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 (Donkink, 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 (Schütz, 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.
(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, Ollkonen, 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 (Attnave, 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.
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/m² and the Gabors had a peak luminance of 35.2 cd/m², a spatial frequency of 3.1 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. Settings were different in Experiments 8 and 11 (see Mazyar et al., 2013): background luminance was 33.1 cd/m² and the Gabors had a peak luminance of 122 cd/m², 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 identical to each other.

Table 1
Overview of Experiments

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

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.

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 equal spacing between −15° and 15°. The subject reported the tilt with respect to vertical of a single oriented stimulus.

Experimental 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.

Experimental 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 $s_{\text{ref}}$ 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 $s_{\text{ref}}$ with a concentration parameter of 10. The subject reported whether the target was oriented clockwise or counterclockwise with respect to $s_{\text{ref}}$.

**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 $-15^\circ$ and $15^\circ$. 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: $0^\circ$, $45^\circ$, $90^\circ$, $135^\circ$, $180^\circ$, $-45^\circ$, $-90^\circ$, and $-135^\circ$. 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 $-5^\circ$ and $5^\circ$.

**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 $20^\circ$ and $20^\circ$. 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 Mazayar 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
Step 1a: Generative Model: Noisy Measurements

Therefore, the most principled choice for the noise distribution is the Gaussian distribution, with \( \mu = 0 \) and a standard deviation \( \sigma \). In this case, the stimulus values can be modeled as independent Gaussian random variables. Hence, the joint probability distribution of the stimulus values can be written as:

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x - \mu)^2}{2\sigma^2}}.
\]

(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,

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

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,

\[
p(x_i|s_i) = \frac{1}{\pi I_0(\kappa)} e^{\kappa \cos(2(x_i - s_i))},
\]

where \( \kappa \) is the concentration parameter, and \( I_0(\cdot) \) is the modified Bessel function of the first kind of order 0.

However, when the stimulus range is narrow relative to the entire \( 2\pi \) 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:

\[
p(x_i|s_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i - s_i)^2}{2\sigma_i^2}}.
\]
A general definition of precision based on $p(x_i|s_i)$ is as Fisher information (Cover & Thomas, 2005). Fisher information, denoted by $J$, is related to the parameters above through

$$J = \frac{1}{\sigma^2} \quad \text{(Gaussian)}$$

$$J = \frac{4\pi I_1(\theta)}{J_0(\theta)} \quad \text{(Von Mises)},$$

(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):

$$\sigma_i = \sigma_o (1 + \beta |\sin(2\chi)|),$$

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

$$J_i = \frac{1}{\sigma_0^2} \frac{1}{(1 + \beta |\sin(2\chi)|)^2}$$

(4)

$$= \frac{1}{\sigma_0^2} \frac{1}{(1 + \beta |\sin(2\chi)|)^2},$$

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 (Fournie 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 assuming a gamma distribution over $J_i$:

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

(5)

where $\bar{J}$ is the mean precision, and $\tau$ 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 $\tau$ (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 $\tau$ (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 $p(x_i|s_i)$, 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 $s$. We assume that the observer’s beliefs about the category distribution $p(C)$ and the category-conditioned stimulus distributions $p(s|C)$ 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 ($p_{\text{right}}$) in the categorization experiments, or that the stimulus was present ($p_{\text{present}}$) in the detection experiments.
- In Experiments 1, 3, 4, 5, and 6 we used discrete stimulus values, for example, 19 values spaced linearly between $-15^\circ$ and $15^\circ$ (Experiments 1 and 3), between $-5^\circ$ and $5^\circ$ (Experiment 4), and between $-20^\circ$ and $20^\circ$ (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:

$$p(s|C) = 2 \cdot N(s; 0, \sigma_C^2)H(C \cdot s_T),$$

(6)

where $\sigma_C$ 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 $x$ 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 $x$:

$$d = \log \frac{p(C = 1|x)}{p(C = -1|x)}.$$
The Bayesian observer without decision noise uses the decision rule to report “C = 1” when \( d \) is positive, or in other words,

\[
\hat{C} = \text{sgn}(d).
\]  

(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:

\[
d = \log \frac{p(C = 1)}{p(C = -1)} + \log \frac{p(x|C = 1)}{p(x|C = -1)}.
\]  

(9)

Now we evaluate the likelihood of each \( C \) by averaging over the stimulus vector \( s \):

\[
p(x|C) = \int p(x|s)p(s|C)ds.
\]  

(10)

This is an example of Bayesian marginalization, in which each possible value of the unknown variable \( s \) is considered.

