Recognition criteria vary with fluctuating uncertainty

Joshua A. Solomon
Optometry Division, Applied Vision Research Centre, City University London, UK

Patrick Cavanagh
Laboratoire Psychologie de la Perception, Université Paris Descartes and CNRS, Paris, France

Andrei Gorea
Laboratoire Psychologie de la Perception, Université Paris Descartes and CNRS, Paris, France

In distinct experiments we examined memories for orientation and size. After viewing a randomly oriented Gabor patch (or a plain white disk of random size), observers were given unlimited time to reproduce as faithfully as possible the orientation (or size) of that *standard* stimulus with an adjustable Gabor patch (or disk). Then, with this *match* stimulus still in view, a recognition *probe* was presented. On half the trials, this probe was identical to the standard. We expected observers to classify the probe (a same/different task) on the basis of its difference from the match, which should have served as an explicit memory of the standard. Observers did better than that. Larger differences were classified as "same" when probe and standard were indeed identical. In some cases, recognition performance exceeded that of a simulated observer subject to the same matching errors, but forced to adopt the single most advantageous criterion difference between the probe and match. Recognition must have used information that was not or could not be exploited in the reproduction phase. One possible source for that information is observers' confidence in their reproduction (e.g., in their memory of the standard). Simulations confirm the enhancement of recognition performance when decision criteria are adjusted trial-by-trial, on the basis of the observer's estimated reproduction error.

Keywords: orientation, size, reproduction, recognition, confidence, doubly stochastic noise, decision criteria, accuracy

Citation: Solomon, J. A., Cavanagh, P., & Gorea, A. (2012). Recognition criteria vary with fluctuating uncertainty. *Journal of Vision*, *12*(8):2, 1–13, http://www.journalofvision.org/content/12/8/2, doi:10.1167/12.8.2.

Introduction

In conventional recognition experiments, observers try their best to remember a standard stimulus, but memory is not perfect. The precision of those memories may be inferred from their recognition responses using Signal Detection Theory (SDT) (Green & Swets, 1966), but with one crucial assumption; namely that observers use the same decision strategy (i.e., criterion) on every trial. Sophisticated theorists (notably Wickelgren, 1968) have already acknowledged that decision criteria may fluctuate. The problem remains that this fluctuation is indistinguishable from memory noise in most paradigms. Evidence that decision criteria depend systematically on trial-by-trial feedback was first presented by Tanner, Rauk, and Atkinson (1970). Treisman and Williams (1984) later codified several previously reported sequential effects in their Criterion Setting Theory. Our own research is situated within this alternative tradition. We introduce a paradigm in which criterion fluctuation demonstrably improves recognition performance.

In our paradigm, criterion fluctuation can be advantageous only if it complements a similar fluctuation in the precision of memory. After 150 years (see p. 82 of Fechner, 1860 & 1966), this idea of variable precision is only now finding its way into signal-detection models (notably that of van den Berg, Shin, Chou, George, & Ma, 2012). To get a handle on the trial-by-trial fluctuation in memory noise, we required observers to match their memory of each standard with an adjustable stimulus. We then computed the best possible recognition performances, assuming that reproduction errors reflected a memory noise with constant precision. The fact that our observers' recognition performances exceeded these predictions allows us to infer that memory noise has variable precision. This would be insufficient to allow better-than-prediction recognition if observers' decision criteria did not covary with their memory noise.

Our empirical approach (summed up in the Abstract and detailed in the Methods) consisted of a matching (recollection) task followed by a same/different (recognition) task. This approach allows us to characterize decision strategies for recognition with respect to explicit representations of the imperfectly recollected

doi: 10.1167/12.8.2 Received December 19, 2011; published August 6, 2012 ISSN 1534-7362 © 2012 ARVO

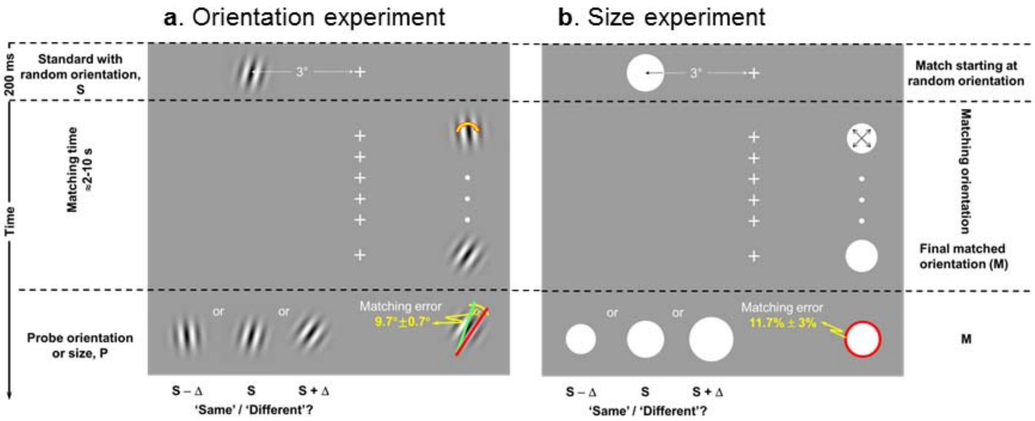
a. Orientation experiment
b. Size experiment
200 ms
Standard with random orientation, S
3°
Match starting at random orientation
Time
Matching time ≈2-10 s
Matching orientation
Final matched orientation (M)
Probe orientation or size, P
or
or
Matching error 9.7°±0.7°
Matching error 11.7% ± 3%
M
S – Δ
S
S + Δ
'Same' / 'Different'?

