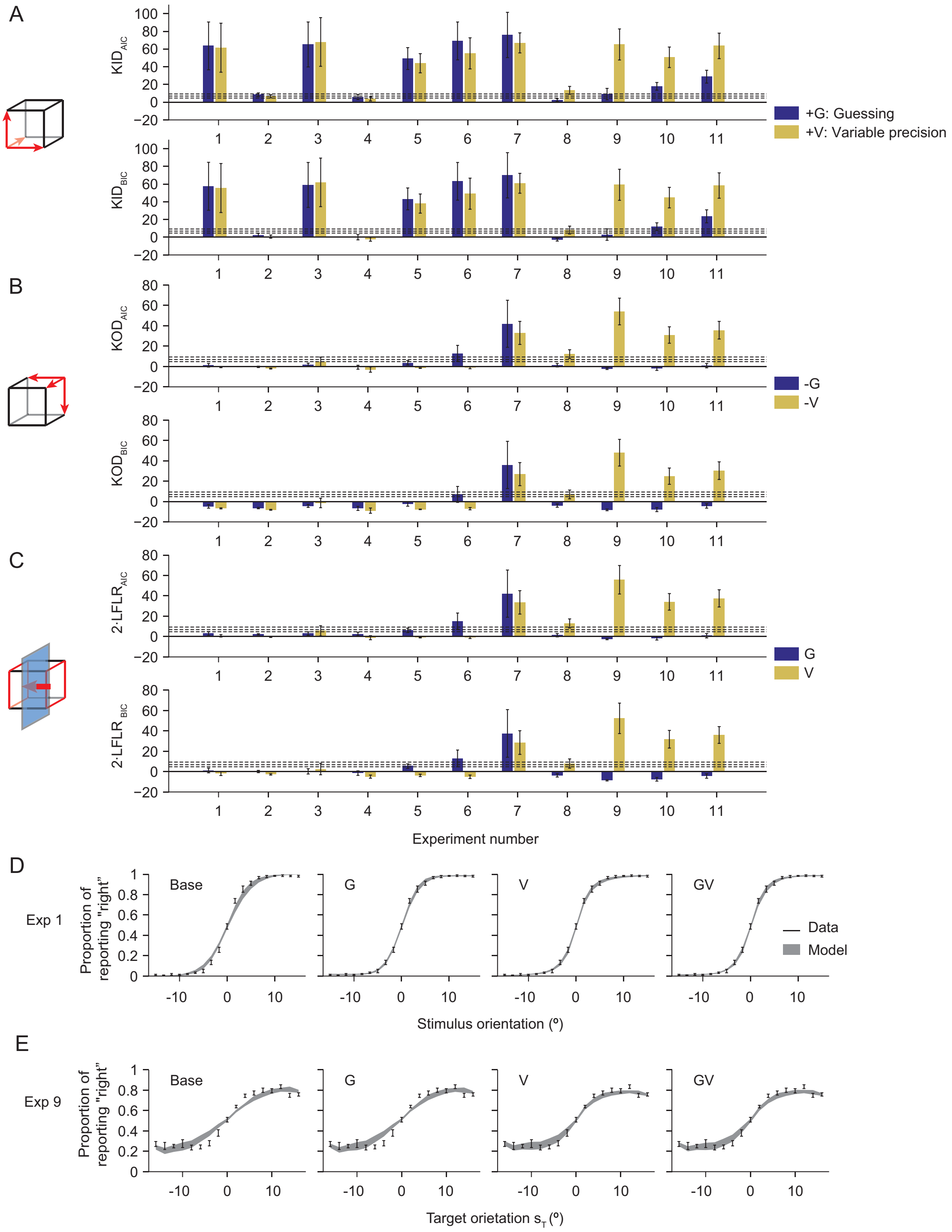


Figure 4. Factor importance: guessing (G) and variable precision (V). Here and in other factor importance plots, dashed black lines mark the boundaries of our interpretation of the strength of the evidence (>9.2: very strong, >6.8: strong, >4.6: moderate). (A–C) Mean and SEM of Knock-In Difference (KID) (A), Knock-Out Difference (KOD) (B), and 2 × LFLR (C) based on Akaike Information Criterion (AIC) (top) or Bayesian Information Criterion (BIC) (bottom) for the factors G and V in all experiments. (D) Model fits to the proportion of reporting "right" as a function of target orientation in Experiment 1. In all model fit plots, we use error bars and shaded areas to represent ±1 *SEM* in the data and the model fits, respectively. The G, V, and GV models fit the data equally well, and better than the Base model. (E) Model fits in Experiment 9. The V and GV models fit the data almost equally well, and better than the Base and G models. See the online article for the color version of this figure.

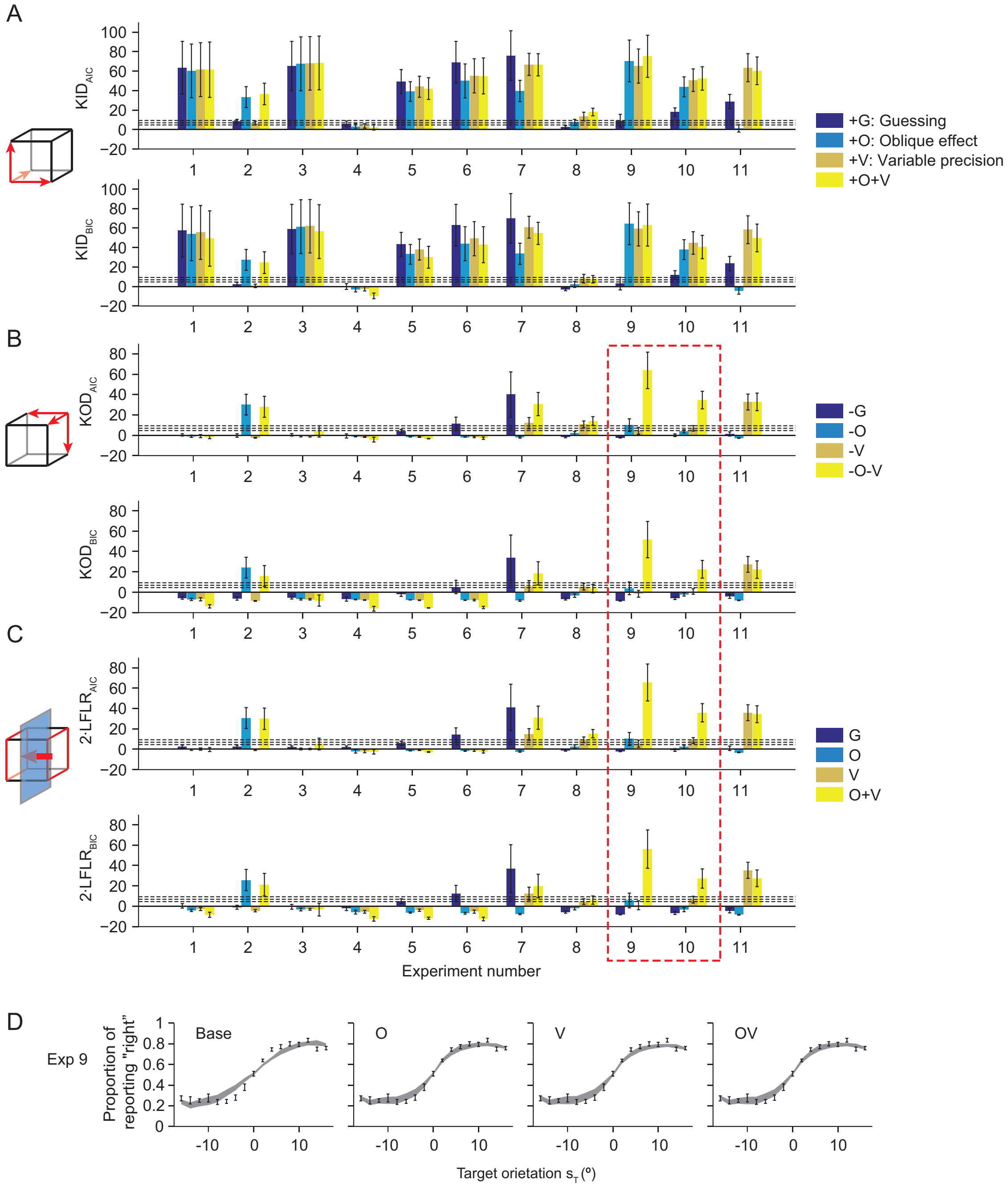


Figure 5. Factor importance among guessing (G), oblique effect (O), and residual variable precision (V). The red (grey) dashed box marks the major changes (compared with Figure 4) in the evidence for the importance of factor V when taking factor O into consideration. (A–C) Mean and *SEM* of Knock-In Difference (KID) (A), Knock-Out Difference (KOD) (B), and 2 · LFLR (C) based on Akaike Information Criterion (AIC; top) or Bayesian Information Criterion (BIC; bottom) for the factors G, O, V, and the OV combination, in all experiments. (D) Model fits in Experiment 9. The O, V, and OV models fit the data almost equally well, and better than the base model. See the online article for the color version of this figure.

± 3.3 (BIC) in Experiment 9, and 6.7 ± 2.7 (AIC) and 0.7 ± 2.7 (BIC) in Experiment 10 (Figure 5B), indicating little or no evidence that factor V is necessary. Consistently, 2 · LFLP(V) was 5.4 ± 3.5 (AIC) and 0.9 ± 4.4 (BIC) in Experiment 9 (Figure 5C), indicating little or no evidence for the presence of factor V. 2 · LFLP(V) was 8.5 ± 2.9 (AIC) and 6.4 ± 3.3 (BIC) in Experiment 10 (Figure 5C), indicating moderate evidence for the presence of factor V.

In Experiments 9 and 10, evidence for the presence of factor V was reduced because factor O could also explain the data well. KID(O) was 70 ± 21 (AIC) and 64 ± 21 (BIC) in Experiment 9, and 44 ± 10 (AIC) and 38 ± 10 (BIC) in Experiment 10 (Figure 5A). This can also be seen in the model fits of model O (Figure 5D). However, knocking out factor O did not cause large changes in AIC or BIC: KOD(O) was 9.8 ± 6.3 (AIC) and 3.8 ± 6.3 (BIC) in Experiment 9, and 3.7 ± 1.5 (AIC) and −2.3 ± 1.5 (BIC) in Experiment 10 (Figure 5B), indicating little or no evidence that factor O is necessary. Knocking out both factors O and V was disastrous, resulting in a KOD(OV) of 64 ± 18 (AIC) and 52 ± 18 (BIC) in Experiment 9, and of 34.5 ± 8.7 (AIC) and 22.5 ± 8.7 (BIC) in Experiment 10 (Figure 5B). This "nonlinear" phenomenon seemed to occur because factors O and V could stand in for each other in explaining the data, therefore neither was necessary, but having at least one of them was important. This is consistent with the model fits to the psychometric curves (Figure 5D, Appendix Figures B9A, B10B). Consistently, 2 · LFLR was not high for either factor O or factor V, but was high for their combination: 65 ± 18 (AIC) and 56 ± 19 (BIC) in Experiment 9, and 36.9 ± 9.0 (AIC) and 27.2 ± 9.4 (BIC) in Experiment 10 (Figure 5C).

In the visual STM literature, Pratte et al. (2017) compared models similar with G, V, GV, GO, OV, and GOV; again however, the guessing was set size-dependent in a specific way (dictated by an item limit). They found that V fitted better than G but worse than GV. Adding factor O flipped the first result: GO fitted better than OV; however, both still fitted much worse than GOV. Although they did not test the base and O models, this last result suggests evidence for the presence of V. The experiments by Pratte et al. (2017) were again most similar to our Experiments 10 and 11, featuring multiple set sizes and heterogeneous distractors. In Experiment 10, we also found that accounting for factor O changed our evidence for the presence of factor V (V fitted better than G but GO fitted as well as OV, Appendix B Figure B10), and (using LFLR) we found moderate evidence for the presence of V. In Experiment 11, considering factor O did not change our evidence for the importance of factor V, and we still found strong evidence for the presence of factor V.

The Importance of Variable Precision (V) When Taking Into Account Guessing (G), the Oblique Effect (O), and Decision Noise (D)

Besides guessing and the oblique effect, another confounding factor might be noise in the decision stage (or suboptimal inference, which can look like decision noise). We examined evidence for the importance of V when accounting for guessing (G), the oblique effect (O), and decision noise (D). We modeled D as Gaussian noise added to the log posterior ratio (Equation 17).

In Experiments 1–10, the inclusion of factor D did not change the evidence for the importance of factor V much (Figure 6A–C). In Experiment 8, evidence for the presence of factor V changed from strong to moderate (Figure 6C). In Experiment 10, the inclusion of factor D slightly reduced the evidence for the presence of factor V. 2 · LFLP(V) was 2.3 ± 2.3 (AIC) and 1.9 ± 3.3 (BIC), indicating little or no evidence for the presence of factor V (Figure 6C).

In Experiment 11, however, the inclusion of D greatly reduced the importance of factor V. KOD_{AIC}(V) and KOD_{BIC}(V) were negative (Figure 6B), indicating no evidence that factor V is necessary. Consistently, 2 · LFLP(V) was −0.6 ± 1.6 (AIC) and 0.7 ± 3.4 (BIC) in Experiment 11 (Figure 6C), indicating little or no evidence for the presence of factor V. The reason is probably that factor D can also explain the data: KID(D) was 60 ± 14 (AIC) and 55 ± 14 (BIC; Figure 6A), consistent with the model fits of the D model (Figure 7). However, knocking out factor D did not cause large KODs (Figure 6B), suggesting that factor D is useful, but not necessary.

In summary, when accounting for guessing, the oblique effect, and decision noise, we only found very strong evidence for the presence of the residual variable precision in Experiment 7. Experiment 7 was the only orientation categorization task in which the distractors were homogeneous but varied across trials. In Experiment 8, which was a target detection task also with homogeneous variable distractors, we found moderate evidence for residual variable precision, suggesting that homogeneous variable distractors might induce residual variable precision.

To our knowledge, no previous studies have compared models containing all four factors G, O, D, and V.

Relationship Between Task Features and Importance of Factors Other Than the Residual Variable Precision

The variation of the designs of our experiments also enables us to relate the features of tasks to the importance of factors other than the residual variable precision. The experiments differed in the following design features (Table 5): set size greater than 1 (divided attention), set size variability, number of targets greater than 1, task type (categorization 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 target-distractor category distributions).

By examining the importance of factors in all these 11 experiments, we found that some factors were important when certain features were present. We now summarize the evidence for the importance of each factor across experiments and attempt to make a connection to the features of the experiments; Table 5 lists the evidence for the presence of each factor (2 · LFLR) in each experiment.

Guessing (G)

Consistent with the notion that guessing is widespread in psychophysical tasks, we found that in many experiments (Experiments 1, 3, 5, 6, 7), KID(G) was greater than 9.2 (Figure 6A) and the G model provided clearly better fits to the psychometric curves than the base model (Figure 8A, Figures B1, B3, B5, B6, B7, panel B in Appendix B). Among these experiments, in Experiment 6, factor G was necessary (Figure 6B) and had 2 · LFLR of 13.1 ± 6.4 (AIC) and 12.3 ± 7.9 (BIC; Figure 6C), indicating very strong evidence for the presence of factor G. In this experiment, the target orientation took values between −20° and 20°, which was the largest range among all experiments. Moreover, set size could be low (1 or 2). A mistake on a trial with low set size and a strongly tilted target could only be explained by guessing. In Experiment 5, in which