The derivation so far applies to all experiments. \( p(x|s) \) has been evaluated in Equation (1) or (2). We now evaluate \( p(s|C) \) 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 \( p(s|C) \) as a marginalization over the scalar target orientation \( s_T \):

\[
p(s|C) = \int p(s|s_T)p(s_T|C)ds_T.
\]  

(11)

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

\[
p(s|s_T) = \prod_{L=1}^{N} \delta(s_L - s_T).
\]  

(12)

where \( s_L \) is the stimulus orientation at the \( L \)th location, and \( \delta \) 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 \( p(s|C) \) as a marginalization over the \( N \)-dimensional vector of distractor orientations \( s_D \):

\[
p(s|C = -1) = \int p(s|s_D)p(s_D)ds_D.
\]  

Because all stimuli are distractors, we can evaluate \( p(s|s_D) \) as:

\[
p(s|s_D) = \prod_{L=1}^{N} \delta(s_L - s_{DL}),
\]  

where \( s_{DL} \) is the distractor orientation at the \( L \)th location. The distractor distribution \( p(s_D) \) 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 \( p(s|C) \) as a marginalization over both the scalar target orientation \( s_T \) and the \( (N-1) \)-dimensional vector of distractor orientations \( s_D \):

\[
p(s|C) = \int \int p(s|s_T, s_D)p(s_T|C)p(s_D)ds_Tds_D.
\]  

(13)

Here, \( p(s|s_T, s_D) \) represents the distribution of the stimulus vector \( s \) given \( s_T \) and \( s_D \). Given that exactly one stimulus is the target, we can write this distribution as:

\[
p(s|s_T, s_D) = \frac{1}{N} \sum_{L=1}^{N} \delta(s_L - s_T)\delta(s_L - s_D).
\]  

(14)

where \( s_D \) denotes the vector of all stimuli except the \( L \)th 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 \( p(s_T|C) \) and the distractor distribution \( p(s_D) \) 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 \( p(s|C) \) as a marginalization over both the scalar target orientation \( s_T \) and the scalar reference orientation \( s_{ref} \):

\[
p(s|C) = \int \int p(s|s_T, s_{ref})p(s_T|C)p(s_{ref})ds_Tds_{ref}.
\]  

(15)

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

\[
p(s|s_T, s_{ref}) = \delta(s_{left} - s_T)\delta(s_{right} - s_{ref}).
\]  

(16)

Finally, \( p(s_{ref}) \) is a uniform distribution and \( p(s_T|s_{ref}, C) \) 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 \( \sigma \) when evaluating the decision variable \( d \) (Equation 7), but does not “realize” that there is a relationship between \( \sigma \) and orientation, \( s \). Therefore, the observer does not infer \( s \) from \( \sigma \), for example by marginalizing over \( \sigma \). 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 \( \sigma_d \).
\[ (\tilde{d}|d) = N(\tilde{d}, d, \sigma_\tilde{d}^2). \]  

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

\[ p(\hat{C}|s) = \int p(\hat{C}|\mathbf{x}) p(\mathbf{x}|s) d\mathbf{x}. \]  

Here, \( p(\hat{C}|\mathbf{x}) \) is deterministic and given by Step 2, and \( p(\mathbf{x}|s) \) is given by the measurement distributions in Step 1a. To approximate this integral, we sampled, for each trial in the experiment, a large number of measurement vectors \( \mathbf{x} \) based on the physical stimuli \( s \) on that trial. For each \( \mathbf{x} \), we applied the decision rule from Step 2, and counted the outcomes. The proportions of either category response serve as our approximation of \( p(\hat{C}|s) \). The number of samples of \( \mathbf{x} \) 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 \( \lambda \), so that the probability of reporting \( \hat{C} \) given \( s \) becomes

\[ p_{\text{with guessing}}(\hat{C}|s) = 0.5\lambda + (1 - \lambda)p(\hat{C}|s). \]  

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 combination 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