Figure 1. Spatiotemporal layout of the displayed and matched stimuli in the orientation (a) and size (b) experiments. In this diagram the size of the fixation cross and its distance from the standard have been scaled to twice their actual values.

standards. Such a feat is impossible with standard yes/no procedures. As will be shown, our observers' decision behavior is consistent with a doubly stochastic memory noise model where observers modulate their decision criterion in the recognition phase in inverse proportion with their confidence in their match. To our knowledge, the present experiments are the first to provide empirical evidence of a systematic, trial-by-trial modulation of recognition strategy in accordance with observers' confidence in each memory of the standard stimulus.

Methods

Units

In this paper, all Gabor orientations (s, m, m_0, and p; see below) have units of degree, clockwise with respect to horizontal. All disk sizes (also s, m, m_0, and p) are quantified as logarithms (base 10) of diameter.

Stimuli

Gabor patches or white disks (see Figure 1) were presented on a 15" MacBook Pro computer at a comfortable viewing distance of about 0.5 m. At this distance, each Gabor was centered 3° of visual angle to the left or right of a central fixation spot. The white disks were displayed at the same mean eccentricity but their positions were randomly jittered, across and within trials, ±1°. Each Gabor was the product of a sinusoidal luminance grating and a Gaussian luminance envelope. The grating had a spatial frequency of 1.5 cycles per degree and a random spatial phase. The Gaussian envelope had a space constant (σ) of 0.5 degrees of visual angle. Both grating and envelope had maximum contrast. The white disks had a luminance of 260 cd/m^2. Stimuli were presented on a grey, 60 cd/m^2 background.

Procedure

On each trial, the *standard*, a randomly oriented Gabor (or a random size white disk), was briefly (200 ms) presented on one side of fixation. The observer then attempted to reproduce its orientation (or size, s) by manipulating another stimulus (the *match*), subsequently presented on the opposite side of fixation. Just like the standard, the match's initial orientation (or size, m_0) was randomly selected from a uniform distribution over all orientations (or diameters between 1.5° and 3.0°). Each press of the "c" key rotated the match 2° anticlockwise (or reduced its diameter by 2%) and each press of the "m" key rotated it 2° clockwise (or increased its diameter by 2%).[1] Gabor phase was randomly reselected with each keypress. To indicate satisfaction with the match's orientation (or size, m), the observer pressed the space bar, initiating the trial's second, recognition phase. With the match still in view, a *probe* Gabor (or disk) was presented at the location of the standard. On 50% of trials the orientation (or size, p) of the probe was identical to s. In the remaining trials the orientation (or size) of p was changed with respect to s by a value, $\pm\Delta$s.[2] The value of Δs was held constant within each block of trials. In the orientation experiment, Δs took values of 3°, 5°, 7°, 14°, and 21°. In the size experiment, Δs took values of 0.04, 0.06, and 0.08. Observers had to classify p as either "same" or "different" with respect to their memory of s. No feedback was given. Observers performed two blocks of 100 trials at each level of difficulty in a random order. Two additional 50-trial blocks (one with $\Delta s = 3°$, the other with $\Delta s = 5°$) were run by the last author in the