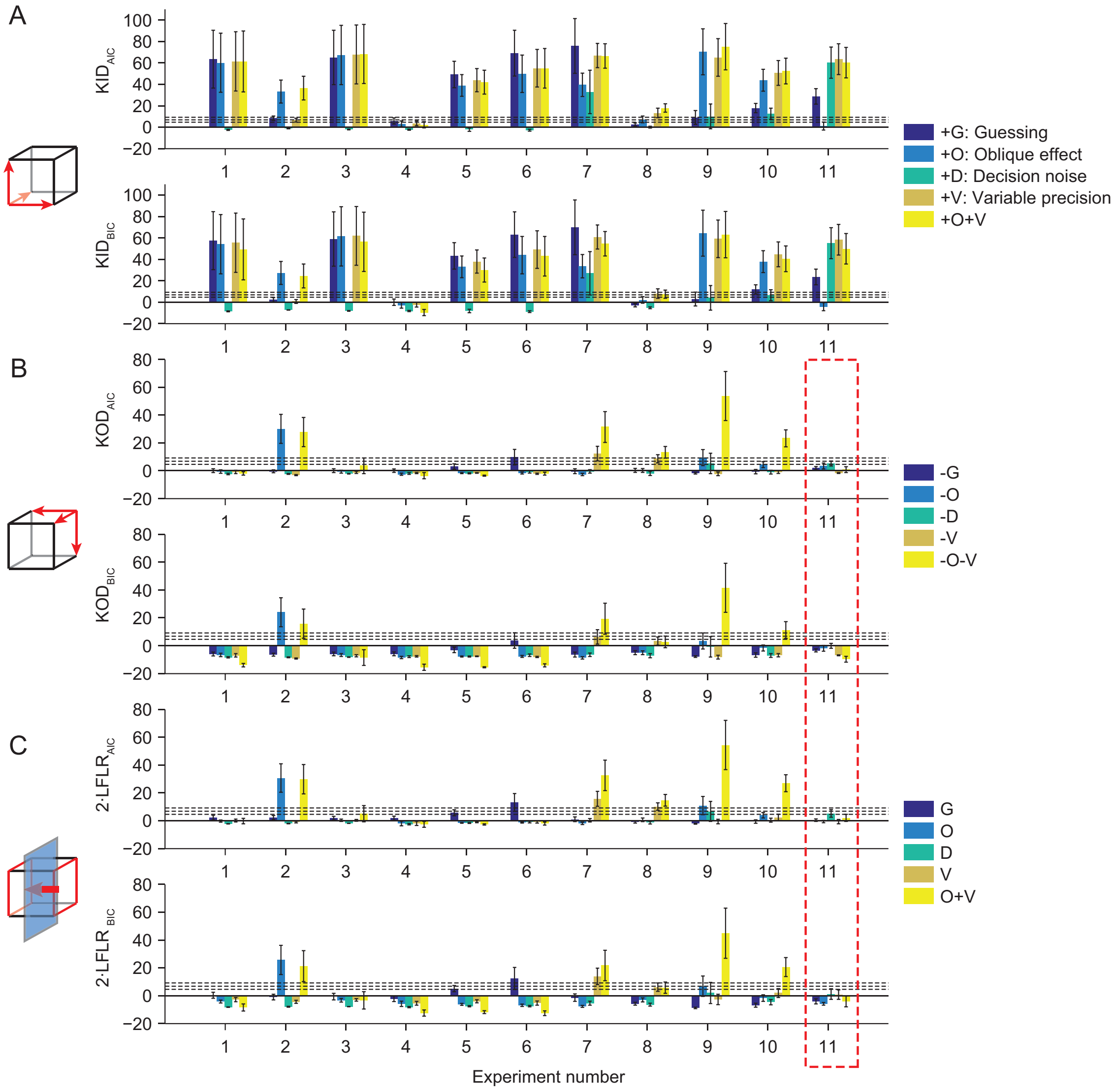


Figure 6. Factor importance: guessing (G), oblique effect (O), decision noise (D) and the residual variable precision (V). The red (grey) dashed box marks the major changes (compared with Figure 5) in the evidence for the importance of factor V when taking factor D into consideration. (A–C) Mean and *SEM* of Knock-In Difference (KID) (A), Knock-Out Difference (KOD) (B), and 2 · LFLR (C) based on Akaike Information Criterion (AIC; top) or Bayesian Information Criterion (BIC; bottom) for the factors G, O, D, V, and the OV combination, in all experiments. See the online article for the color version of this figure.

the stimulus range was the same but set size was equal to 4, factor G was no longer necessary, but 2 · LFLR(G) was 5.7 ± 2.1 (AIC), and 5.1 ± 2.2 (BIC), respectively, indicating moderate evidence for the presence of factor G. Across all experiments, it seems that the larger the proportion of easy trials, the higher the evidence for the importance of factor G. With fewer easy trials, models without factor G fitted the data equally well as models with factor G, by estimating a lower encoding precision (Figure 8A). For example, in Experiment 4, where the target orientation range was narrow (between −5° and 5°), the base model fitted as well as the G model, but the estimated precision was lower.

Oblique Effect (O)

We expected that factor O would be easier to detect when the stimulus distribution covered a larger orientation range. Indeed, we found very strong evidence for the presence of factor O in Experiment 2 and strong evidence in Experiment 9 (Figure 6C), in which the stimulus distribution covered the entire orientation space (Figure 1, Experiment 2, Experiment 9). This can also be seen in the model fits (Figure 8B). In Experiments 10 and 11, however, although the distractor distribution also covered the entire space, the evidence for factor O was weak. In Experiment 10, this might be because the experiment also contained a set size 1 condition, in

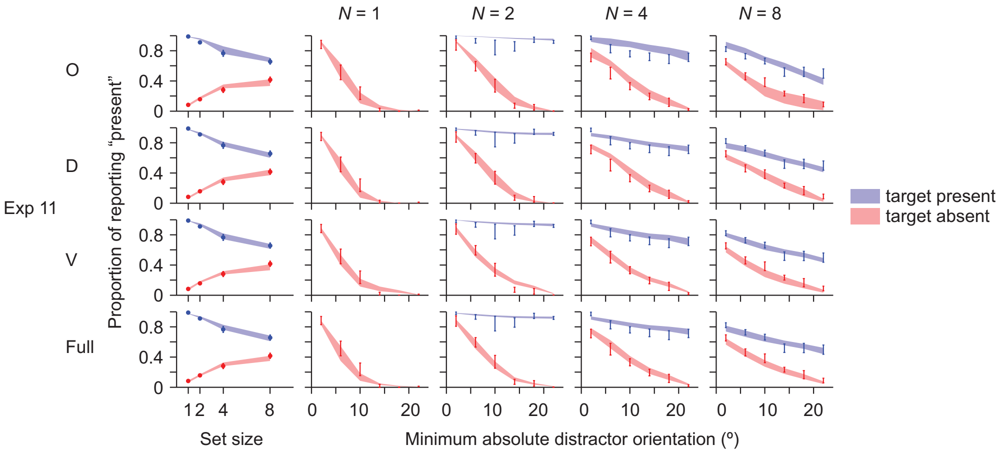


Figure 7. Model fits in Experiment 11. Proportion of reporting "target present" as a function of set size (left) and the smallest circular distance (right) in orientation space between the target and any of the distractors. Target present trials and target absent trials are shown with blue (dark grey) and red (light grey), respectively. The D and V models fit the data almost as well as the full model, and better than the O model. See the online article for the color version of this figure.

which there were no distractors. In Experiment 11, stimuli with large tilts were informative of factor O, but because of the task structure, these stimuli were weighted less in the optimal decision rule (Appendix A, Experiment 11), therefore perhaps making factor O harder to detect. Furthermore, because strongly tilted stimuli were less relevant to the task, subjects may have paid less attention to them. The weak evidence for the presence of factor O is consistent with previous findings that the oblique effect is weaker when stimuli are unattended (Kelly & Matthews, 2011; Takács, Sulykos, Czigler, Barkaszi, & Balázs, 2013).

Decision Noise (D)

Decision noise might reflect random variability or systematic suboptimality in the decision stage (Beck et al., 2012). We found little or no evidence for the presence of factor D in any of our experiments (Figure 6C). This is consistent with the conclusion of our previous article (Shen & Ma, 2016), where we compared many suboptimal decision rules with the optimal rule in an orientation categorization task (Experiment 7 in this article) and found that the more similar a suboptimal rule was to the optimal rule, the better it fitted the data. However, the lack of evidence for factor D might be a result of factor or parameter trade-off, which we will illustrate in the following section.

Causes for False Negatives

Overall, we were conservative when claiming evidence for a certain factor. Therefore, there were many "negative" results. Some of these results could be false negatives, where a factor was present but not detected. One potential source of false negatives is a lack of informative trials for an individual factor. For example, easy trials on the ends of the psychometric curve tend to be informative about the presence of guessing; thus, having too few easy trials may have prevented us from detecting guessing. Similarly, a narrow orientation range may have prevented us from detecting the oblique effect. A second potential source is trade-offs between parameters. For example, a nonzero guessing rate can be mimicked by a zero guessing rate and a lower (mean) precision parameter. To illustrate this trade-off, we generated 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 computed 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 (Figure 9A). In such a scenario, the $\mathrm{LL}_{\max}$ of the with-factor model (G) could be identical to the $\mathrm{LL}_{\max}$ 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 (Experiments 9 and 10) or D (Experiment 11). To illustrate this scenario, we generated a synthetic data set with the V model for Experiment 9, with a scale parameter $\tau = 0.05$, and computed 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 nonzero β and a smaller τ (Figure 9B). A smaller fitted τ indicated weaker evidence for factor V, because the data were partly explained by the factor O. Trade-offs can happen with any model comparison metric, but AIC and BIC are specifically known to be insensitive to trade-offs between factors (Gelman, Hwang, & Vehtari, 2014).

Suboptimality

So far, we have only considered the optimal decision rule in our models. However, the possibility that the observer uses a suboptimal decision rule must be considered, as it could change our conclusions.

In models based on signal detection theory, it is common to consider "simple heuristic" rules (although simplicity is hard to define). A well-known example is the max rule (Baldassi & Burr, 2000; Eckstein, 1998; Green & Swets, 1966; Nolte & Jaarsma, 1967; Palmer, 1990). A second way to construct suboptimal rules is more principled. An assumption behind the optimal rule is that

Table 5
Features of Experiments and the Evidence for the Presence of Factors

Experimental design								2·log factor likelihood ratio ($2\cdot\text{LFLP}_{\text{AIC}}$; $2\cdot\text{LFLP}_{\text{BIC}}$)					
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)	Oblique effect (O)	Decision noise (D)	"Residual" variable precision (V)
1	0	0	0	Catg	[−15, 15]	—	None	—	0	2.2 ± 1.7; .3 ± 2.2	−.32 ± .79; −4.2 ± 1.1;	−2.24 ± .12; −8.22 ± .19	.1 ± 1.0; −2.8 ± 1.5
2	0	1	0	Catg	N (s_{ref}, 9.1)	U (−90, 90)	High	Low	0	2.3 ± 1.5; −1.0 ± 2.0	**31 ± 10; 26 ± 10**	−2.03 ± .21; −7.93 ± .29	−.95 ± .57; −4.4 ± 1.1
3	0	1	1	Catg	[−15, 15]	—	None	Low	0	1.8 ± 1.3; −.7 ± 2.3	.28 ± .86; −3.3 ± 1.3	−1.90 ± .09; −7.86 ± .07	.23 ± .68; −3.00 ± .89
4	1	1	1	Catg	[−5, 5]	—	None	Low	0	1.8 ± 1.5; −2.2 ± 2.1	−2.1 ± 1.6; −5.8 ± 1.2	−2.48 ± .50; −8.22 ± .45	1.9 ± 1.5; −5.5 ± 1.4
5	0	1	0	Catg	[−20, 20]	δ (0)	None	Low	0	***5.7 ± 2.1; 5.1 ± 2.2***	−1.59 ± .21;−6.22 ± .72	−1.62 ± .32; −7.54 ± .40	−1.10 ± .39; −4.0 ± 1.1
6	1	1	0	Catg	[−20, 20]	δ (0)	None	Low	0	**13.1 ± 6.4; 12.3 ± 7.9**	−1.32 ± .39; −6.86 ± .84	−.97 ± .55; −7.44 ± .52	−.87 ± .87; −5.3 ± 1.5
7	0	1	0	Catg	N (0, 9.1)	N (0, 9.1)	Low	Low	0	1.1 ± 1.7; −1.4 ± 2.7	−1.93 ± .89; −7.69 ± .75	.3 ± 1.3; −5.4 ± 1.4	**15.7 ± 5.4; 13.8 ± 6.0**
8	1	1	0	Det	0	N (0, 5.1)	Low	Low	0	−.90 ± .77; −5.8 ± 1.1	.7 ± 1.3; −3.0 ± 1.4	−1.38 ± .94; −6.4 ± 1.1	***10.1 ± 2.7; 6.1 ± 3.0***
9	0	1	0	Catg	N (0, 9.1)	U (−90, 90)	High	High	1	−2.17 ± .31; −8.80 ± .56	*10.9 ± 6.5; 6.8 ± 7.4*	6.7 ± 7.2; 2.0 ± 7.6	−.6 ± 1.6; −2.7 ± 3.7
10	1	1	0	Catg	N (0, 9.1)	U (−90, 90)	High	High	1	−.7 ± 1.3; −6.8 ± 1.5	3.9 ± 2.0; −1.7 ± 2.4	.4 ± 1.4; −4.6 ± 1.9	2.3 ± 2.3; 1.9 ± 3.3
11	1	1	0	Det	0	U (−90, 90)	High	High	0	.41 ± .92; 4.1 ± 1.8	−.61 ± .84; −5.81 ± .99	5.3 ± 2.8; 1.2 ± 3.8	−.6 ± 1.6; .7 ± 3.4

Note. Catg = categorization; Det = detection; Ref = reference; S_{ref} = reference orientation; $U(a, b)$ denotes a continuous uniform distribution on the interval $[a, b]$; $N(s_0, \sigma)$ denotes a Gaussian distribution with a mean of s_0 and a standard deviation of σ; $\delta(a)$ denotes a dirac delta function at a. the unit of orientation is degrees, and we "converted" the (highly concentrated) Von Mises distributions to Gaussian distributions to make the comparison across experiments easier. Bold, italics, and bold and italic mark the *very strong*, *strong*, and *moderate* evidence for the presence of a factor, respectively, based on our criteria in Tables 3 and 4.

the observer has learned and correctly incorporates the experimental statistics, in our case the joint distribution $p(C, \mathbf{s})$. This assumption is common in Bayesian modeling of perception and often justified by arguing that the stimulus distribution is simple enough for subjects to learn quickly (e.g., Gaussian or uniform). The assumption is to some extent validated by the success of the resulting models (Geisler, 2011). However, in some cases, there is evidence that observers do not use accurate estimates of the parameters of the stimulus distribution (Acerbi, Ma, & Vijayakumar, 2014; Honig, Ma, & Fougnie, 2018), or instead use natural statistics (Adams, Graf, & Ernst, 2004; Girshick et al., 2011; Zhu & Ma, 2017).

In the present section, we undertake a limited exploration of both kinds of suboptimality.

Heuristic Decision Rules in Experiment 7

In the article where Experiment 7 was originally presented (Shen & Ma, 2016), we compared the optimal decision rule against 24 suboptimal rules, including "simple heuristic" rules such as the Max rule; however, of the four factors G, O, D, and V, we only included factor G. In the present article, we tested all factors, but so far assumed an optimal decision rule. Considering the suboptimal rules and all models in the "GODV family" simultaneously could in principle undermine the conclusions of both studies: First, if a suboptimal rule that fitted poorly in Shen and Ma (2016) were to fit the data substantially better when combined with a different model in the GODV family, that could change the conclusions of that article. Second, if a model without factor V in the GODV family were to fit the data substantially better when combined with a suboptimal rule, it would imply a change in the evidence for factor V in the current article.

To examine these possibilities, we crossed the suboptimal rules from Shen and Ma (2016) with all members of the GODV family in the current study, eliminating logically inconsistent combinations. This led to 292 extra models (Figure 10A, for a more detailed description, see Appendix C). The results confirmed the conclusions from Shen and Ma (2016) that human behaviors are closer to optimality than to simplicity in this task (Figure 10A, Appendix B Figure B12): Regardless of which GODV family member the decision rule was crossed with (a) simple rules (Class I and Class II) fitted the data worse than the optimal decision rules, with mean AIC or BIC differences greater than 40; and (b) the more similar a suboptimal rule was to the optimal rule, the better it fitted the data (Figure 10A). By contrast, considering the suboptimal decision rules in Experiment 7 changed the earlier conclusion of the current paper about the presence of factor V. We computed 2 × LFLR_{AIC}(V) and 2 × LFLR_{BIC}(V) by marginalizing over all models, including those with a suboptimal decision rule (Figure 10B). The evidence for the presence of factor V decreased to 2.1 ± 1.1 (AIC) and 0.3 ± 1.6 (BIC). Thus, the strong evidence we found for factor V in Experiment 7 disappeared when considering suboptimal decision rules.

A Heuristic Decision Rule in Experiments 9 and 10

In Experiments 9 and 10, the distribution of the target orientation was narrower than that of each distractor orientation. Therefore, an intuitive alternative to the optimal decision rule is to report the tilt of the least tilted stimulus. Following Shen and Ma (2016),

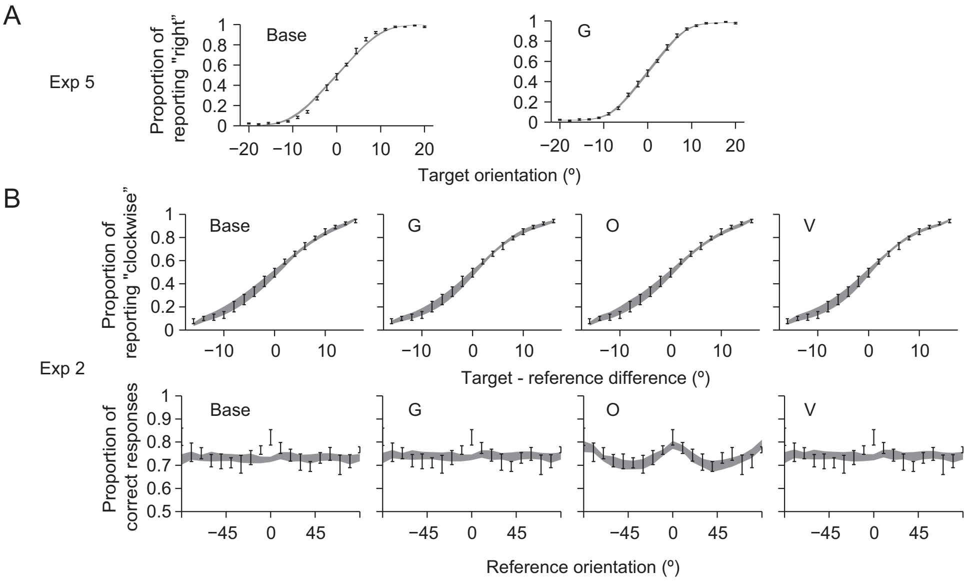


Figure 8. Model fits show that factor G and factor O are important in some experiments. (A) Model fits in Experiment 5 show the importance for factor G. The G model fits better than the Base model. (B) Model fits in Experiment 2 show the importance of factor O. Top: Proportion of reporting "clockwise" as a function of the orientation difference between the target and the reference, collapsed across reference orientations. Bottom: Proportion of reporting "clockwise" as a function of the reference orientation, collapsed across target orientations. The O model fits better than the Base, G, and V models.

we call this rule the "Min" rule. This rule is not just intuitive but can also be considered a Bayesian "two-step" rule: first pick the target by maximizing $p(L|\mathbf{x})$, where L is the hypothesized target location, and then report the tilt at the best location $\hat{L}$, which is equivalent to maximizing $p(C|\hat{L})$. It turns out that the best location $\hat{L}$ is the location of the least tilt stimulus. Here we will give an intuitive example to show the difference between the optimal rule and the Min rule. For a set of measurements [−5°, 2°, −15°, −85°], the Min rule would simply report "right" because of the 2° measurement. By contrast, the optimal rule would take into account both uncertainty over target location and uncertainty due to sensory noise. Uncertainty over target location: although −5° and −15° do not have the minimum tilt, it is well possible that one of them came from the target distribution. Uncertainty due to sensory noise: suppose $\sigma = 3°$. Then, even if the 2° measurement came from the target, the evidence that the target tilts right would not be very high. However, if −5° or −15° were the target, the evidence that the target tilts left would be higher. Therefore, the optimal decision-maker would report "left" in this case.