<table>
<thead>
<tr>
<th>Models, Model Factors, and Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model or model factor</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Model: Base</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Factor: Guessing, G</td>
</tr>
<tr>
<td>Factor: Oblique effect, O</td>
</tr>
<tr>
<td>Factor: Decision noise, D</td>
</tr>
<tr>
<td>Factor: Residual variable precision, V</td>
</tr>
<tr>
<td>Model: Full or GODV</td>
</tr>
</tbody>
</table>

**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:

\[ \text{AIC}(M) = 2k_M - 2\text{LL}_{\text{max}}(M), \]

\[ \text{BIC}(M) = \log(N_{\text{trials}})k_M - 2\text{LL}_{\text{max}}(M), \]

where \( k_M \) is the number of parameters of model \( M \) and \( N_{\text{trials}} \) 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 + \( F \)”: \[
\text{KID}_{AIC}(F) = AIC(\text{Base}) - AIC(\text{Base} + F).
\] (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}_{AIC}(F) = AIC(\text{Full-}\text{\( F \))} - AIC(\text{\( Full \))).
\] (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:

\[
\text{LFLR}(F) = \log \frac{p(M | F, \text{data})}{p(M | \text{data})}.
\]
\[ L(F \text{ present}) = p(\text{data} | F \text{ present}) = \sum_{\text{all models } M} p(\text{data} | M) p(M | F \text{ present}). \]

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

\[ L(F \text{ present}) \approx \sum_{M \text{ tested}} \frac{1}{\text{number of tested models that contain } F} p(\text{data} | M), \]

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

\[ \log \text{marginal likelihood} = -0.5 \text{AIC}(M) \]

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 \(\text{LFLR}_{\text{AIC}}\) as an example. Starting from Equation (23),

\[ \text{LFLR}_{\text{AIC}}(F) = \log \sum_{M_{F=1}} e^{-0.5 \text{AIC}(M)} = \log \max_{M_{F=1}} e^{-0.5 \text{AIC}(M)} . \]

If furthermore, the lowest-AIC model is the most highly parameterized model, then \(2 \cdot \text{LFLR}_{\text{AIC}}\) becomes identical to \(\text{KOD}_{\text{AIC}}\).
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 Factor Importance Metrics $\text{FIM}_{\text{AIC}}$ and $\text{FIM}_{\text{BIC}}$; 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), and 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 $\text{KD}_{\text{AIC}}(V)$ and mean $\text{KD}_{\text{BIC}}(V)$ were both greater than 9.2 (Figure 4A), indicating very strong evidence that factor V is useful. 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 results of the Figures 1 to 11 in Appendix B correspond to the experiments numbers). In Experiments 2 and 4, $\text{KID}_{\text{AIC}}(V)$ and $\text{KID}_{\text{BIC}}(V)$ were much smaller, indicating little or no evidence that factor V is useful.

In Experiments 7, 9, 10, and 11, mean $\text{KD}_{\text{AIC}}(V)$ and mean $\text{KD}_{\text{BIC}}(V)$ were both greater than 9.2, indicating very strong evidence that factor V is necessary to explain the data (Figure 4B).

<table>
<thead>
<tr>
<th>$\text{min}(\text{FIM}<em>{\text{AIC}}(F), \text{FIM}</em>{\text{BIC}}(F))$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&gt;9.2$</td>
<td>Very strong evidence that $F$ is important</td>
</tr>
<tr>
<td>6.8 to 9.2</td>
<td>Strong evidence that $F$ is important</td>
</tr>
<tr>
<td>4.6 to 6.8</td>
<td>Moderate evidence that $F$ is important</td>
</tr>
<tr>
<td>$&lt;4.6$ (including $&lt;0$)</td>
<td>Little or no evidence that $F$ is important</td>
</tr>
</tbody>
</table>

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

### 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 easier to discriminate 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 1–6, with mean $\text{KOD}_{\text{AIC}}(V)$ and mean $\text{KOD}_{\text{BIC}}(V)$ both smaller than 4.6. The common feature of these experiments was that there were either no distractors (Experiments 1–6) 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.
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 2LFLR (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.
± 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). 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 GV. 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
the stimulus range was the same but set size was equal to 4, factor G was no longer necessary, but \(2 \cdot \text{LFLR}(G)\) was \(5.7 \pm 2.1\) (AIC), and \(5.1 \pm 2.2\) (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 \(-5^\circ\) and \(5^\circ\)), 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
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 LL_{max} of the with-factor model (G) could be identical to the LL_{max} of the without-factor model (base) even though the factor is present. In another example, in Experiments 9–11, V might trade off against O and/or D, and the weaker evidence for factor V might be due to stronger evidence for factors O (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 τ = 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...
the observer has learned and correctly incorporates the experimental statistics, in our case the joint distribution \( p(C, s) \). 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, & Fougne, 2018), or instead use natural statistics (Adams, Graf, & Ernst, 2004; Girshick et al., 2011; Zhu & Ma, 2017).