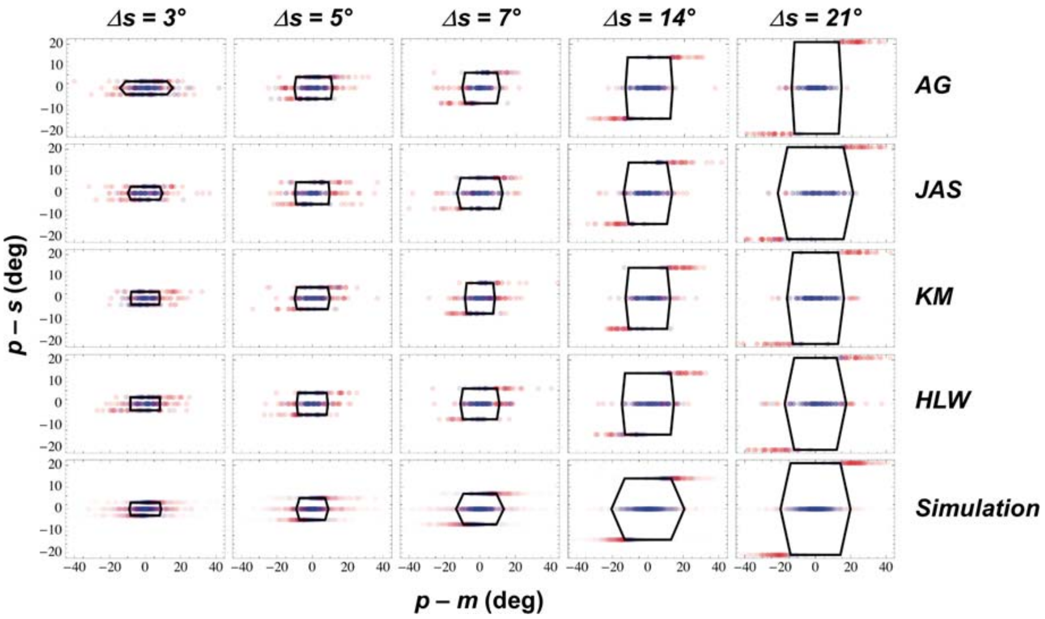

Figure 2. Orientation experiment. Distributions of "same" and "different" responses (blue and red symbols respectively) of each of the four observers (top four rows of panels) and of a variable precision model (bottom row) as a function of the orientation difference ($p - m$) between probe and match for each of the five difficulty levels (columns) and for the cases where $p > s$, $p = s$, and $p < s$ (rows of symbols from top to bottom within each panel). Each panel in the top four rows shows 200 responses/trials. Each panel in the bottom row shows 2000 simulated trials. Each regular hexagon connects the points where Pr("same") = Pr("different"), i.e., the decision criteria, as estimated from a maximum-likelihood fit of two cumulative Normal distributions (see text).

orientation experiment and were included in all subsequent analyses.

Observers

Two authors and two naïve observers were run in the orientation experiment. The same two authors and three naïve observers (one of which, KM, also participated in the orientation experiment) were run in the size experiment.

Results

Gross indices of reproduction can be obtained by computing the standard deviation of each observer's matching errors ($s - m$). Across observers, these indices had a mean and standard deviation (SD) of 9.6° and 0.7°, respectively, for orientation,[3] and 0.048 and 0.013, respectively, for size (see Tables SM1 and SM2 in Supplementary material). Comparable values for orientation have been obtained under similar conditions (e.g., Tomassini, Morgan, & Solomon, 2010), but more precise reproduction has also been reported with different stimuli and procedures (e.g., Vandenbussche, Vogels, & Orban, 1986). Comparable values for size have also been obtained under similar conditions (e.g., Solomon, Morgan, & Chubb, 2011; note that the value 0.048 corresponds to a Weber fraction of 12% for diameter).

Better indices of reproduction can be obtained by segregating variable error from constant error. This segregation is described in detail in Appendix A. Here, we merely offer the following summary: the variable error for orientation depended upon the standard orientation, such that the error tended to be larger when the standard was further from the cardinal axes. This finding has been coined the *oblique effect* (e.g., Appelle, 1972). The variable error for size, on the other hand, neither increased nor decreased with standard size. This invariance with standard size has become known as Weber's Law (Fechner, 1860 & 1966).

Figures 2 and 3 present all of the raw recognition data (in the orientation and size experiments, respectively), together with simulated data (bottom row in each figure) from the variable precision model described below. Each panel shows the 200 "same" (blue symbols) and "different" (red symbols) responses obtained from a single observer at a single level of difficulty (Δ*s*).

If the recognition process accessed only the very same information as that used during the reproduction phase, then recognition performance should depend exclusively on the difference $p - m$. As an alternative to this (null) hypothesis, we considered the possibility that observers' same/different decisions also depended upon