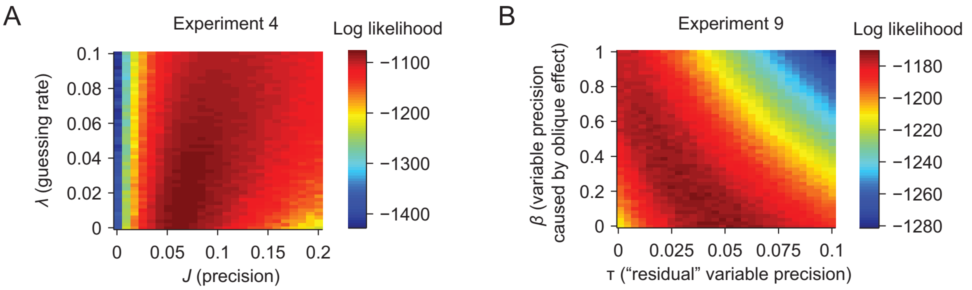


Figure 9. Trade-offs between parameters. (A) Trade-off between precision J and guessing rate λ. We generated a synthetic data set from the G model in Experiment 4 with $J = 0.08$ deg^{-2} and $\lambda = 0.02$, and fitted the data with the G model. The color plot shows the log likelihood of combinations of J and λ. Many combinations have a high log likelihood, including a combination of $\lambda = 0$ and a value of J lower than the true value. (B) Trade-off between the factors O (parameterized by β) and V (parameterized by τ). We generated a synthetic data set from the V model in Experiment 9, with $\tau = 0.05$ (and $\beta = 0$), and fitted the data with the OV model. The color plot shows the log marginal likelihood of combinations of β and τ. Many combinations have a high log likelihood, including a combination of non-zero β and a value of τ lower than the true value. See the online article for the color version of this figure.

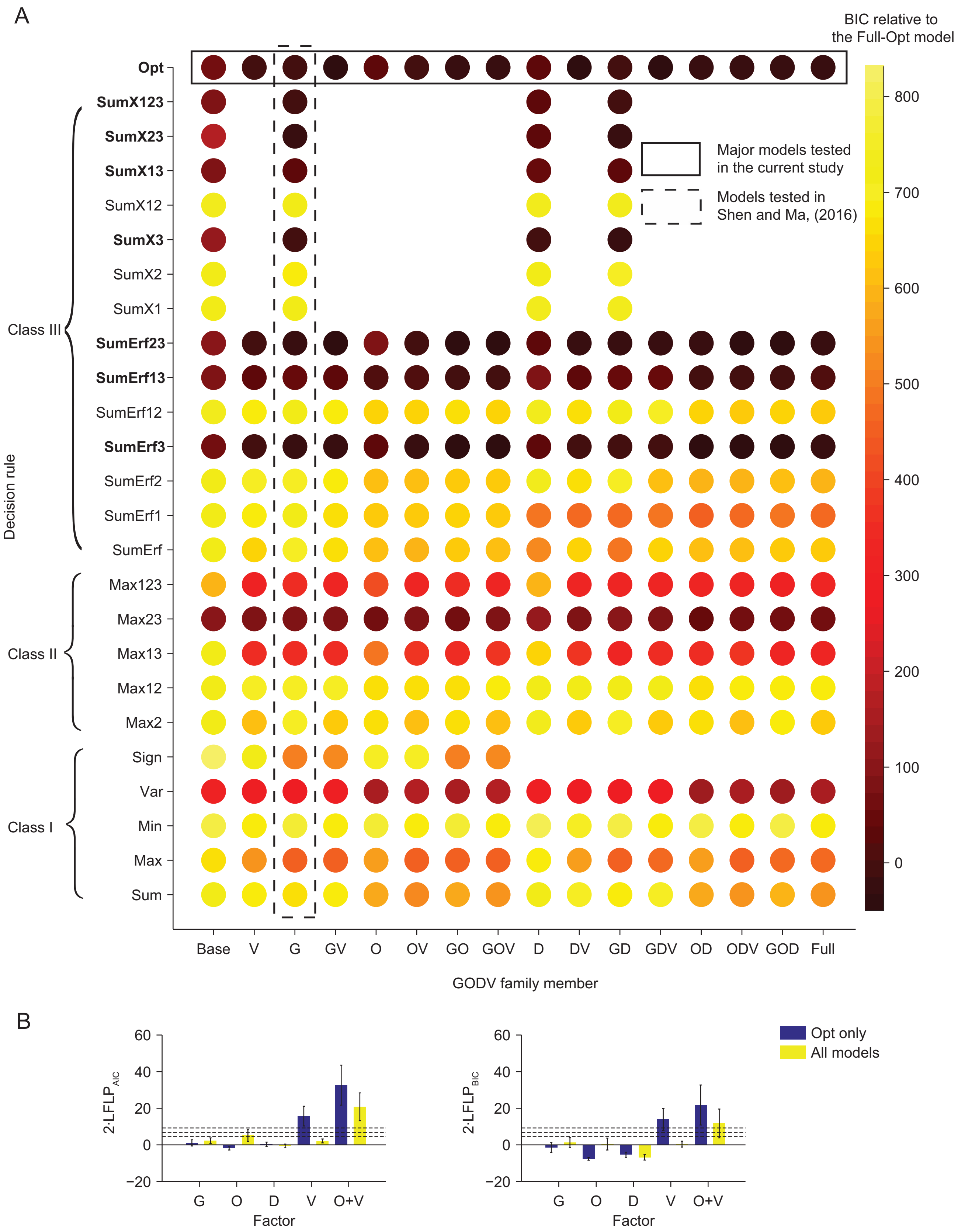


Figure 10(*opposite*)

We tested combinations between the Min rule and all models in the GODV family in both Experiment 9 and 10. We found that the Min rule fitted much worse than the optimal rule regardless of the GODV family member it was crossed with (Figure 11A, C). This result suggests that subjects do not use the Min rule in these two experiments. Consistently, considering both optimal rule and Min rule did not change the evidence for the presence of factors in Experiments 9 and 10 (Figure 11B, D).

Incorrect Assumption About the Stimulus Statistics in Experiment 5

In Experiments 1 and 3–6, where the orientation of the target was discrete, we did not model the prior over target orientation as the true target orientation distribution, because it is unlikely that subjects learned a dense, discrete distribution. Instead, we modeled the prior to be a zero-mean Gaussian distribution with the same standard deviation of the true distribution (Equation 6), which was a form of suboptimality. Alternatively, we could assume a boxcar prior of the target distribution that covered the orientation range of the true distribution. We tested all GODV family members with this prior and compared the results to the GODV family members with the Gaussian prior in Experiment 5. We found that both AIC and BIC were similar between the Gaussian prior and the boxcar prior (Figure 12A), and changing the prior did not change the evidence for the factors (Figure 12B). This result suggests that our conclusions are not sensitive to what prior we use when the target orientation is discrete.

Although our results in Experiment 5 were not sensitive to the class-conditioned stimulus distributions we assumed, we cannot in general rule out mismatch between the stimulus statistics assumed by the observer and the true ones, and we cannot rule out that such mismatch would affect our conclusions about the presence of the factors.

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, as estimated in the full model, decreased strongly with set size in Experiments 8, 10, and 11 (repeated-measures ANOVAs: $p < .05$); in these experiments, the distractors were variable across trials. There were no obvious differences between detection (Experiments 8 and 11) and categorization (Experiment 10). There was no significant effect of set size in Experiment 6, $F(3, 6) = 1.1$, $p = .38$, where the distractors were fixed at vertical (Figure 13). In Experiment 4, all stimuli were targets but with an orientation that was unpredictable across trials. Although performance increased with set size (Appendix B Figure B13A; $F(3, 6) = 7.25$, $p < .01$), because more stimuli gave more information about the correct answer, we found that mean precision *decreased* with set size (Figure 13; $F(3, 6) = 4.18$, $p = .013$). Given the weak evidence found for factor G in Experiment 4, we also estimated the precision with the ODV model, but the set size effect was similar (Appendix B Figure B13B: $F(3, 6) = 6.07$, $p <$ .01).

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 article (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 article 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. However, in all of this, one needs to keep in mind the possibility that the estimates of the precision parameters are affected by trade-offs with guessing (Figure 9A).

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 would lead to a weaker set size effect.

Discussion

Summary

We asked whether variable precision exists in visual perception. Specifically, we varied the complexity of the distractor context. We analyzed data from 11 visual experiments that used very similar oriented stimuli, and performed factorial model comparison with six factor importance metrics. Overall, we found little evidence for residual variable precision (V) when accounting for guessing (G), the oblique effect (O), and decision noise (D). In

Figure 10 (*opposite*). Crossing the suboptimal decision rules with the factor models in Experiment 7. (A) The *x*-axis lists GODV family members, and the *y*-axis lists different decision rules from Shen and Ma (2016). Decision rules marked in bold face are the rules similar to the Opt rule (Shen & Ma, 2016). The color of the dot represents the Bayesian Information Criterion (BIC) of a hybrid model with a certain decision rule and factor model. Some combinations are missing because those models are invalid (Appendix C). Akaike Information Criterion (AIC) version of the results is shown in Appendix Figure B12. (B) Mean and *SEM* of 2 · LFLR based on AIC (left) or BIC (right) for factors G, O, D, V, and the OV combination in Experiment 7. Blue bars: only models with the optimal decision rule are included. Yellow bars: all models except for those crossed with the Sign rule and SumX rules are included; we marginalized over decision rule in the same way as we marginalized over the "missing" GODV factors. See the online article for the color version of this figure.

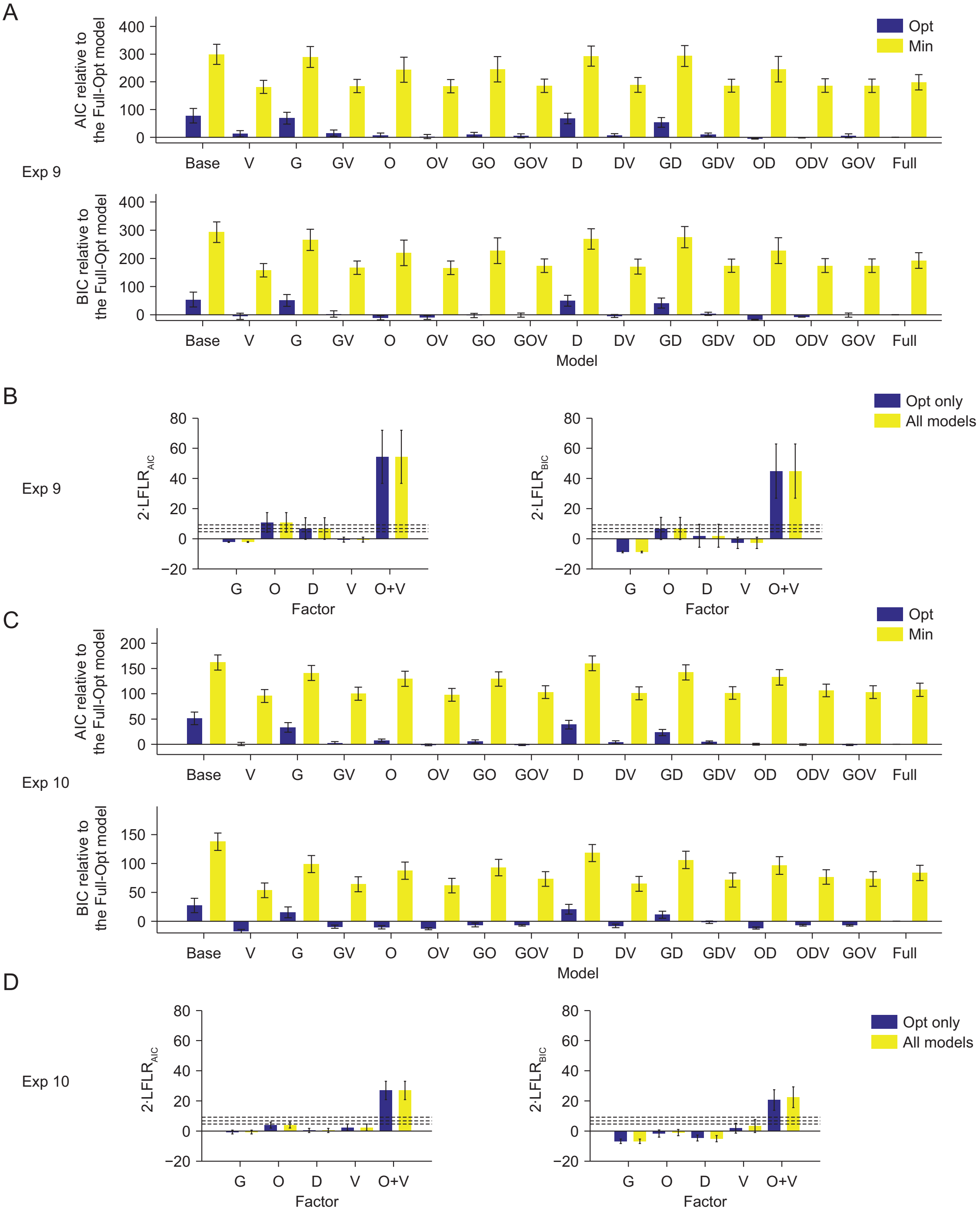


Figure 11. Comparing the optimal with the Min rule in Experiments 9 and 10. (A) Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full-opt model in Experiment 9. Blue (dark grey) bar: models with the optimal decision rule. Yellow (light grey) bar: models with the Min decision rule. (B) Mean and *SEM* of 2 · LFLR based on AIC (left) or BIC (right) for the factors G, O, D, V, and the OV combination in Experiment 9. Blue (dark grey) bars: only models with the optimal decision rule are included. Yellow (light grey) bars: all models are included; we marginalized over decision rule (Opt/Min) in the same way as we marginalized over the "missing" GODV factors. (C–D) Same as (A–B), but for Experiment 10. See the online article for the color version of this figure.

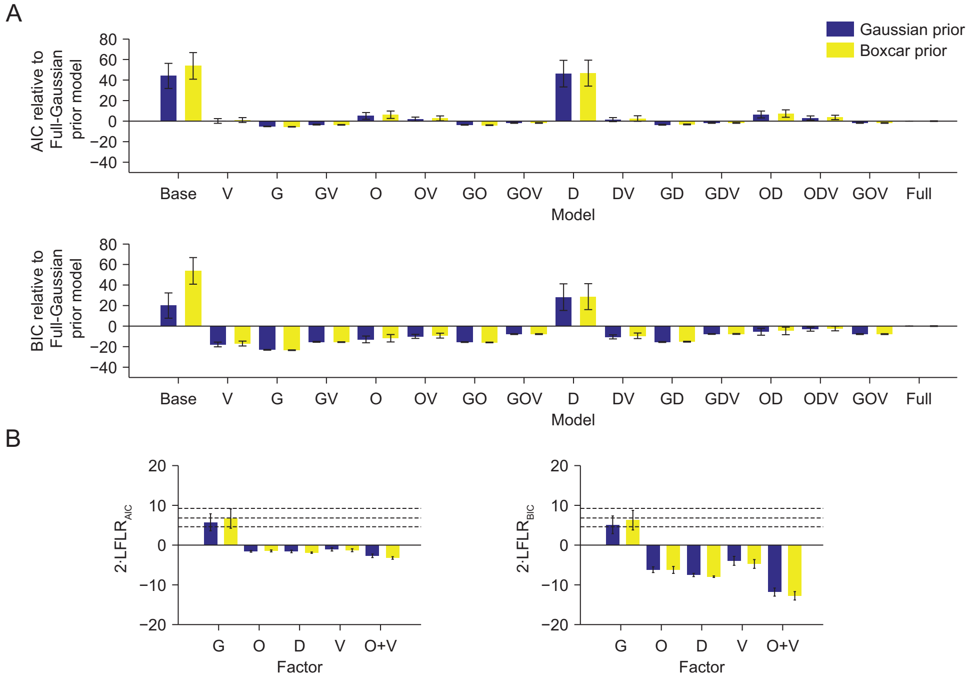


Figure 12. Comparing models with a Gaussian prior and a boxcar prior over orientation in Experiment 5. (A) Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full-Gaussian prior model in Experiment 5. Blue (dark grey) bars: models with a Gaussian prior. Yellow (light grey) bar: models with a boxcar prior. (B) Mean and *SEM* of 2 · LFLR based on AIC (left) or BIC (right) for the factors G, O, D, V, and the OV combination in Experiment 5. Blue (dark grey) bar: models with a Gaussian prior. Yellow (light grey) bars: models with a boxcar prior. See the online article for the color version of this figure.