In the present section, we undertook 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\textsubscript{ASC}(V) and 2 × LFLR\textsubscript{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),

---

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

<table>
<thead>
<tr>
<th>Experiment number</th>
<th>Max set size</th>
<th>Feature set size</th>
<th>Task target</th>
<th>Task distractors</th>
<th>Target distribution or range</th>
<th>Decision noise (D)</th>
<th>Oblique effect (O)</th>
<th>Residual variable (V)</th>
<th>p(C, s)</th>
<th>Residual variable (V)</th>
<th>\text{LFLR}_{\text{BIC}}(V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>Catg 15</td>
<td>Catg 15</td>
<td>((-90, 90))</td>
<td>0.31; 0.12; 0.11</td>
<td>2.8; 1.1</td>
<td>0.32; 0.32</td>
<td>0.12</td>
<td>0.32; 0.32</td>
<td>2.8; 1.1</td>
</tr>
</tbody>
</table>
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 $p(L|x)$, 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 decision-maker 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 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.

Figure 9. Trade-offs between parameters. (A) Trade-off between precision $J$ and guessing rate $\lambda$. 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 $\lambda$. 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 $\beta$) and V (parameterized by $\tau$). 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 $\beta$ and $\tau$. Many combinations have a high log likelihood, including a combination of non-zero $\beta$ and a value of $\tau$ lower than the true value. See the online article for the color version of this figure.
Major models tested in the current study

Models tested in Shen and Ma, (2016)

BIC relative to the Full-Opt model

Class III

Class II

Class I

Decision rule

A

GODV family member

SumX123
SumX23
SumX13
SumX12
SumX3
SumX2
SumX1
SumErf23
SumErf13
SumErf12
SumErf3
SumErf2
SumErf1
SumErf
Max123
Max23
Max13
Max12
Max1
Max
Sign
Var
Min
Max
Sum

Base V G GV O OV GO GOV D DV GD GDV OD ODV GOD Full

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 $5^\circ$, 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 $\ell$-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.
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.

**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.

---

**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.

**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.
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.
<table>
<thead>
<tr>
<th>References</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wellling, C. Cortes, N. D. Lawrence, &amp; K. Q. Weinberger (Eds.), Advances</td>
<td>10.1038/nn.2842</td>
</tr>
<tr>
<td>doi.org/10.1016/j.visres.2005466</td>
<td></td>
</tr>
</tbody>
</table>

Appendix A

Decision Rules

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

Notation:
- $s T$: target orientation
- $s D$ or $s D$: 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 $i$th stimulus
- $\sigma i$: measurement noise of the $i$th stimulus, standard deviation of the Gaussian distribution
- $\sigma T$: measurement noise of the target stimulus, standard deviation of the Gaussian distribution
- $\sigma ref$: measurement noise of the reference stimulus, standard deviation of the Gaussian distribution
- $\sigma$: standard deviation to generate the stimulus orientations
- $J i$: encoding precision of the $i$th stimulus
- $J$: precision to generate stimulus orientations
- $\kappa i$: concentration parameter of Von Mises distribution of the $i$th stimulus
- $\kappa$: concentration parameter of the Von Mises distribution to generate stimulus orientations
- $\kappa T$: concentration parameter of the Von Mises distribution to generate target orientations
- $L$: hypothesized target location
- $N$: set size
- $p right, p clockwise, or p present$: prior probability of reporting “right,” “clockwise,” or “present”
- $\Phi(\cdot)$: standard normal cumulative function
- $\delta(\cdot)$: Dirac delta function
- $N(\cdot; \mu, \sigma^2)$: Gaussian distribution with a mean of $\mu$ and a variance of $\sigma^2$
- $H(\cdot)$: Heaviside step function
- $U(a,b)$: uniform distribution in a range between $a$ and $b$
- $VM(\cdot; s, \kappa)$: Von Mises distribution with a mean of $s$ and concentration parameter of $\kappa$