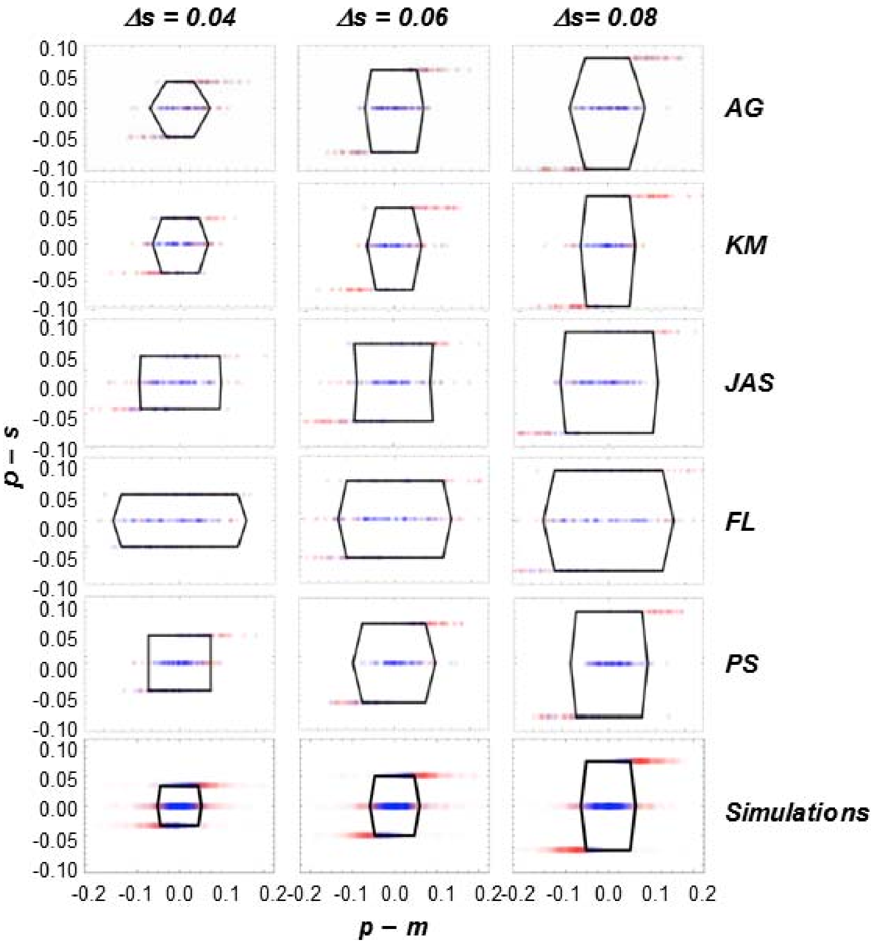

Figure 3. Size experiment. Notation and symbols are as in Figure 2. Data are shown this time for five observers (first five rows of panels) together with the simulations from the variable precision model (bottom) row. As for the orientation experiment, the three rows of data (blue and red symbols) in each panel are for the cases where $p > s$, $p = s$, and $p < s$.

probe identity (i.e., the difference $p - s$). As will be discussed below, an interaction between $p - m$ and $p - s$ may have arisen from observers modulating their decision criteria according to their confidence in each match.

To evaluate these hypotheses, we analyzed observers' responses separately for $p = s$ trials and $p \neq s$ trials. (Within each panel, $p = s$ trials are represented by the middle row of symbols; $p \neq s$ trials are represented by the top and bottom rows.) For each observer, recognition responses were also segregated according to the level of difficulty (Δs). They were then maximum-likelihood fit with psychometric functions based on the cumulative Normal distribution:

$$\Pr(\text{“different”}) = \Phi[(|p - m| - \mu)/\sigma] \tag{1}$$

Finally, a series of hypotheses regarding the *decision criterion* (μ) and the *SD* of its fluctuation (σ) were subjected to chi-square tests based on the generalized likelihood ratio (Mood, Graybill, & Boes, 1974). In the data from both experiments, we found significant (at the $\alpha = 0.05$ level) changes in the criterion with observer, difficulty, and probe identity, but changes in the standard deviation of its fluctuation were significant only when observer and difficulty changed, not when the probe identity changed. Therefore, we constrained $\sigma_{P \neq S} = \sigma_{P=S}$ for all of our fits. The best-fitting parameter values are available in Supplementary Material.

In Figures 2 and 3, each hexagon connects the derived decision criteria, μ (i.e., the $p - m$ values yielding equal proportions of “same” and “different” responses).[4] With one exception out of 20 cases in the orientation experiment (HLW, $\Delta s = 3°$) and two exceptions out of 15 cases in the size experiment (JAS, $\Delta s = 0.06$ and PS, $\Delta s = 0.04$) all hexagons are convex. Their convexity indicates that observers were less inclined to (incorrectly) say “different” when $p = s$ than when $p \neq s$, *whatever the* $p - m$ *difference*. Hence, they did not exclusively base their recognition judgment on the difference between p and m (in which case the hexagons would have had vertical sides). We can therefore reject our null hypothesis and conclude that recognition must have taken advantage of some