Experiments 7–11, if we had only considered factors G and V, we would have claimed evidence for the presence of factor V. However, when we considered factors O and D as well, the evidence weakened in Experiment 8 and disappeared in Experiments 9–11 (consistent with findings by Pratte et al. (2017) that were obtained without factorial model comparison). Evidence for the presence of factor V remained strong in Experiment 7, but then disappeared when considering suboptimal decision rules. Thus, we are not convinced that precision is ever variable in visual perception. On the positive side, this means that modelers of visual perception might not be making a major mistake when they do not include variable precision in their models.

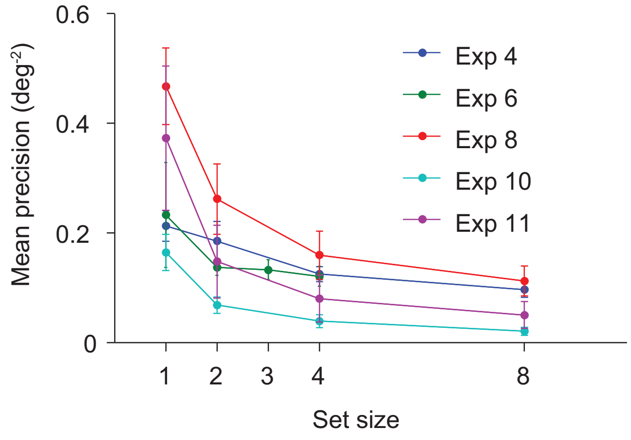


Figure 13. Relationship between mean precision and set size, estimated with the full model, in all experiments with multiple set sizes (mean ±1 *SEM*). The effect of set size is significant in all experiments except Experiment 6. See the online article for the color version of this figure.

Why Did We Not Get Stronger Results?

Group evidence. We quantified all evidence for models or factors by taking the average of AIC, BIC, or derived metrics over subjects. Instead, we could have summed (K. E. Stephan, Marshall, Penny, Friston, & Fink, 2007). We deliberately did not do so, because the underlying assumption would have been that all subjects follow the same model, which is of which we are not convinced. A solution could have been to do group Bayesian model selection (Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009), which marginalizes over assignments of subjects to models. However, we did not trust this method given the low numbers of subjects and large numbers of models in our experiments. Therefore, we decided to describe evidence in an even more conservative way, namely by averaging over subjects. This approach has led us to conclusions that are certainly more cautious than if we had summed evidence, and probably more cautious than if we had used group Bayesian model selection.

Experimental design. We did not optimize our stimulus design for our main question of whether residual variable precision is present. For example, we could have calculated the target or distractor distribution width that would have had the highest expected information gain about the presence of factor V, by simulating large numbers of synthetic data sets. However, it is not clear whether this would have helped a lot, and moreover, it would have depended on the assumptions of subject parameters. In addition, asking subjects to report their confidence could increase model identifiability (van den Berg et al., 2017).

Relation to Previous Work

Relation to work on visual STM (VSTM). Recent studies that claimed to find evidence for variable precision in VSTM (D. T. Devkar et al., 2015; Fougnie et al., 2012; Keshvari et al., 2012, 2013; van den Berg et al., 2012) did not take into account all confounding factors that we considered here: guessing, heteroskedasticity and decision noise. This raises the question of how much residual variability is present in VSTM when accounting for all confounding factors. Given that we found little evidence for residual variable precision in our perceptual tasks, we are left with two possibilities. First, if we still consider residual variability in precision as an established phenomenon in visual working memory, then our study has made several proposed explanations that are not memory-related, such as variability in spike counts for a given gain (Bays, 2014), fluctuations in attention (Cohen & Kohn, 2011; Cohen & Maunsell, 2009), shifts in attention (Lara & Wallis, 2012), less likely. However, it is also possible (and we believe more likely) that the evidence for residual variable precision in visual working memory is not nearly as strong as we originally believed. In visual working memory tasks that previously claimed evidence for factor V, confounding factors are usually not considered (Pratte et al., 2017, being a notable exception). Thus, we consider our work as reason to reconsider the evidence for factor V in STM in the future.

In color STM, estimation precision was found to be much higher for some color configurations than for others, beyond what would be expected from heteroskedasticity (Brady & Alvarez, 2015). This raises the possibility that stimulus context is critical for residual variable precision. In the present work, however, we did not find evidence for the presence of factor V even when the stimulus configuration was very different across trials (in Experiments 7–11). One possible explanation for this discrepancy is that delayed estimation might be more sensitive to the presence of residual variable precision than our binary categorization task. Unfortunately, delayed estimation is not easily adapted to a purely perceptual setting. Another possibility is that residual variability in precision is greater in working memory than in perception.

Relation to work on discriminating noise in different stages. Previous work has characterized noise in human behavior with various approaches. In contrast to detection studies, varying external noise allows one to estimate internal noise (Burgess et al., 1981; Liu et al., 1995; Pelli & Farell, 1999). In Burgess, Wagner, Jennings, and Barlow (1981; Pelli & Farell, 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. "Sampling efficiency" characterizes how close to optimal the decoder is, for example, how well matched Gabor filters are to the stimulus.[1]Like other forms of suboptimality, low sampling efficiency could cause more variability in behavior; in our models, it would be absorbed into decision noise (Beck et al., 2012). More recently, Cabrera et al. (2015) developed an extension of signal detection theory framework to separately estimate encoding noise and decision noise. They found that in a visual detection confidence rating experiment, the decision noise is negligible when there are 2 or 3 response alternatives, which is consistent with our low evidence for decision noise.

Using Bayesian modeling, Drugowitsch et al. (2016) distinguished sources of suboptimality in an evidence accumulation task. The factors they tested included encoding noise, inference noise, selection noise, and deterministic biases. Their encoding noise was equivalent to ours (without O or V). Their inference noise and selection noise were both forms of decision noise, with the former being added at each time step and the latter only once at the end; in our work, these are indistinguishable. They compared models that each had one form of noise with the Base model without noise, similar to our knock-in analysis. They found that a model with inference noise explains the data best. However, they did not do full factorial model comparison and did not compute evidence of factor presence; therefore, their results cannot immediately be compared with ours.

Factor Importance in Factorial Model Comparison

Apart from our scientific question, some of the model comparison methods we used might be useful in other contexts. Although, factorial model comparison (Acerbi, Vijayakumar et al., 2014; van den Berg et al., 2014) helps avoid biases and oversights when deciding which models to compare, its drawbacks include model proliferation and model nonidentifiability. Model proliferation is the phenomenon that the number of models rises exponentially in the number of factors (van den Berg et al., 2014). For example, in Experiment 7, we tested a total of 308 models. This large number of models makes it challenging to sensibly summarize the conclusions of the model comparison. Moreover, many models will be difficult to distinguish, or in other words, they will be *nonidentifiable* (Lehmann & Casella, 1998, Definition 1.5.2; Acerbi, Ma et al., 2014; Shen & Ma, 2016; van den Berg et al., 2014).

Both problems might be alleviated by focusing on the evidence that a factor is important, rather than on the evidence for a specific model. Van den Berg et al. (2014) summarized the results of their factorial model comparison into evidence curves for factors. Here, we introduced three new metrics of factor importance: KID, KOD, and LFLR. All three can be directly computed from the evidence for individual models. LFLR is the most principled and reflects the evidence that a factor is present. However, we made several assumptions in calculating it: that the models tested are "representative", that all models have equal prior probabilities conditioned on factor presence or absence, and that log marginal likelihood can be estimated from AIC or BIC. All these assumptions should be questioned, and the toolkit for quantifying evidence for factor importance will need to be further refined.

[1] Confusingly, Liu et al. (1995) also measure "efficiency" by varying the external noise, but in their terminology, any form of inefficiency is purely a consequence of the internal noise.

References

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, & S. Vishwanathan (Eds.), *Advances in neural information processing systems: Vol. 30. G. R* (pp. 1834–1844). Red Hook, NY: Curran Associates, Inc. http://dx.doi.org/10.1101/150052

Acerbi, L., Ma, W. J., & Vijayakumar, S. (2014). A framework for testing identifiability of Bayesian models of perception. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27, pp. 1026–1034). Red Hook, NY: Curran Associates, Inc.

Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology, 10,* e1003661. http://dx.doi.org/10.1371/journal.pcbi.1003661

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the "light-from-above" prior. *Nature Neuroscience, 7,* 1057–1058. http://dx.doi.org/10.1038/nn1312

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

Akaike, H. (1978). On the likelihood of a time series model. *The Statistician, 27,* 217–235. http://dx.doi.org/10.2307/2988185

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika, 66,* 237–242. http://dx.doi.org/10.1093/biomet/66.2.237

Andrews, D. P. (1965). Perception of contours in the central fovea. *Nature, 205,* 1218–1220. http://dx.doi.org/10.1038/2051218a0

Andrews, D. P. (1967). Perception of contour orientation in the central fovea. II. Spatial integration. *Vision Research, 7,* 999–1013. http://dx.doi.org/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,* 266–278. http://dx.doi.org/10.1037/h0033117

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

Bae, G., 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,* 1–23. http://dx.doi.org/10.1167/14.4.7.doi

Baldassi, S., & Burr, D. C. (2000). Feature-based integration of orientation signals in visual search. *Vision Research, 40,* 1293–1300. http://dx.doi.org/10.1016/S0042-6989(00)00029-8

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT press. http://dx.doi.org/10.1080/15459620490885644

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *The Journal of Neuroscience, 34,* 3632–3645. http://dx.doi.org/10.1523/JNEUROSCI.3204-13.2014

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, 74,* 30–39. http://dx.doi.org/10.1016/j.neuron.2012.03.016

Bhardwaj, M., van den Berg, R., Ma, W. J., & Josić, K. (2016). Do people take stimulus correlations into account in visual search? *PLoS ONE, 11,* e0149402. http://dx.doi.org/10.1371/journal.pone.0149402

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision, 15,* 1–24. http://dx.doi.org/10.1167/15.15.6

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/b97636

Cabrera, C. A., Lu, Z.-L., & Dosher, B. A. (2015). Separating decision and encoding noise in signal detection tasks. *Psychological Review, 122,* 429–460. http://dx.doi.org/10.1037/a0039348

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,* 818–831. http://dx.doi.org/10.1016/j.neuron.2010.12.037

Cohen, M. R., & Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Neuroscience, 14,* 811–819. http://dx.doi.org/10.1038/nn.2842

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

Cover, T., & Thomas, J. (2006). *Elements of information theory* (2nd ed.). Somerset, England: Wiley.

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

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

Devkar, D. T., Wright, A. A., & Ma, W. J. (2015). The same type of visual working memory limitations in humans and monkeys. *Journal of Vision, 15,* 1–18. http://dx.doi.org/10.1167/15.16.13

Devkar, D., Wright, A. A., & Ma, W. J. (2017). Monkeys and humans take local uncertainty into account when localizing a change. *Journal of Vision, 17,* 1–15. http://dx.doi.org/10.1167/17.11.4

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. *Psychological Review, 120,* 873–902. http://dx.doi.org/10.1037/a0034247

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron, 92,* 1398–1411. http://dx.doi.org/10.1016/j.neuron.2016.11.005

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science, 9,* 111–118. http://dx.doi.org/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,* 292–303. http://dx.doi.org/10.1038/nrn2258

Fechner, G. T. (1860). *Elemente Dur Psychophysik* [Elements of psychophysics]. Leipzig, Germany: Breitkopf und Härtel. http://dx.doi.org/10.1111/j.2044-8317.1960.tb00033.x

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

Ganguli, D., & Simoncelli, E. P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation, 26,* 2103–2134. http://dx.doi.org/10.1162/NECO_a_00638

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research, 51,* 771–781. http://dx.doi.org/10.1016/j.visres.2010.09.027

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

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669. http://dx.doi.org/10.1037/0033-295X.103.4.650

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

Goris, R. L. T., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience, 17,* 858–865. http://dx.doi.org/10.1038/nn.3711

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Harvey, L. O. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments, & Computers, 18,* 623–632. http://dx.doi.org/10.3758/BF03201438

Honig, M., Ma, W. J., & Fougnie, D. (2018). *Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions*. http://dx.doi.org/10.1101/306225

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York NY: Oxford University Press.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572

Kelly, J. G., & Matthews, N. (2011). Attentional oblique effect when judging simultaneity. *Journal of Vision, 11,* 1–15. http://dx.doi.org/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,* e40216. http://dx.doi.org/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,* e1002927. http://dx.doi.org/10.1371/journal.pcbi.1002927

Lara, A. H., & Wallis, J. D. (2012). Capacity and precision in an animal model of visual short-term memory. *Journal of Vision, 12,* 1–12. http://dx.doi.org/10.1167/12.3.13

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York, NY: Springer.

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology, 90,* 204–217. http://dx.doi.org/10.1152/jn.00954.2002

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research, 35,* 549–568. http://dx.doi.org/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,* 123–127. http://dx.doi.org/10.1038/nature09086

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

Ma, W. J., Shen, S., Dziugaite, G., & van den Berg, R. (2015). Requiem for the max rule? *Vision Research, 116,* 179–193. http://dx.doi.org/10.1016/j.visres.2014.12.019

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

Mazyar, H., van den Berg, R., & Ma, W. J. (2012). Does precision decrease with set size? *Journal of Vision, 12,* 1–16. http://dx.doi.org/10.1167/12.6.10

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

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

Nolte, L. W., & Jaarsma, D. (1967). More on the detection of one of *M* orthogonal signals. *The Journal of the Acoustical Society of America, 41,* 497–505. http://dx.doi.org/10.1121/1.1910360

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review, 124,* 21–59. http://dx.doi.org/10.1037/rev0000044

Orhan, A. E., Sims, C. R., Jacobs, R. A., & Knill, D. C. (2014). The adaptive nature of visual working memory. *Current Directions in Psychological Science, 23,* 164–170. http://dx.doi.org/10.1177/0963721414529144

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

Pelli, D. G., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 16,* 647–653. http://dx.doi.org/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. *Journal of Experimental Psychology: Human Perception and Performance, 43,* 6–17. http://dx.doi.org/10.1037/xhp0000302

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *NeuroImage, 84,* 971–985. http://dx.doi.org/10.1016/j.neuroimage.2013.08.065

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. http://dx.doi.org/10.1016/j.visres.2016.02.002

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. http://dx.doi.org/10.1214/aos/1176344136

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

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

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63,* 129–138. http://dx.doi.org/10.1037/h0042769

Soltani, A., & Wang, X. J. (2006). A biophysically based neural model of matching law behavior: Melioration by stochastic synapses. *The Journal of Neuroscience, 26,* 3731–3744. http://dx.doi.org/10.1523/JNEUROSCI.5159-05.2006

Stephan, K. E., Marshall, J. C., Penny, W. D., Friston, K. J., & Fink, G. R. (2007). Interhemispheric integration of visual processing during task-driven lateralization. *The Journal of Neuroscience, 27,* 3512–3522. http://dx.doi.org/10.1523/JNEUROSCI.4766-06.2007

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage, 46,* 1004–1017. http://dx.doi.org/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,* 591.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286. http://dx.doi.org/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, 23,* 775–785. http://dx.doi.org/10.1016/0042-6989(83)90200-6

Townsend, J. T. (1972). Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathematical & Statistical Psychology, 25,* 168–197. http://dx.doi.org/10.1111/j.2044-8317.1972.tb00490.x

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121,* 124–149. http://dx.doi.org/10.1037/a0035234 van den Berg, R., & Ma, W. J. (2017). A rational theory of the limitations of working memory and attention. 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 of the United States of America, 109,* 8780–8785. http://dx.doi.org/10.1073/pnas.1117465109

van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review, 124,* 197–214. http://dx.doi.org/10.1037/rev0000060

Van Horn, K. S. (2003). Constructing a logic of plausible inference: A guide to Cox's theorem. *International Journal of Approximate Reasoning, 34,* 3–24. http://dx.doi.org/10.1016/S0888-613X(03)00051-3

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196. http://dx.doi.org/10.3758/BF03206482

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

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33,* 113–120. http://dx.doi.org/10.3758/BF03202828

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience, 18,* 1509–1517. http://dx.doi.org/10.1038/nn.4105

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

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4,* 1120–1135. http://dx.doi.org/10.1167/4.12.11

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453,* 233–235. http://dx.doi.org/10.1038/nature06860

Zhu, J. E., & Ma, W. J. (2017). Orientation-dependent biases in length judgments of isolated stimuli. *Journal of Vision, 17,* 20. http://dx.doi.org/10.1167/17.2.20

Appendix A

Decision Rules

Following the theory part, all decision rules can be derived from Equations (9) and (10). In Equation (10), $p(\mathbf{x}|\mathbf{s})$ is evaluated in either Equation (1) or (2). We now specify the evaluation of $p(\mathbf{s}|C)$ for each experiment.