(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 \( s_x \). The distribution \( p(s|C) \) 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

\[
d = \log \frac{\Phi\left(\frac{y}{\sqrt{J + y}}\right)}{\Phi\left(\frac{y}{\sqrt{J + y}}\right)} + \log \frac{p_{\text{right}}}{1 - p_{\text{right}}}.\]

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:

\[
p(\Delta s|C) = 2 \cdot N(\Delta s; 0, \sigma_2^2)H(C \cdot \Delta s)
\]

\[
p(s_{\text{ref}}) = U(-90, 90)
\]

\[
p(x_t|s_{\text{ref}} + \Delta s) = \delta(s_x - s_{\text{ref}} - \Delta s)
\]

\[
(25), \quad \text{and} \quad p(x_t|s_{\text{ref}}) = N(x_t; s_{\text{ref}}, \sigma_1^2).
\]

We use Gaussian distributions to model \( p(\Delta s|C), p(x_{\text{ref}}|s_{\text{ref}}) \) and \( p(x_t|s_{\text{ref}}) \), 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:

\[
p(x|C) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_t|s_{\text{ref}})p(x_{\text{ref}}|s_{\text{ref}})p(s_{\text{ref}}|s_{\text{ref}})\times p(\Delta s|C)ds_{\text{ref}}ds_{\text{ref}}d(\Delta s)
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_t|s_{\text{ref}})p(x_{\text{ref}}|s_{\text{ref}})\delta(s_x - s_{\text{ref}} - \Delta s)p(s_{\text{ref}})\times p(\Delta s|C)ds_{\text{ref}}ds_{\text{ref}}d(\Delta s)
\]

\[
= \int_{-\infty}^{\infty} p(x_t|s_{\text{ref}} + \Delta s)p(x_{\text{ref}}|s_{\text{ref}})p(s_{\text{ref}})\times p(\Delta s|C)ds_{\text{ref}}ds_{\text{ref}}d(\Delta s)
\]

\[
= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} p(x_t|s_{\text{ref}} + \Delta s)p(x_{\text{ref}}|s_{\text{ref}})p(s_{\text{ref}})ds_{\text{ref}} \right)\times p(\Delta s|C)d(\Delta s).
\]

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

\[
p(x|C) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} N(x_t; s_{\text{ref}} + \Delta s, \sigma_2^2)N(x_{\text{ref}}; s_{\text{ref}}, \sigma_1^2)\frac{1}{180} ds_{\text{ref}} \right)^2 N(\Delta s; 0, \sigma_2^2)H(C \cdot \Delta s)d(\Delta s).
\]

Numerically computing the double integral is not computation-ally feasible: It happens inside a Monte Carlo simulation of \( x \), which happens inside a parameter optimization, which happens inside a large model comparison. Therefore, we approximate partial integrals with the full integral:

\[
p(x|C) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} N(x_t; s_{\text{ref}} + \Delta s, \sigma_2^2)N(x_{\text{ref}}; s_{\text{ref}}, \sigma_1^2)\frac{1}{180} ds_{\text{ref}} \right)^2 N(\Delta s; 0, \sigma_2^2)H(C \cdot \Delta s)d(\Delta s).
\]

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

\[
p(x|C) = \frac{1}{180} \int_{-\infty}^{\infty} N(x_t - s_{\text{ref}}; \Delta s, \sigma_2^2 + \sigma_1^2)2 \cdot N(\Delta s; 0, \sigma_2^2)H(C \cdot \Delta s)d(\Delta s).
\]

By further simplification, the final decision variable is

\[
d = \log \frac{\Phi\left(\frac{\Delta s}{\sqrt{2J_T + J_c}}\right)}{\Phi\left(\frac{\Delta s}{\sqrt{2J_T + J_c}}\right)} + \log \frac{p_{\text{clockwise}}}{1 - p_{\text{clockwise}}}.
\]

where \( J_c = \frac{1}{J_T + J_c} \).

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 \( p(s|C) \) takes the form

\[
p(s|C) = \int p(s|x_t)p(s_t|C)ds_t
\]

\[
= \delta(s - s_t)p(s_t|C)ds_t,
\]

where \( 1 \) is the vector of 1s and \( p(s_t|C) \) is a truncated Gaussian distribution (Equation 6).