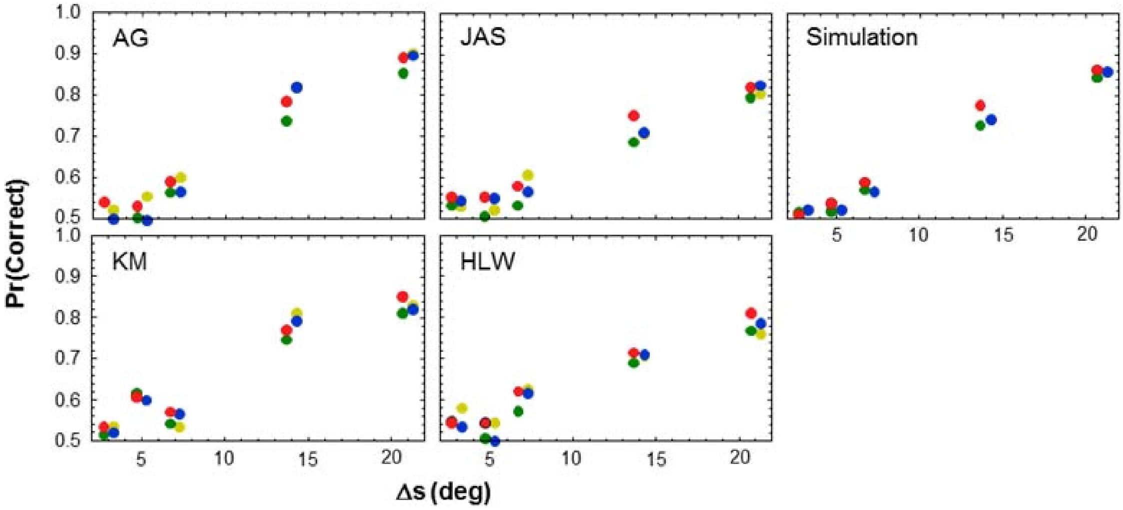

Figure 4. Measured (red symbols) and predicted orientation recognition accuracy as a function of task difficulty ($s - p = \Delta s$) for four observers and the variable-precision model (simulation; see section: The variable-precision, ideal-observer model). Predicted accuracies were obtained from three hypothetical observers subject to the same matching errors. One of these observers (green symbols) adopted criteria sampled at random from Normal distributions equal to the best fitting psychometric functions of $|p - m|$. Another observer (blue symbols) adopted the single most advantageous criterion with respect to $|p - m|$. The third hypothetical observer (dark yellow symbols) adopted the most advantageous criteria, not only with respect to $|p - m|$, but also with respect to the human observer's expected matching error (as determined from the regression analyses in Appendix A), given the effects of standard (s) and starting (m_0) orientations. Red and green symbols have been nudged leftward and blue and yellow symbols have been nudged rightward to facilitate legibility.

additional information, which was not used for reproduction.

To quantify the advantage of that additional information, we can compare each observer's overall performance in the recognition phase (red symbols in Figure 4) with the performance of a psychometrically matched observer not privy to any such information. The latter performance (green symbols) was computed using decision criteria that were sampled at random from the Normal distribution equal to the two-parameter psychometric function (i.e., $\mu_{P \neq S} = \mu_{P=S} = \mu$ and $\sigma_{P \neq S} = \sigma_{P=S} = \sigma$ in Equation 1) that best fit the human observer's data. Each green symbol in Figure 4 shows the psychometrically matched observer's overall performance after 1000 trials with *each* of the human observer's 200 matching errors. In 18 of 20 cases (the exceptions were KM, $\Delta s = 5°$ and HLW, $\Delta s = 3°$) our human observers' performances exceeded those of psychometrically matched model observers (i.e., significantly more than 50%, using a binomial test; $P < 0.001$). We must infer that the two-parameter psychometric functions did not capture all of the information used by our human observers in the recognition task. Not only were their decisions based on something besides $|p - m|$, that additional information enhanced overall performances. In the size experiment, human recognition performance (Figure 5, red symbols) was better than that of psychometrically matched observers (green symbols) in 13 of 15 cases (i.e., also significantly more than 50%; same test; the exceptions were PS, $\Delta s =$ 0.04 and $\Delta s = 0.08$). A description of the blue and yellow symbols appears below, under Regression-based Models.

The fact that our human observers out-performed psychometrically matched observers implies that the former used information besides $|p - m|$ when deciding "same" or "different." We will now demonstrate that the present recognition results are consistent with an observer whose criterion varies from trial-to-trial with the precision of each memory trace. Although we have no evidence that this criterion placement reflects a conscious strategy, we will use the term *confidence* to describe the underlying variable. When observers are confident that their still-visible match is good (i.e., close to the standard), they effectively label all but the most similar probes as "different." When observers have low confidence in their match, they show a greater willingness to accept some of those same probes as "same."