Notation:

s_T: target orientation

s_D or $\mathbf{s}_\mathrm{D}$: distractor orientation or distractor orientation vector

x_T: internal measurement of the target

x_ref: internal measurement of the reference

x_i: internal measurement of the ith stimulus

σ_i: measurement noise of the ith stimulus, standard deviation of the Gaussian distribution

σ_T: measurement noise of the target stimulus, standard deviation of the Gaussian distribution

σ_ref: measurement noise of the reference stimulus, standard deviation of the Gaussian distribution

σ_s: standard deviation to generate the stimulus orientations

J_i: encoding precision of the ith stimulus

J_s: precision to generate stimulus orientations

κ_i: concentration parameter of Von Mises distribution of the ith stimulus

κ_s: concentration parameter of the Von Mises distribution to generate stimulus orientations

κ_T: concentration parameter of the Von Mises distribution to generate target orientations

L: hypothesized target location

N: set size

p_right, $p_\mathrm{clockwise}$, or p_present: prior probability of reporting "right," "clockwise," or "present"

$\Phi(x)$: standard normal cumulative function

$\delta(x)$: Dirac delta function

$N(x;\ \mu,\ \sigma^2)$: Gaussian distribution with a mean of μ and a variance of σ^2

$H(x)$: Heaviside step function

U(a, b): uniform distribution in a range between a and b.

$\mathrm{VM}(x;\ s,\ \kappa)$: Von Mises distribution with a mean of s and concentration parameter of κ

(*Appendices continue*)

Experiment 1: Single Stimulus, Four Possible Locations

All trials in Experiment 1 belong to Case 1, but the stimulus vector is now a scalar s_T. The distribution $p(\mathbf{s}|C)$ in Equation (11) becomes $p(s|C)$ and takes the form of a truncated Gaussian distribution (Equation 6).

Putting Equations (6), (9), and (10) together and simplifying, the final decision variable is

$$d = \log\frac{\Phi\left(\frac{xJ}{\sqrt{J+J_\mathrm{s}}}\right)}{\Phi\left(\frac{-xJ}{\sqrt{J+J_\mathrm{s}}}\right)} + \log\frac{p_\mathrm{right}}{1-p_\mathrm{right}}.$$

Experiment 2: Categorization of a Single Target With Respect to a Variable Reference Orientation

The generative model of Experiment 2 is shown in Appendix Figure A14. The distributions are as follows:

$$\begin{aligned}
p(\Delta s|C) &= 2\cdot N\big(\Delta s;\, 0,\, \sigma_\mathrm{s}^2\big)H(C\cdot\Delta s)\\
p(s_\mathrm{ref}) &= U(-90,\, 90)\\
p(s_\mathrm{T}|s_\mathrm{ref}, \Delta s) &= \delta(s_\mathrm{T} - s_\mathrm{ref} - \Delta s)\\
p(x_\mathrm{T}|s_\mathrm{T} = s_\mathrm{ref} + \Delta s) &= N\big(x_\mathrm{T};\, s_\mathrm{ref} + \Delta s,\, \sigma_\mathrm{T}^2\big)\\
p(x_\mathrm{ref}|s_\mathrm{ref}) &= N\big(x_\mathrm{ref};\, s_\mathrm{ref},\, \sigma_\mathrm{ref}^2\big).
\end{aligned} \tag{25}$$

We use Gaussian distributions to model $p(\Delta s|C)$, $p(x_\mathrm{T}|s_\mathrm{T})$ and $p(x_\mathrm{ref}\ |s_\mathrm{ref})$, because the variable and its mean are relatively close in all three distributions.

All trials in Experiment 2 belong to Case 4. Putting together Equations (9), (10), (15), and (16), we get:

$$\begin{aligned}
p(\mathbf{x}|C) &= \iiint p(x_\mathrm{T}|s_\mathrm{T})p(x_\mathrm{ref}|s_\mathrm{ref})p(s_\mathrm{T}|s_\mathrm{ref}, \Delta s)p(s_\mathrm{ref})\\
&\quad\times p(\Delta s|C)ds_\mathrm{T}ds_\mathrm{ref}d(\Delta s)\\
&= \iiint p(x_\mathrm{T}|s_\mathrm{T})p(x_\mathrm{ref}|s_\mathrm{ref})\delta(s_\mathrm{T} - s_\mathrm{ref} - \Delta s)p(s_\mathrm{ref})\\
&\quad\times p(\Delta s|C)ds_\mathrm{T}ds_\mathrm{ref}d(\Delta s)\\
&= \iint p(x_\mathrm{T}|s_\mathrm{T} = s_\mathrm{ref} + \Delta s)p(x_\mathrm{ref}|s_\mathrm{ref})p(s_\mathrm{ref})\\
&\quad\times p(\Delta s|C)ds_\mathrm{ref}d(\Delta s)\\
&= \int\left(\int p(x_\mathrm{T}|s_\mathrm{T} = s_\mathrm{ref} + \Delta s)p(x_\mathrm{ref}|s_\mathrm{ref})p(s_\mathrm{ref})ds_\mathrm{ref}\right)\\
&\quad\times p(\Delta s|C)d(\Delta s).
\end{aligned}$$

Now we substitute the probability distributions as in Equation (25), and $p(\mathbf{x}|C)$ becomes

$$p(\mathbf{x}|C) = \int_{-90}^{90}\left(\int_{-90}^{90} N\big(x_\mathrm{T};\, s_\mathrm{ref} + \Delta s,\, \sigma_\mathrm{T}^2\big)N\big(x_\mathrm{ref};\, s_\mathrm{ref},\, \sigma_\mathrm{ref}^2\big)\frac{1}{180}ds_\mathrm{ref}\right)2\cdot$$
$$N\big(\Delta s;\, 0, \sigma_\mathrm{s}^2\big)H(C\cdot\Delta s)d(\Delta s).$$

Numerically computing the double integral is not computationally feasible: It happens inside a Monte Carlo simulation of $\mathbf{x}$, which happens inside a parameter optimization, which happens inside a large model comparison. Therefore, we approximate partial integrals with the full integral:

$$p(\mathbf{x}|C) \approx \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} N\big(x_\mathrm{T};\, s_\mathrm{ref} + \Delta s,\, \sigma_\mathrm{T}^2\big)N\big(x_\mathrm{ref};\, s_\mathrm{ref},\, \sigma_\mathrm{ref}^2\big)\frac{1}{180}ds_\mathrm{ref}\right)2\cdot$$
$$N\big(\Delta s;\, 0,\, \sigma_\mathrm{s}^2\big)H(C\cdot\Delta s)d(\Delta s).$$

We simplify the inner integral by making use of the fact that the convolution of two Gaussian probability density functions is also Gaussian and get

$$p(\mathbf{x}|C) \approx \frac{1}{180}\int_{-\infty}^{+\infty} N\big(x_\mathrm{T} - x_\mathrm{ref};\, \Delta s,\, \sigma_\mathrm{T}^2 + \sigma_\mathrm{ref}^2\big)2\cdot$$
$$N\big(\Delta s;\, 0,\, \sigma_\mathrm{s}^2\big)H(C\cdot\Delta s)d(\Delta s).$$

We define $\Delta x = x_\mathrm{T} - x_\mathrm{ref}$, then get

$$p(\mathbf{x}|C) \approx \frac{1}{180}\int_{-\infty}^{+\infty} N\big(\Delta x;\, \Delta s,\, \sigma_\mathrm{T}^2 + \sigma_\mathrm{ref}^2\big)2\cdot$$
$$N\big(\Delta s;\, 0,\, \sigma_\mathrm{s}^2\big)H(C\cdot\Delta s)d(\Delta s).$$

By further simplification, the final decision variable is

$$d = \log\frac{\Phi\left(\frac{\Delta xJ_\mathrm{c}}{\sqrt{2(J_\mathrm{c}+J_\mathrm{s})}}\right)}{\Phi\left(\frac{\Delta xJ_\mathrm{c}}{\sqrt{2(J_\mathrm{c}+J_\mathrm{s})}}\right)} + \log\frac{p_\mathrm{clockwise}}{1-p_\mathrm{clockwise}},$$

where $J_\mathrm{c} = \dfrac{1}{\dfrac{1}{J_\mathrm{T}}+\dfrac{1}{J_\mathrm{ref}}}$.

Experiments 3 and 4: Categorization With All Stimuli Being Targets

All trials of both experiments belong to Case 1. Because all stimuli are targets, the distribution $p(\mathbf{s}|C)$ takes the form

$$\begin{aligned}
p(\mathbf{s}|C) &= \int p(\mathbf{s}|s_\mathrm{T})p(s_\mathrm{T}|C)ds_\mathrm{T}\\
&= \int \delta(\mathbf{s} - s_\mathrm{T}\mathbf{1})p(s_\mathrm{T}|C)ds_\mathrm{T},
\end{aligned} \tag{26}$$

where $\mathbf{1}$ is the vector of 1s and $p(s_\mathrm{T}|C)$ is a truncated Gaussian distribution (Equation 6).

Putting Equations (6), (9), (10), and (26) together and simplifying, the final decision variable is

$$d = \log\frac{\Phi\left(\frac{\sum_{i=1}^{N}x_iJ_i}{\sqrt{\left(\sum_{i=1}^{N}J_i\right)+J_\mathrm{s}}}\right)}{\Phi\left(\frac{-\sum_{i=1}^{N}x_iJ_i}{\sqrt{\left(\sum_{i=1}^{N}J_i\right)+J_\mathrm{s}}}\right)} + \log\frac{p_\mathrm{right}}{1-p_\mathrm{right}}.$$

(*Appendices continue*)

Experiments 5 and 6: Categorization of a Single Target With Vertical Distractors

All trials of both experiments belong to Case 3. Combining Equations (13) and (14), the distribution $p(\mathbf{s}|C)$ takes the form

$$\begin{aligned} p(\mathbf{s}|C) &= \iint p(\mathbf{s}|s_\mathrm{T},\mathbf{s}_\mathrm{D})p(s_\mathrm{T}|C)p(\mathbf{s}_\mathrm{D})ds_\mathrm{T}d\mathbf{s}_\mathrm{D} \\ &= \iint \frac{1}{N}\sum_{L=1}^{N}\delta(s_L - s_\mathrm{T})\delta(\mathbf{s}_{\backslash L} - \mathbf{s}_\mathrm{D})p(s_\mathrm{T}|C)p(\mathbf{s}_\mathrm{D})ds_\mathrm{T}d\mathbf{s}_\mathrm{D}. \end{aligned} \tag{27}$$

Given that all distractors are vertical, $p(\mathbf{s}_\mathrm{D})$ is equal to $\prod_{i\neq L}\delta(s_{\mathrm{D}i})$. Therefore, Equation (27) is further simplified as

$$p(\mathbf{s}|C) = \frac{1}{N}\sum_{L=1}^{N}\int\delta(s_L - s_\mathrm{T})\delta(\mathbf{s}_{\backslash L})p(s_\mathrm{T}|C)ds_\mathrm{T}, \tag{28}$$

where $p(s_\mathrm{T}|C)$ is a truncated Gaussian distribution (Equation 6).

Putting Equations (6), (9), (10), and (28) together and simplifying, the final decision variable becomes

$$d = \log\frac{\sum_{L=1}^{N}\Phi\left(\frac{x_L J_L}{\sqrt{J_L + J_\mathrm{s}}}\right)\exp\left(\frac{x_L^2 J_L^2}{2(J_L + J_\mathrm{s})}\right)\sqrt{\frac{J_\mathrm{s}}{J_L + J_\mathrm{s}}}}{\sum_{L=1}^{N}\Phi\left(\frac{-x_L J_L}{\sqrt{J_L + J_\mathrm{s}}}\right)\exp\left(\frac{x_L^2 J_L^2}{2(J_L + J_\mathrm{s})}\right)\sqrt{\frac{J_\mathrm{s}}{J_L + J_\mathrm{s}}}} + \log\frac{p_\mathrm{right}}{1 - p_\mathrm{right}}.$$

Experiment 7: Categorization of a Single Target With Homogeneous Distractors

All trials in Experiment 7 belong to Case 3, so $p(\mathbf{s}|C)$ takes the same form as in Equation (27). Given that all distractors are identical and take the value s_D, $p(\mathbf{s}_\mathrm{D})$ equals to $\prod_{i\neq L}\delta(s_{\mathrm{D}i} - s_\mathrm{D})$. Therefore, Equation (27) is further simplified as

$$p(\mathbf{s}|C) = \frac{1}{N}\sum_{L=1}^{N}\iint\delta(s_L - s_\mathrm{T})\delta(\mathbf{s}_{\backslash L} - s_\mathrm{D})p(s_\mathrm{T}|C)p(s_\mathrm{D})ds_\mathrm{T}ds_\mathrm{D}, \tag{29}$$

where $p(s_\mathrm{T}|C)$ is a truncated Gaussian distribution (Equation 6) and $p(s_\mathrm{D})$ is a Gaussian distribution with a mean of 0 and a standard deviation of σ_s.

Putting Equations (6), (9), (10), and (29) together and simplifying, the final decision variable becomes

$$d = \log\frac{\sum_{L=1}^{N}\sqrt{\frac{1}{J_L + J_\mathrm{s}}}\sqrt{\frac{1}{\left(\sum_{i\neq L}J_i\right) + J_\mathrm{s}}}\Phi\left(\frac{x_L J_L}{\sqrt{J_L + J_\mathrm{s}}}\right)\exp\frac{-x_L^2}{2\left(\frac{1}{J_L} + \frac{1}{J_\mathrm{s}}\right)}\exp\left(\frac{1}{2}\frac{\left(\sum_{i\neq L}x_L J_L\right)^2}{\left(\left(\sum_{i\neq L}J_i\right) + J_\mathrm{s}\right)} - \frac{1}{2}\sum_{i\neq L}x_L^2 J_L\right)}{\sum_{L=1}^{N}\sqrt{\frac{1}{J_L + J_\mathrm{s}}}\sqrt{\frac{1}{\left(\sum_{i\neq L}J_i\right) + J_\mathrm{s}}}\Phi\left(\frac{-x_L J_L}{\sqrt{J_L + J_\mathrm{s}}}\right)\exp\frac{-x_L^2}{2\left(\frac{1}{J_L} + \frac{1}{J_\mathrm{s}}\right)}\exp\left(\frac{1}{2}\frac{\left(\sum_{i\neq L}x_L J_L\right)^2}{\left(\left(\sum_{i\neq L}J_i\right) + J_\mathrm{s}\right)} - \frac{1}{2}\sum_{i\neq L}x_L^2 J_L\right)} + \log\frac{p_\mathrm{right}}{1 - p_\mathrm{right}}.$$

Experiment 8: Detection of a Single Target With Homogeneous Distractors

Target present trials belong to Case 3 and all distractors are identical and take the value s_D, so $p(\mathbf{s}|C = 1)$ takes the same form as in Equation (29).

Target absent trials belong to Case 2 and all stimuli share the same orientation s_D, therefore $p(\mathbf{s}|C = -1)$ takes the form:

$$\begin{aligned} p(\mathbf{s}|C = -1) &= \int p(\mathbf{s}|s_\mathrm{D})p(s_\mathrm{D})ds_\mathrm{D} \\ &= \int \delta(\mathbf{s} - s_\mathrm{D}\mathbf{1})p(s_\mathrm{D})ds_\mathrm{D}, \end{aligned} \tag{30}$$

where $p(s_\mathrm{D})$ is a Gaussian distribution with a mean of 0 and a standard deviation of σ_s.

Putting together Equations (6), (9), (10), (29), and (30) together and after simplifying, the final decision variable becomes

$$d = \log\frac{\frac{1}{N}\sum_{L=1}^{N}\frac{1}{\sqrt{\left(\sum_{i\neq L}J_i\right) + J_\mathrm{s}}}\exp\left(\frac{1}{2}\left(\sum_{i\neq L}x_i J_i\right)^2\frac{1}{\left(\sum_{i\neq L}J_i\right) + J_\mathrm{s}}\right)}{\frac{1}{\sqrt{\left(\sum_{i=1}^{N}J_i\right) + J_\mathrm{s}}}\exp\left(\frac{1}{2}\left(\sum_{i=1}^{N}J_i\right)^2\frac{1}{\left(\sum_{i=1}^{N}J_i\right) + J_\mathrm{s}}\right)} + \log\frac{p_\mathrm{present}}{1 - p_\mathrm{present}}.$$

(*Appendices continue*)

Experiments 9 and 10: Categorization of a Single Target With Heterogeneous Distractors

All trials in Experiments 9 and 10 belong to Case 3, so $p(\mathbf{s}|C)$ takes the same form as in Equation (27), where $p(s_\mathrm{T}|C)$ is a truncated Von Mises distribution:

$$p(s_\mathrm{T}|C) = 2 \cdot \mathrm{VM}(2s_\mathrm{T};\, 0,\, \kappa_\mathrm{T})H(C \cdot s_\mathrm{T}) \tag{31}$$

and $p(\mathbf{s}_\mathrm{D})$ is a product of N-1 uniform distributions.