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

\[
d = \log \frac{\Phi\left(\sum_{t=1}^{\infty} x_t\right)}{\sqrt{\left(\sum_{t=1}^{\infty} J_t\right) + J_c} + \log \frac{p_{\text{right}}}{1 - p_{\text{right}}}.
\]

(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 \( p(s|C) \) takes the form

\[
p(s|C) = \int \int p(s|s_I, s_D)p(s_D|s_I)p(s_I)ds_I ds_D = \int \frac{1}{N} \sum_{l=1}^{N} \delta(s_l - s_I)\delta(s_d - s_D)p(s_I|C)p(s_D)ds_I ds_D.
\]

(27)

Given that all distractors are vertical, \( p(s_D) \) is equal to \( \prod_i \delta(s_D) \). Therefore, Equation (27) is further simplified as

\[
p(s|C) = \frac{1}{N} \sum_{l=1}^{N} \int \delta(s_l - s_I)\delta(s_d - s_D) p(s_I|C) ds_I.
\]

(28)

where \( p(s_I|C) \) is a truncated Gaussian distribution (Equation 6).

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

\[
d = \log \frac{\prod_i \exp \left( \frac{1}{2} \left( \sum_{l=1}^{N} \frac{s_l}{\sqrt{J_l + J_s}} \right)^2 \right)}{1 - \prod_i \exp \left( \frac{1}{2} \left( \sum_{l=1}^{N} \frac{s_l}{\sqrt{J_l + J_s}} \right)^2 \right)} + \log \frac{P_{\text{right}}}{1 - P_{\text{right}}}.
\]

Experiment 7: Categorization of a Single Target With Homogeneous Distractors

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

\[
p(s|C) = \frac{1}{N} \sum_{l=1}^{N} \int \delta(s_l - s_I)\delta(s_d - s_D)p(s_I|C)p(s_D)ds_l ds_D.
\]

(29)

where \( p(s_I|C) \) is a truncated Gaussian distribution (Equation 6) and \( p(s_D) \) is a Gaussian distribution with a mean of 0 and a standard deviation of \( \sigma_v \).

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

\[
d = \log \frac{\prod_i \exp \left( \frac{1}{2} \left( \sum_{l=1}^{N} \frac{s_l}{\sqrt{J_l + J_s}} \right)^2 \right)}{1 - \prod_i \exp \left( \frac{1}{2} \left( \sum_{l=1}^{N} \frac{s_l}{\sqrt{J_l + J_s}} \right)^2 \right)} + \log \frac{P_{\text{right}}}{1 - P_{\text{right}}}.
\]

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 \( p(s|C = 1) \) 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 \( p(s|C = -1) \) takes the form:

\[
p(s|C = -1) = \int p(s|s_I)p(s_D)ds_D = \int \delta(s - s_D) p(s_D)ds_D,
\]

(30)

where \( p(s_D) \) is a Gaussian distribution with a mean of 0 and a standard deviation of \( \sigma_v \).

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

\[
d = \log \frac{\prod_i \exp \left( \frac{1}{2} \left( \sum_{l=1}^{N} \frac{s_l}{\sqrt{J_l + J_s}} \right)^2 \right)}{1 - \prod_i \exp \left( \frac{1}{2} \left( \sum_{l=1}^{N} \frac{s_l}{\sqrt{J_l + J_s}} \right)^2 \right)} + \log \frac{P_{\text{present}}}{1 - P_{\text{present}}}.
\]

(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 $p(s|C)$ takes the same form as in Equation (27), where $p(s_t|C)$ is a truncated Von Mises distribution:

$$p(s_t|C) = 2 \cdot \text{VM}(2x_T; 0, \kappa_T)H(C \cdot s_t)$$

(31)

and $p(s_D)$ is a product of $N$-1 uniform distributions.