Of course, if observers' confidence in each match bore no relation to its actual accuracy, then we would not expect any advantage of a trial-by-trial decision criterion modulation in the recognition task. Therefore, as an initial investigation into the viability of our idea, we refit the aforementioned psychometric functions when recognition decisions were segregated into two equal-sized subsets: those following large matching errors and those following small matching errors. Fitting psychometric functions to each of these subsets separately (see Appendix C) confirmed that our

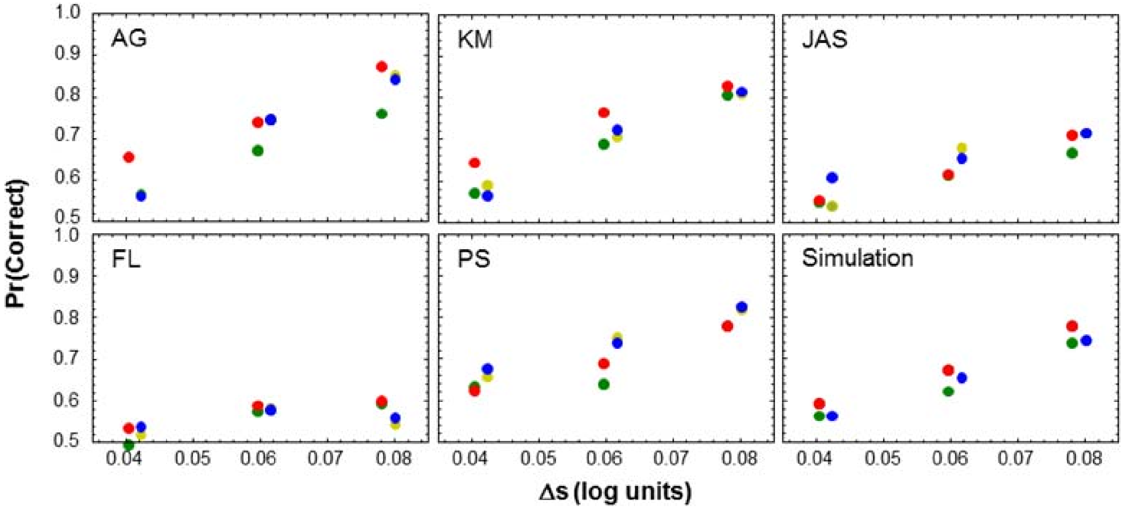

Figure 5. As in Figure 4, but for the size experiment and five observers.

observers adopted larger criterion values of $|p - m|$ when their matching errors were large. This is in line with studies reporting that confidence (at the time of the test) and accuracy are based (at least partly) on the same source of information (such as memory strength; Busey, Tunncliff, Loftus, & Loftus, 2000) and that subjects have (conscious or unconscious) access to the fidelity of their memory (the *trace access theory*; Hart, 1967; Burke, MacKay, Worthley, & Wade, 1991).

The variable-precision, ideal-observer model

To demonstrate that a greater reluctance to say "different" when memory fidelity is low translates into an advantage for recognition over the case where observers' same/different responses are independent of their confidence, we simulated the performance of an observer whose matching errors ($s - m$) were randomly selected from a an inverse Gamma distribution:[5]

Let $s - m \sim N(0, Y)$, where $Y \sim \text{Inv-Gamma}(a, b)$, $a > 1$, $b > 0$.

Essentially, this means that memory noise (or some correlated variable such as confidence) fluctuates from trial to trial. We selected an inverse Gamma distribution for Y as a matter of convenience. For one thing, all samples from it are positive, a requirement for variances. Furthermore, integrating over all possible values of Y yields a relatively familiar distribution for $s - m$: the non-standardized version of Student's t. When the inverse Gamma distribution is described by shape and scale parameters a and b, respectively:[6]

$$\text{var}(s - m) = \frac{b}{a - 1}, \tag{2}$$

which is guaranteed to be greater than zero. Although this formula contains two parameters, we really have only one free parameter, because $\text{var}(s - m)$ is something we measure:

$$b = (a - 1)\text{var}(s - m) \tag{3}$$

We can approximate ordinary Signal Detection Theory by adopting a large value for a. Fluctuations in memory noise will be largest when we adopt a small value for a.

We simulated the behavior of the ideal observer (see Appendix B), who adopts the most advantageous criterion on each trial, given that trial's sample of Y. The one free parameter in our model is a. As noted above, it describes the shape of the variance distribution. For the simulations illustrated in Figures 2-5, we selected $a = 2$. For Figure 2 and 4: $b = (10°)^2$. For Figure 3: $b = (0.04)^2$.

Just like our human observers, this variable-precision ideal observer selected larger criteria (on average) when $p = s$. Consequently, psychometric fits to its data form hexagons, when plotted in the format of Figures 2 and 3. As can be seen from the red symbols in Figures 4 and 5 (simulation panels), the model's overall performance is similar to that of our human observers. It also exceeds that of a psychometrically matched observer (green symbols) by an amount similar to that seen in our human observers' data.

Regression-based models

The main point of our paper is that decision strategies covary with uncertainty, which fluctuates over trials. A separate but interesting issue is the degree to which uncertainty fluctuations are due to external factors such as the oblique effect. Once that question has been answered, any remaining variability must be ascribed to internal factors such as arousal and attention (Shaw, 1980). We are not in a position to unequivocally measure the influence of external factors, nonetheless we believe that the regression models described in Appendix A represent a good first step in that direction. They provide estimates for the precision (the reciprocal of variable error) of each observer's memory given any values of s and m_0.