Putting Equations (9), (10), (27), and (31) together and after simplifying, the final decision variable becomes

$$d = \log\frac{\sum_{L=1}^{N}\int_{0}^{\frac{\pi}{2}} \mathrm{VM}(2x_L;\, 2s_\mathrm{T}, \kappa_L)\mathrm{VM}(2s_\mathrm{T};\, 0,\, \kappa_\mathrm{T})ds_\mathrm{T}}{\sum_{L=1}^{N}\int_{-\frac{\pi}{2}}^{0} \mathrm{VM}(2x_L;\, 2s_\mathrm{T},\, \kappa_L)\mathrm{VM}(2s_\mathrm{T};\, 0,\, \kappa_\mathrm{T})ds_\mathrm{T}} + \log\frac{p_\mathrm{right}}{1 - p_\mathrm{right}}.$$

Experiment 11: Detection of a Single Target With Heterogeneous Distractors

Target present trials in Experiment 11 belong to Case 3, and $p(\mathbf{s}|C = 1)$ takes the same form as in Equation (27), where $p(s_\mathrm{T}|C = 1)$ is a truncated Von Mises distribution (Equation 31) and $p(\mathbf{s}_\mathrm{D})$ is a product of N-1 uniform distribution. Target absent trials in Experiment 11 belong to Case 2, and $p(\mathbf{s}|C = -1)$ takes the form as in Equation (14).

Putting Equations (9), (10), (14), (27), and (31) together and after simplifying, the final decision variable becomes

$$d = \log\left(\frac{1}{N}\sum_{L=1}^{N} 2 \cdot \mathrm{VM}(2x_L;\, 0,\, \kappa_L)\right) + \log\frac{p_\mathrm{present}}{1 - p_\mathrm{present}}.$$

(*Appendices continue*)

Appendix B

Additional Model Fits and Model Comparisons

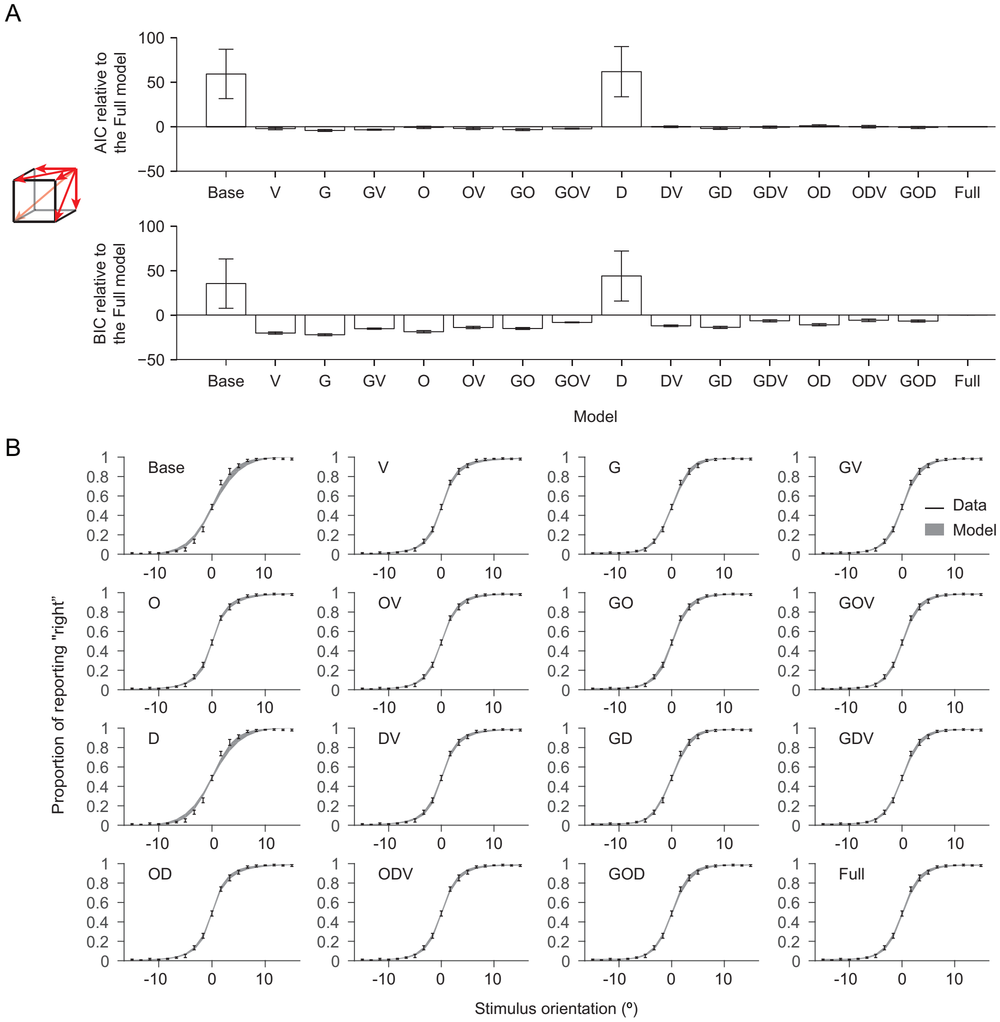


Figure B1. Experiment 1. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the Full model. (B) Proportion of reporting "right" as a function of stimulus orientation: data and model fits. See the online article for the color version of this figure.

(*Appendices continue*)

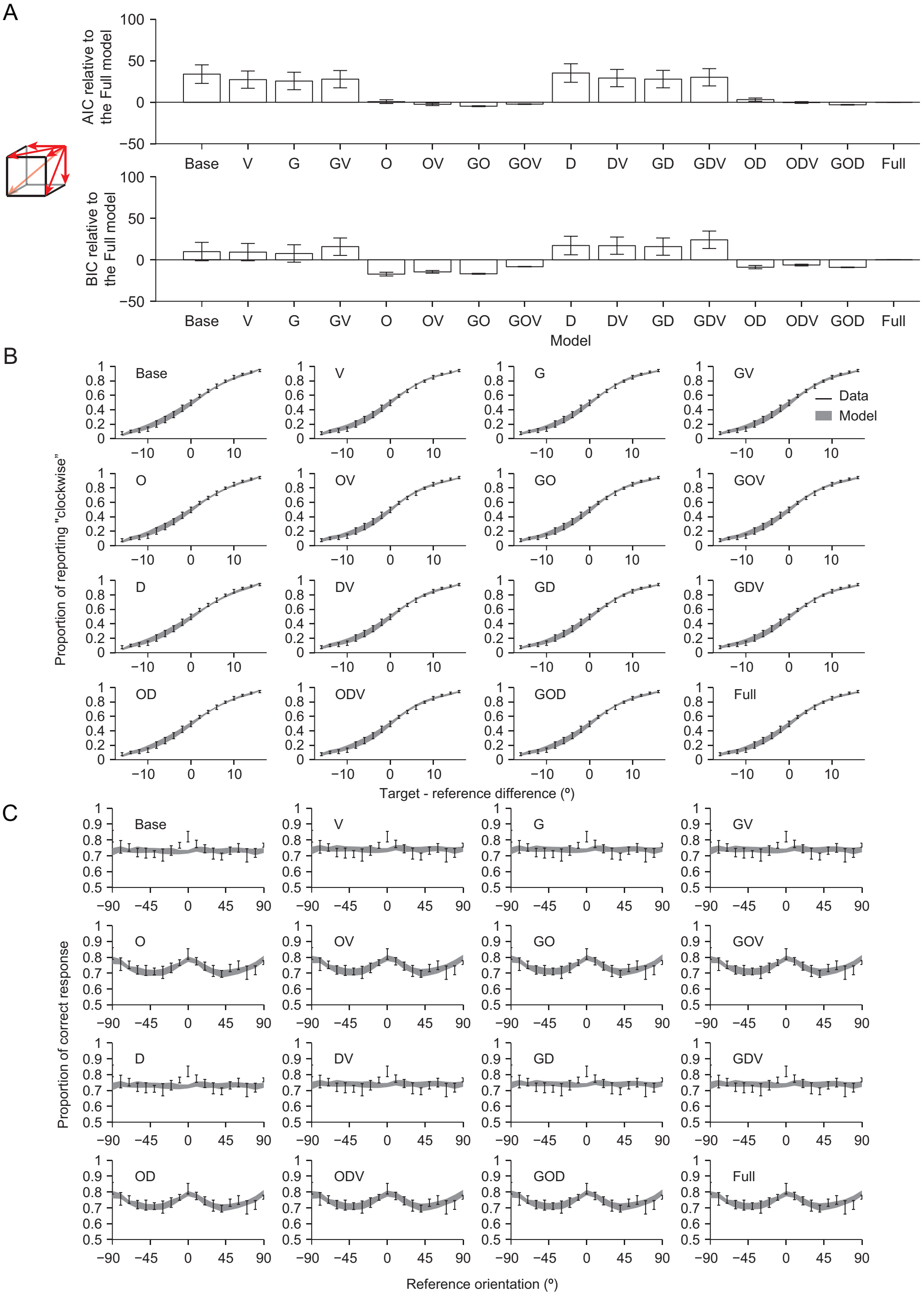


Figure B2. Experiment 2. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting "clockwise" as a function of orientation difference between the target and the reference. (C) Proportion correct as a function of the reference orientation: data and model fits. See the online article for the color version of this figure.

(*Appendices continue*)

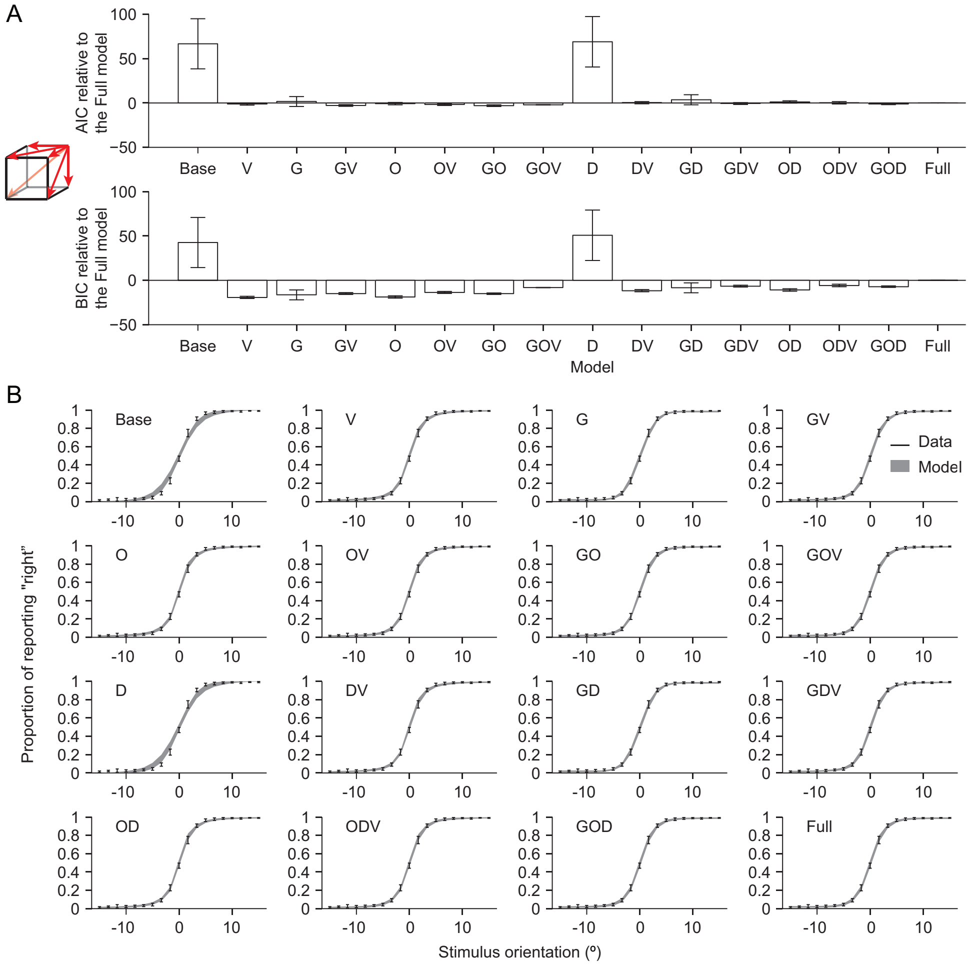


Figure B3. Experiment 3. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting "right" as a function of target orientation: data and model fits. See the online article for the color version of this figure.

(*Appendices continue*)

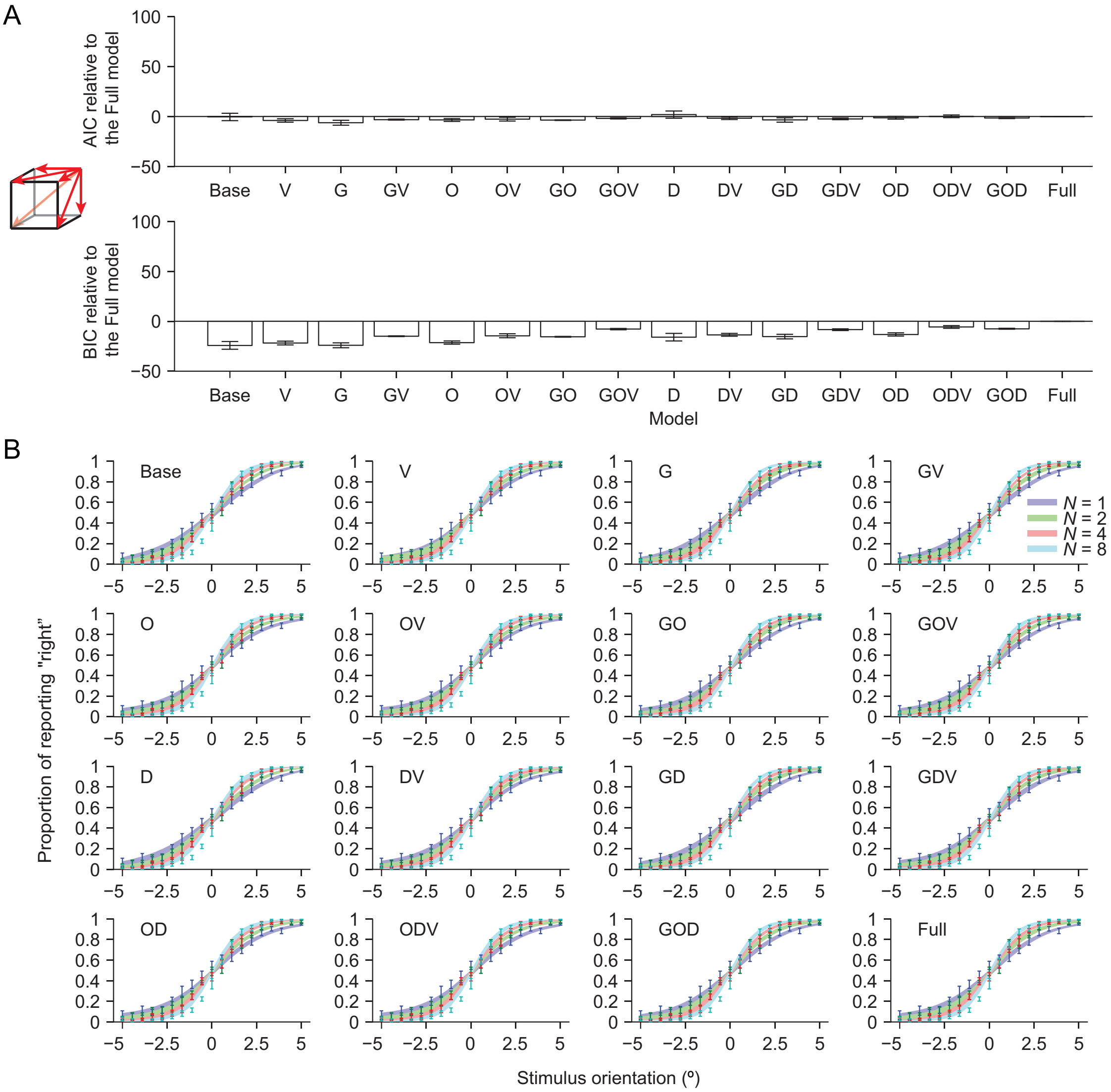


Figure B4. Experiment 4. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting “right” as a function of set size and target orientation: data and model fits. See the online article for the color version of this figure.