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

$$d = \log \frac{\sum_t \int \frac{1}{2} \text{VM}(2x_T; 2x_T, \kappa_T) \text{VM}(2x_T; 0, \kappa_T) ds_T}{\sum_t \int \frac{1}{2} \text{VM}(2x_T; 2x_T, \kappa_T) \text{VM}(2x_T; 0, \kappa_T) ds_T} + \log \frac{p_{\text{left}}}{1 - p_{\text{left}}}.$$ 

Experiment 11: Detection of a Single Target With Heterogeneous Distractors

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

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

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

(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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.

\[
p(\Delta s|C) = 2 \cdot N(\Delta s; 0, \sigma^2) H(C \cdot \Delta s) \\
p(s_{ref}) = U(-90, 90) \\
p(s_{ref}, \Delta s) = \delta(s_{ref} - s_{ref} - \Delta s) \\
p(x_T|s_{ref} + \Delta s) = N(x_T; s_{ref} + \Delta s, \sigma^2) \\
p(x_{ref}|s_{ref}) = N(\Delta s; s_{ref}, \sigma^2)
\]

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;" \(\Delta s\) = difference between target orientation and the reference orientation, "clockwise" when positive; \(s_{ref}\) = reference orientation; \(s_T\) = target orientation; \(s_{ref}\) = 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 \(\mu\) and a variance of \(\sigma^2\). \(H(x)\) 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 \(x_T\).
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 \( d/d_{1022} < 0 \), 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 \( \sigma_d \) (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

<table>
<thead>
<tr>
<th>Exp. no.</th>
<th>( p_{\text{prior}} )</th>
<th>( J ) (deg(^{-2}))</th>
<th>( \lambda )</th>
<th>( \beta )</th>
<th>( \sigma_d )</th>
<th>( \tau ) (deg(^{-2}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.485 ± .029</td>
<td>.234 ± .042</td>
<td>.022 ± .012</td>
<td>.96 ± .36</td>
<td>.76 ± .39</td>
<td>.023 ± .013</td>
</tr>
<tr>
<td>2</td>
<td>.454 ± .054</td>
<td>.144 ± .011</td>
<td>.079 ± .018</td>
<td>1.32 ± .48</td>
<td>1.5 ± 1.2</td>
<td>.010 ± .007</td>
</tr>
<tr>
<td>3</td>
<td>.490 ± .039</td>
<td>.073 ± .011</td>
<td>.025 ± .017</td>
<td>1.68 ± .51</td>
<td>.207 ± .062</td>
<td>.018 ± .011</td>
</tr>
<tr>
<td>4</td>
<td>.467 ± .026</td>
<td>.212 ± .028</td>
<td>.015 ± .007</td>
<td>.27 ± .18</td>
<td>.39 ± .21</td>
<td>.048 ± .019</td>
</tr>
<tr>
<td>5</td>
<td>.492 ± .023</td>
<td>.153 ± .018</td>
<td>.034 ± .006</td>
<td>.31 ± .10</td>
<td>1.05 ± .19</td>
<td>.010 ± .004</td>
</tr>
<tr>
<td>6</td>
<td>.534 ± .039</td>
<td>.233 ± .096</td>
<td>.065 ± .023</td>
<td>.27 ± .17</td>
<td>.98 ± .25</td>
<td>.013 ± .007</td>
</tr>
<tr>
<td>7</td>
<td>.496 ± .010</td>
<td>.075 ± .015</td>
<td>.053 ± .023</td>
<td>.20 ± .15</td>
<td>.30 ± .12</td>
<td>.073 ± .017</td>
</tr>
<tr>
<td>8</td>
<td>.514 ± .007</td>
<td>.468 ± .070</td>
<td>.039 ± .019</td>
<td>1.13 ± .68</td>
<td>.105 ± .038</td>
<td>.177 ± .031</td>
</tr>
<tr>
<td>9</td>
<td>.505 ± .006</td>
<td>.120 ± .076</td>
<td>.0004 ± .0004</td>
<td>.187 ± .55</td>
<td>.41 ± .18</td>
<td>.002 ± .002</td>
</tr>
<tr>
<td>10</td>
<td>.504 ± .005</td>
<td>.164 ± .033</td>
<td>.012 ± .004</td>
<td>.94 ± .29</td>
<td>.182 ± .076</td>
<td>.034 ± .026</td>
</tr>
<tr>
<td>11</td>
<td>.470 ± .013</td>
<td>.37 ± .13</td>
<td>.012 ± .003</td>
<td>.146 ± .073</td>
<td>.743 ± .060</td>
<td>.17 ± .16</td>
</tr>
</tbody>
</table>

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.