To see how well our observers could have done, given knowledge of these external effects on uncertainty, we simulated the recognition behavior of model observers whose memory noise was affected in the same way by s and m_0. These regression-based model observers adopted ideal criteria on the basis of each trial's combination s and m_0.

We have illustrated the performances of these regression-based models using dark yellow symbols in Figures 4 and 5. In 20 of 34 cases these performances were inferior compared to the human performances from which they were derived. (There was one tie.) This result suggests, in these 20 cases at least, that some uncertainty fluctuation should be ascribed to internal factors, and that our observers were able to adjust their criteria in concert with these fluctuations.

The question remains whether our human observers were able to exploit the systematic variations in their matching errors, which were revealed by our regression analyses, for the purposes of making better recognition responses. To address this question, we correlated their responses with those of the aforementioned regression-based models, as well as with the responses of another model observer, whose criteria were fixed with respect to $|p - m|$, but otherwise ideal, i.e., they optimized overall performance, given the sample variance in each condition's matching errors (i.e., $\text{Var}[s - m]$).

Any excess correlation between our observers' responses and those of the regression-based models (relative to the *ideal fixed-criterion* model) would indicate at least some criterion adjustment on the basis of external factors. As can be seen in Table 1, excess correlation was present in just three out of nine cases. In no cases did the two correlations differ by more than (0.03). Thus we have little evidence in favor of our observers exploiting the oblique effect or other external influences on the precisions of their memories when adjusting criteria for recognition. Therefore, we would like to suggest that the bulk of their uncertainty-based criterion fluctuations (Shaw, 1980) were due to internal factors (e.g., attention and arousal).

As neither the variable-precision ideal-observer's constant errors nor its variable errors could have been affected by either s or m_0, we felt a regression analysis of its data would be unnecessary. That is why there are no yellow symbols in the simulation panels of Figures 4 and 5, and that is why there are two empty cells in Table 1. On the other hand, we did feel it would be interesting to correlate the variable-precision model's recognition responses with those of the fixed-criterion otherwise-ideal observer. Those correlations (0.743 and 0.753) were quite a bit higher than any derived from our humans' data. This suggests that some criterion fluctuation is independent of uncertainty.

Experiment	Subject	Regression	Fixed criterion
Orientation	AG	0.641	0.656
	JAS	0.654	0.672
	KM	0.705	0.699
	HLW	0.654	0.65
	Simulation		0.743
Size	AG	0.535	0.526
	KM	0.655	0.665
	JAS	0.589	0.629
	FL	0.531	0.55
	PS	0.623	0.633
	Simulation		0.753

Table 1. Correlations between human recognition responses and those of two model observers derived from human matching data. Correlations between the variable-precision ideal-observer's recognition responses with those of the fixed-criterion ideal-observer are also provided.

Discussion

The present experimental paradigm, consisting in the successive measurement of reproduction and recognition performances, allowed the assessment of the trial-by-trial decisional behavior in recognition of two stored visual features, orientation and size. Our results show that recognition is better than can be expected from the prior explicit retrieval of the standard (through reproduction) under the usual assumption that subjects use a single decision criterion: the decision criterion for a *different* response when probe and standard were actually the same was more stringent (i.e., required a larger difference between the probe and the visible match) than when the probe and standard were not the same.

Our data support the view that observers do not maintain stable criteria for recognition. It should be understood that the observed criterion changes are *not* the consequence of the probe presentation but are modulated prior to it according to observers' confidence in their match. In its turn, the latter is related to the noisiness of the memory trace (or of the coding process) as reflected by the difference between match and standard. When the probe is identical to the standard (i.e., *signal* trials), absolute differences between standard (= probe) and match are a direct reflection of this noise (memory strength or coding efficiency, hence of the confidence) associated with that trial (Busey et al., 2000). Thus, large differences between probe and match will be associated with large criterion settings. When the probe differs from the standard, probe vs. match differences will be less correlated with the memory/coding noise and hence will not correlate with observers' criteria (see also footnote 5). As a consequence, both data and modeling show that observers uniformly demonstrate a greater willingness to accept probes as identical to the standard when they really are, regardless of their similarity to the match.

Simulations confirm that such criterion-setting strategy is consistent with a modulation of the recognition decision behavior in accord with observers' confidence in the fidelity of their prior reproduction of the standard. Recent studies demonstrate that such confidence can be coded on a trial-by-trial basis (e.g., Denève, 2012) by both parietal (Kiani & Shadlen, 2009) and midbrain dopamine neurons (de Lafuente & Romo, 2011). Further evidence favoring this model is the fact that observers adopt higher decision criteria when the differences between standard and match are large, i.e., for less accurate reproductions presumably associated with lower confidence levels. This finding supersedes the possibility that the observed advantage of human's recognition over an ideal observer using a unique decision criterion was due to the fact that our probe was presented at the same retinal location as the standard, while reproduction was performed at a different location (Dwight, Latrice, & Chris, 2008). Running the whole experiment with probe and match locations swapped would provide a direct test of this conjecture and is one of our future experiments.