(*Appendices continue*)

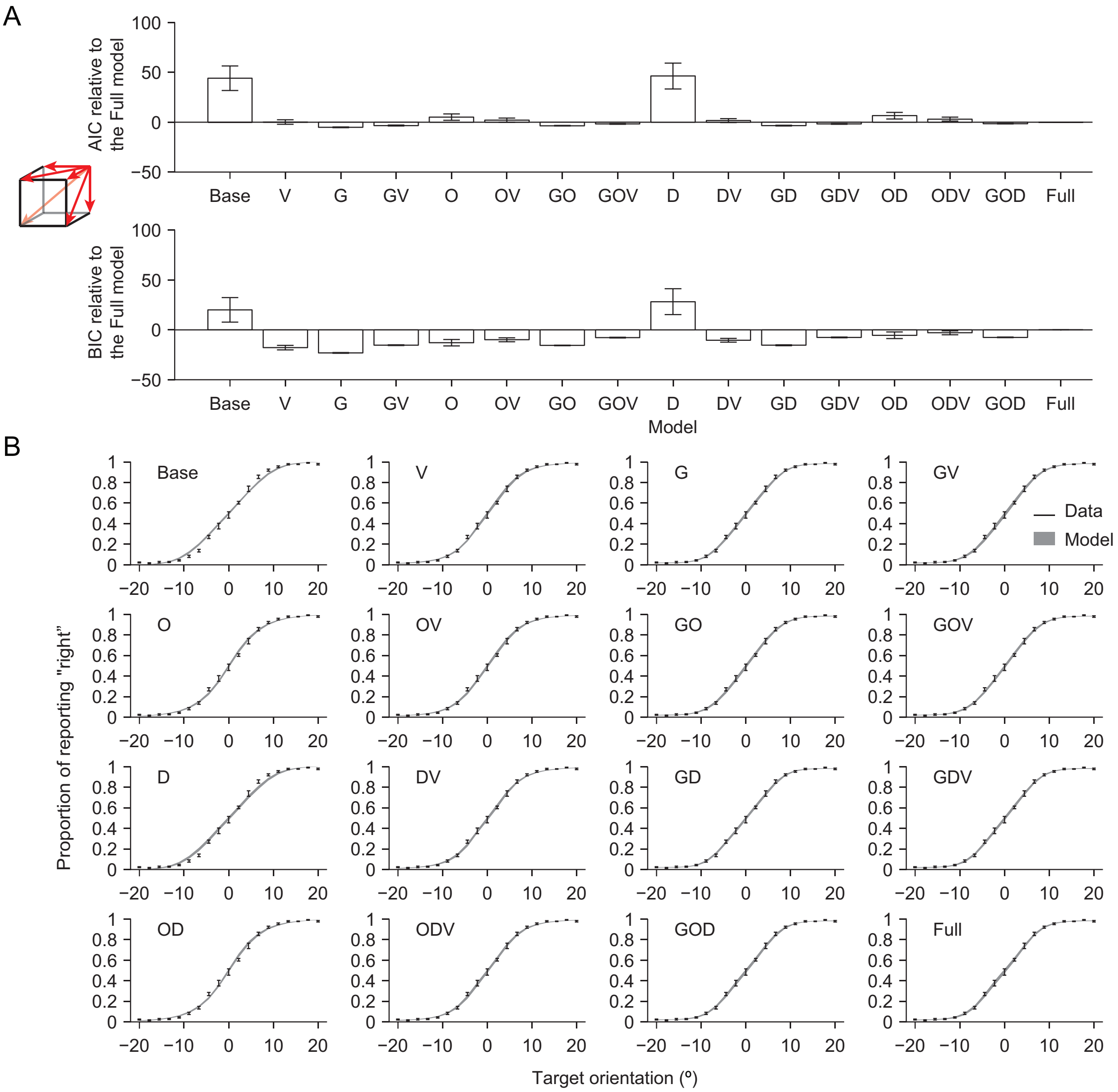


Figure B5. Experiment 5. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting "right" as a function of target orientation. See the online article for the color version of this figure.

(*Appendices continue*)

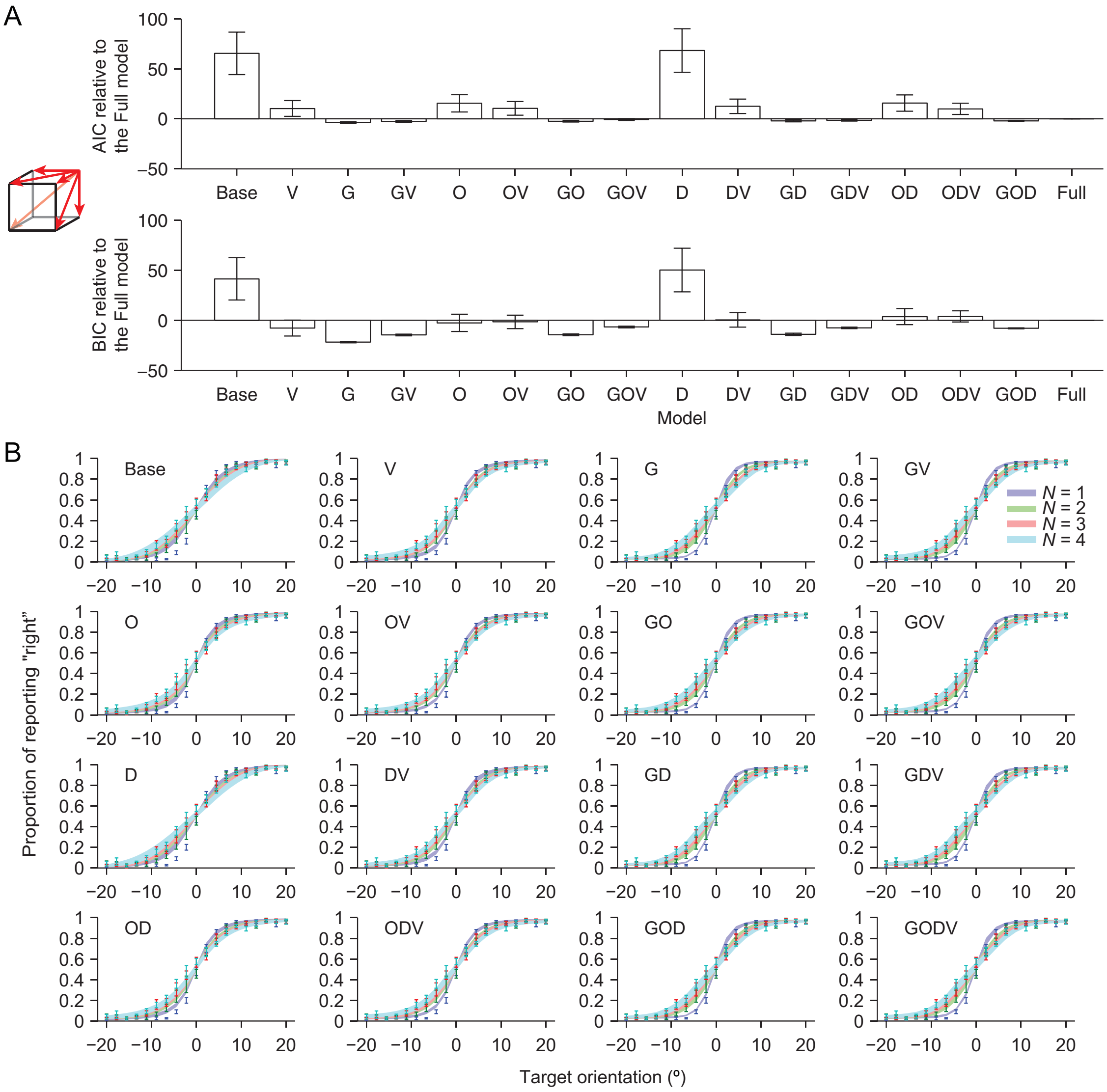


Figure B6. Experiment 6. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting "right" as a function of set size and target orientation: data and model fits. See the online article for the color version of this figure.

(*Appendices continue*)

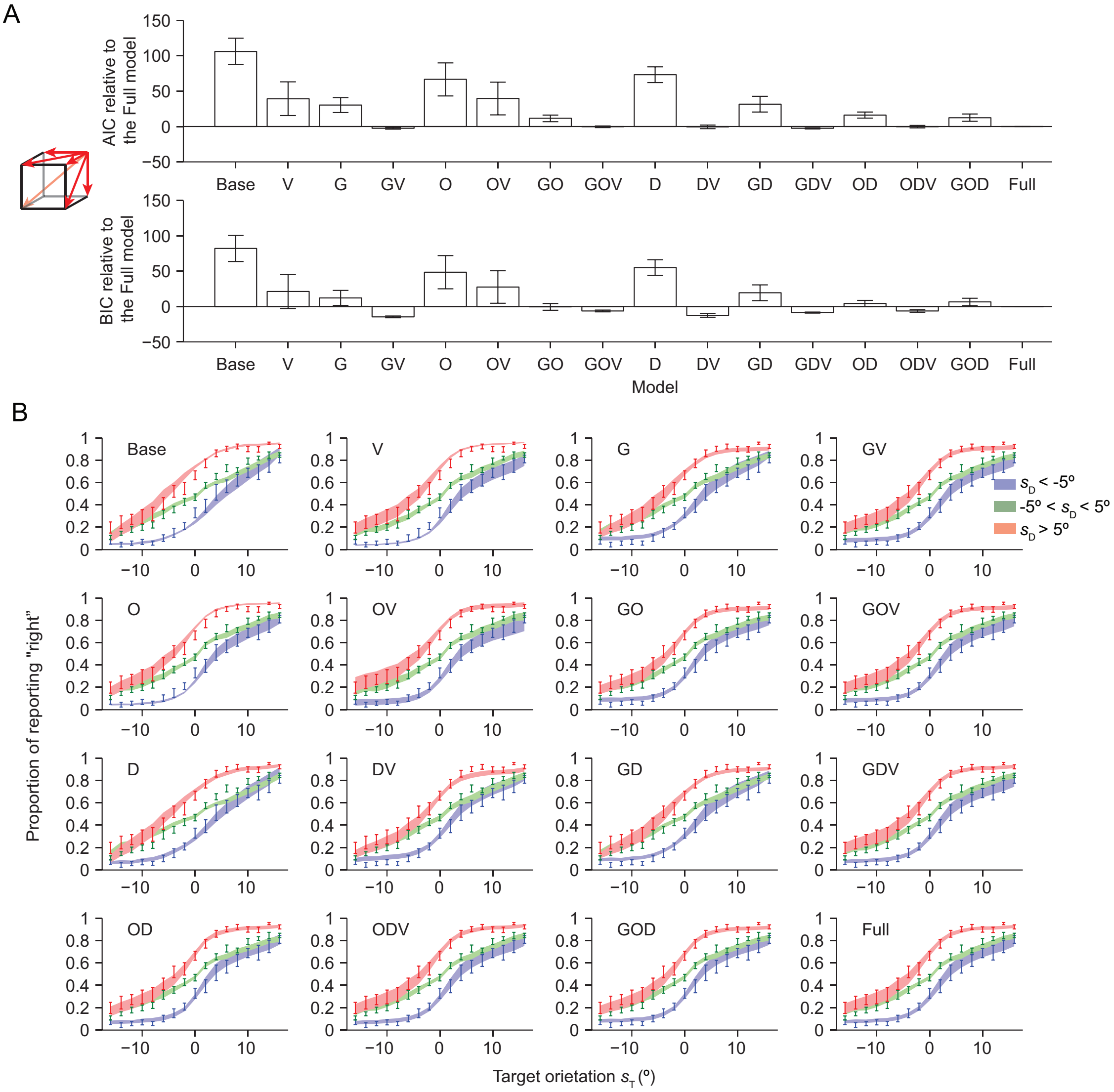


Figure B7. Experiment 7. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting “right” as a function of target orientation: data and model fits. See the online article for the color version of this figure.

(*Appendices continue*)

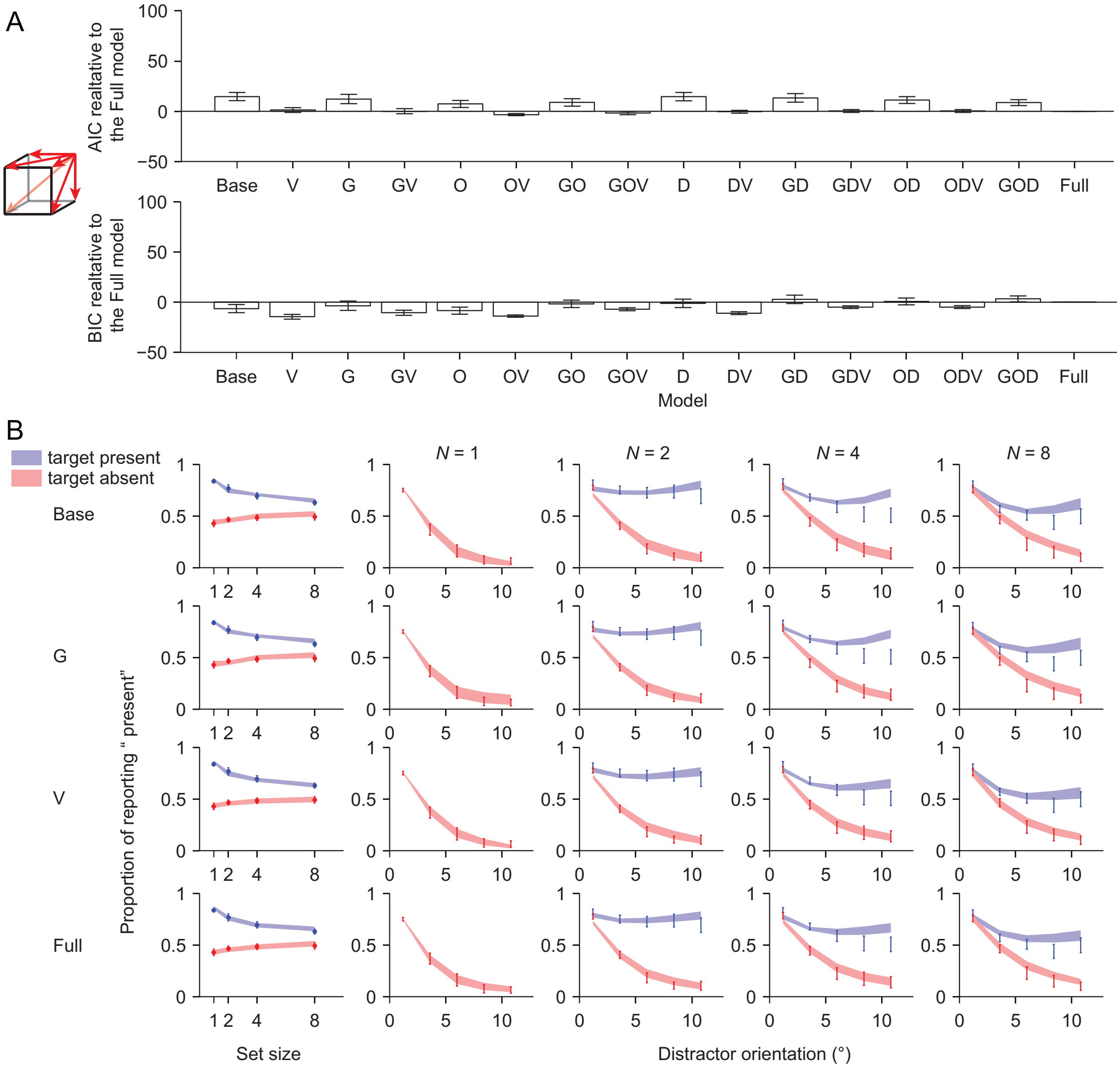


Figure B8. Experiment 8. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting "present" as a function of set size, target presence, and the common orientation of the distractors: data and model fits. See the online article for the color version of this figure.

(*Appendices continue*)

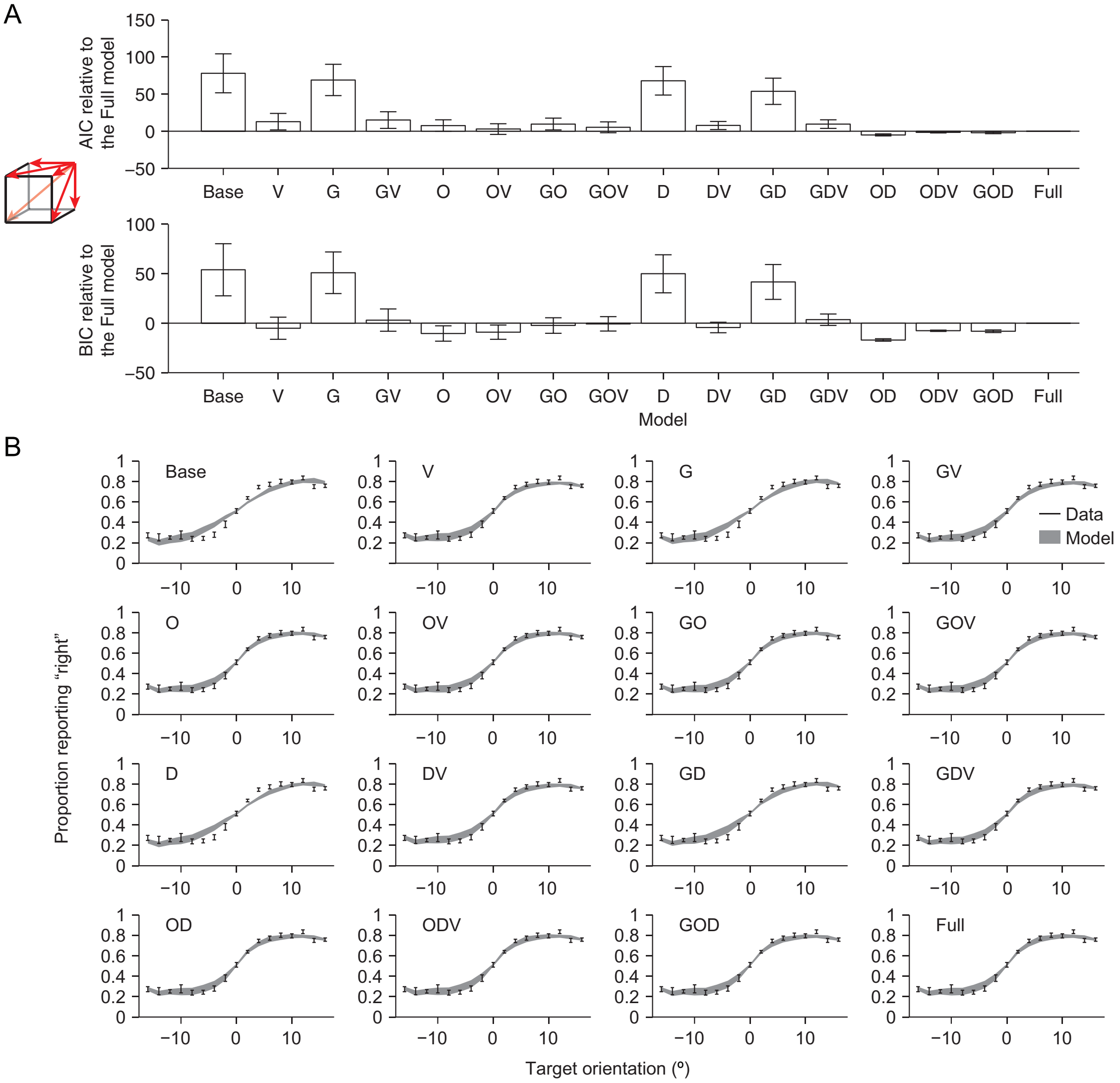


Figure B9. Experiment 9. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting "right" as a function of target orientation. See the online article for the color version of this figure.

(*Appendices continue*)

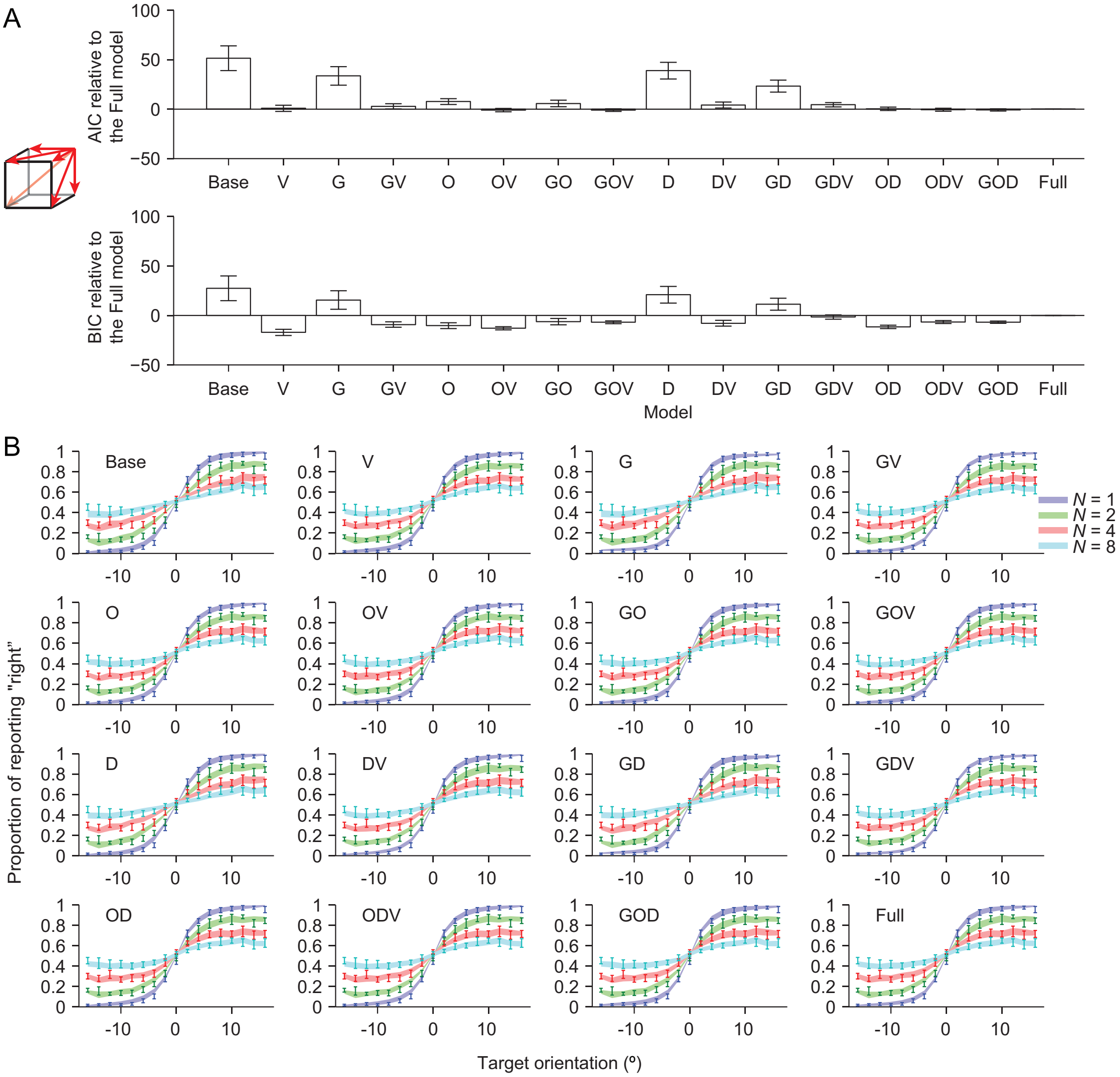


Figure B10. Experiment 10. (A) Complete model comparison. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. (B) Proportion of reporting “right” as a function of target orientation. See the online article for the color version of this figure.

(*Appendices continue*)

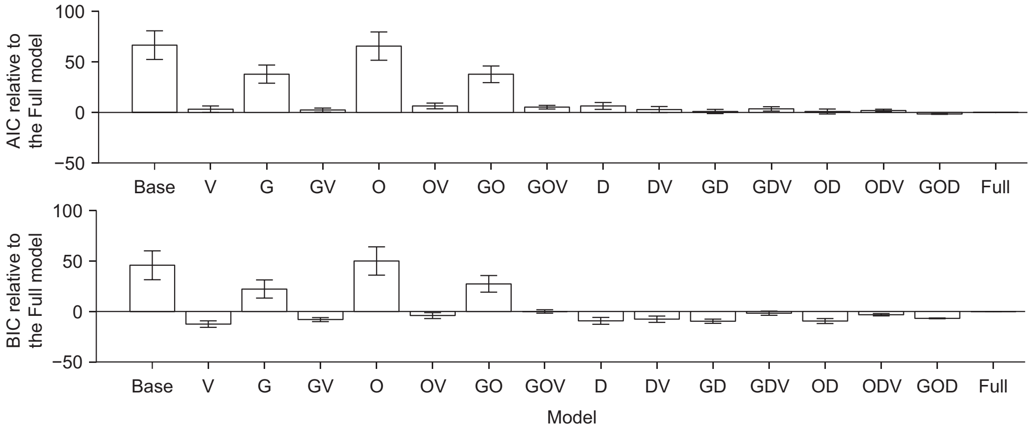


Figure B11. Complete model comparison in Experiment 11. Mean and *SEM* of the difference in Akaike Information Criterion (AIC; top) and Bayesian Information Criterion (BIC; bottom) between each model and the full model. See the online article for the color version of this figure.

(*Appendices continue*)

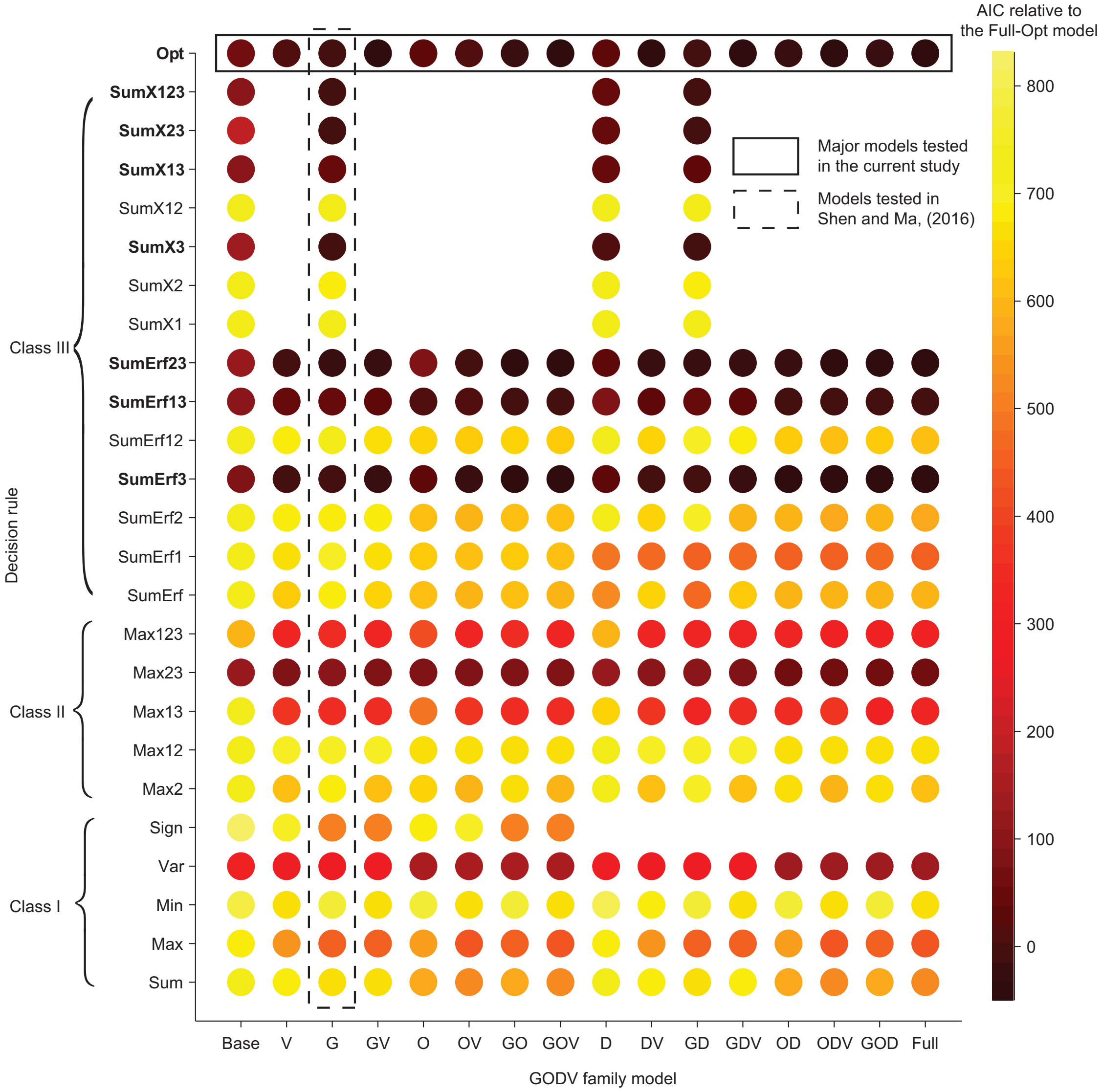


Figure B12. Crossing the suboptimal decision rules with the GODV factor models in Experiment 7. As Figure 10A, but computed with Akaike Information Criterion (AIC). Results are similar to those with Bayesian Information Criterion (BIC). See the online article for the color version of this figure.

(*Appendices continue*)

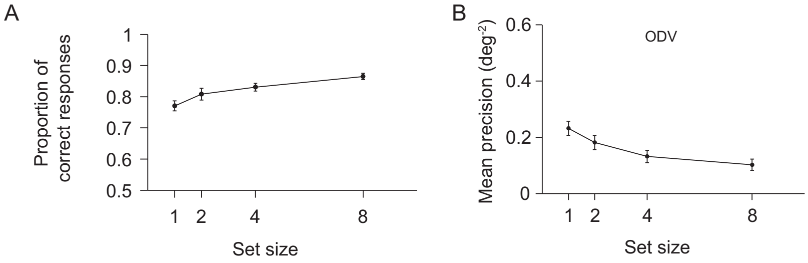
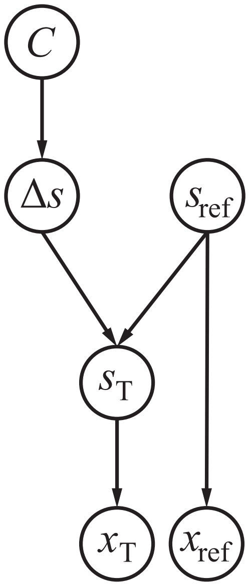


Figure B13. Effects of set size in Experiment 4. Even though proportion correct increases as a function of set size (A), mean precision *decreases* with set size both when estimated with the Full model (Figure 13) and with the ODV model (B). Error bars denote ±1 *SEM*



$$p\left(\Delta s \middle| C\right) = 2 \cdot N\left(\Delta s;\ 0,\ \sigma_{\mathrm{s}}^{2}\right) H\left(C \cdot \Delta s\right)$$

$$p\left(s_{\mathrm{ref}}\right) = U\left(-90,\ 90\right)$$

$$p\left(s_{\mathrm{T}} \middle| s_{\mathrm{ref}}, \Delta s\right) = \delta\left(s_{\mathrm{T}} - s_{\mathrm{ref}} - \Delta s\right)$$

$$p\left(x_{\mathrm{T}} \middle| s_{\mathrm{T}} = s_{\mathrm{ref}} + \Delta s\right) = N\left(x_{\mathrm{T}};\ s_{\mathrm{ref}} + \Delta s,\ \sigma_{\mathrm{T}}^{2}\right)$$

$$p\left(x_{\mathrm{ref}} \middle| s_{\mathrm{ref}}\right) = N\left(\Delta s;\ s_{\mathrm{ref}},\ \sigma_{\mathrm{ref}}^{2}\right)$$

Figure B14. Generative model of Experiment 2. Each node represents a random variable, each arrow a conditional probability distribution. Notations of variables are as follows. C = nature of the world, “clockwise” or “counterclockwise; ” Δs = difference between target orientation and the reference orientation, “clockwise” when positive; s_{ref} = reference orientation; s_{T} = target orientation; x_{ref} = reference measurement; x_{T} = target measurement. Distributions are shown in the equations on the side. $N(x;\ \mu,\ \sigma^2)$ denotes a Gaussian distribution with a mean of μ and a variance of σ^2. $H(x)$ denotes the Heaviside step function. U(a, b) denotes the uniform distribution in a range between a and b. $\delta(x)$ is the Dirac delta function. This diagram specifies the distribution of the measurements, x_{ref} and x_{T}. The optimal observer inverts the generative model and computes the conditional probability of C given x_{ref} and x_{T}.

(*Appendices continue*)

Appendix C

Combinations of the GODV Family With Suboptimal Rules in Experiment 7

In Shen and Ma (2016) where Experiment 7 was first published, we tested three classes of suboptimal rules: Class I contained “simple” suboptimal rules, Class II contained “two-step” rules in which the observer first decides on target location and then reports the tilt of the purported target (thereby ignoring target uncertainty), and Class III encompassed variations of the optimal rule. All decision rules took the form “report” “right” when $d > 0$,” where d is the decision variable.

Here, we created new models by combining these suboptimal decision rules with the GODV factors. Moreover, we included in the derivation of the decision rule a prior probability that the target was tilted right in the models where this was possible (the SumErf models in Class III). To combine the suboptimal rules with factor D, we added Gaussian noise with standard deviation σ_d (a free parameter) to d. We left out several invalid combinations: (a) we combined the sign rule in Class I only with models without factor D, because d in the sign model is a small integer rather than continuous; (b) we combined the SumX rules in Class II only with models base, G, D, and GD, because those rules are only compatible with fixed precision.

Appendix D

Mean and SEM of the Parameter Estimates in the Full-Opt Model for All Experiments

Exp. no.	p_{prior}	$\bar{J}$ (deg^{-2})	λ	β	σ_d	τ (deg^{-2})
1	.485 ± .029	.234 ± .042	.022 ± .012	.96 ± .36	.76 ± .39	.023 ± .013
2	.454 ± .054	.144 ± .011	.079 ± .018	1.32 ± .48	1.5 ± 1.2	.010 ± .007
3	.490 ± .039	.073 ± .011	.025 ± .017	1.68 ± .51	.207 ± .062	.018 ± .011
4	.467 ± .026	.212 ± .028 .185 ± .036 .125 ± .014 .096 ± .014	.015 ± .007	.27 ± .18	.39 ± .21	.048 ± .019
5	.492 ± .023	.153 ± .018	.034 ± .006	.31 ± .10	1.05 ± .19	.010 ± .004
6	.534 ± .039	.233 ± .096 .137 ± .015 .133 ± .018 .121 ± .018	.065 ± .023	.27 ± .17	.98 ± .25	.013 ± .007
7	.496 ± .010	.075 ± .015	.053 ± .023	.20 ± .15	.30 ± .12	.073 ± .017
8	.514 ± .007	.468 ± .070 .261 ± .064 .159 ± .044 .112 ± .027	.039 ± .019	1.13 ± .68	.105 ± .038	.177 ± .031
9	.505 ± .006	.120 ± .076	.0004 ± .0004	1.87 ± .55	.41 ± .18	.002 ± .002
10	.504 ± .005	.164 ± .033 .068 ± .015 .039 ± .012 .020 ± .007	.012 ± .004	.94 ± .29	.182 ± .076	.034 ± .026
11	.470 ± .013	.37 ± .13 .148 ± .066 .080 ± .043 .050 ± .025	.012 ± .003	.146 ± .073	.743 ± .060	.17 ± .16

Note. Confusingly, Liu et al. (1995) also measure “efficiency” by varying the external noise, but in their terminology, any form of inefficiency is purely a consequence of the internal noise.

Received June 22, 2017
Revision received July 11, 2018
Accepted July 15, 2018 ■