Our modeling of the present results is based on two critical premises. The first premise is that the storage (or coding) of visual features is a doubly stochastic process. While this premise cannot be tested directly, it has recently received strong support from a study (van den Berg et al., 2012) having tested four memory models, of which the one positing a variable precision across trials provided the best fits to human performance in four sets of experiments. These experiments involved either estimating the color or orientation of one among N memorized items or localizing the change in a color or orientation among N locations. Neurophysiological studies support the notion of a variable noise (typically attributed to attentional fluctuations) and of its trial-by-trial impact on the decision behavior (Cohen & Maunsell, 2010; Churchland, Kiani, Chaudhuri, Wang, Pouget, & Shadlen, 2011; David, Hayden, Mazer, & Gallant, 2008; Nienborg & Cumming, 2009).

Our second premise is that, consciously or not, recognition strategies vary from trial-to-trial with memory precision. In standard SDT (Green & Swets, 1966; Macmillan & Creelman, 2005), observers set their decision criterion within a few trials and stick to it inasmuch as the internal and decision noise permit. Confidence is thought of as reflecting the distance between the current internal response and this criterion, the rationale underlying Receiver Operating Characteristics (ROC) functions. In the context of the present memory task, we propose the inverse scheme whereby, based on a trial-by-trial estimation of the memory trace strength (or noisiness), confidence is established first, and the criterion is set accordingly: the lower the confidence, the higher the criterion. This sequence implies a variable-precision memory trace with this precision (or some correlated variable) accessible on each trial (e.g., Denève, 2012; Hart, 1967; Burke et al., 1991; Koriat, 1993, 1995). Future experiments are needed to establish the empirical link between noise and confidence.

Our data and analyses do not allow an unequivocal distinction between internal and external effects on uncertainty. We constructed models for how matching errors (in particular, their variances) might depend on the external factors of starting error and standard value (i.e., the oblique effect). According to these models, the external influences on uncertainty were not large enough to fully account for the observed criterial fluctuations. Nonetheless, alternative models can be formulated. If such models can account for more of the variance in matching errors, then it is possible they might also account for more, if not all of the criterial fluctuation apparent in our data. However, the main point of our paper would remain valid, namely that recognition criteria co-vary with uncertainty, and uncertainty fluctuates over trials.[7]

In conclusion, the present study has revealed a decisional mechanism by means of which recognition improves over predictions based on reproduction performances. Our simulations confirm that an observer who relies on his confidence in his memory's fidelity to adjust his recognition strategy is more inclined to say "different" when a to-be remembered standard and probe actually *are* different. This was the behavior of our observers.

Acknowledgments

We thank Anna Ma-Wyatt and Ed Vul for helpful discussions. We are particularly indebted to our editor Al Ahumada and to one of our reviewers who developed at length the idea of the relationship between observers' confidence and their decision criteria. Authors AG and JAS were supported by Royal Society grant IE111227.

Commercial relationships: none.
Corresponding author: Andrei Gorea.
Email: andrei.gorea@parisdescartes.fr.

Address: Laboratoire Psychologie de la Perception, Université Paris Descartes & CNRS, Paris, France.

Footnotes

[1] Perfect matches (i.e., where $m = s$) were consequently impossible, but note that these step sizes were much smaller than the average matching error (i.e., $SD[s - m]$).

[2] In other words, $p \in \{s - \Delta s, s, s, + \Delta s\}$. However, due to a programming error in the size experiment, for observers AG and KM (not the others) $p \in \{s + \log(2\text{-}10^{\Delta S}) \ s, s, + \Delta s\}$. As can be seen from Figure 3, $s + \log(2\text{-}10^{\Delta S})$ is very similar to $s - \Delta s$.

[3] In this paper, all angles (such as $s - m$) are signed, acute, and analyzed arithmetically. For comparison, the average *SD* of our axial data (Fisher, 1993; pp. 31–37) was 2.6.

[4] The frequency of trials in which $\mathrm{sgn}(p - s) = \mathrm{sgn}(p - m)$ naturally decreases as Δs increases. Nonetheless, using the aforementioned chi-square test, in almost every case we confirmed that there would be no significant increase in the maximum likelihood of cumulative normal fits when one set of parameter values (μ and σ) was used for those trials and another set was allowed for the remaining $p \neq s$ trials. [The sole exception was JAS's size data with $\Delta s = 0.06$. In this condition, JAS responded "same" on all 13 trials for which $\mathrm{sgn}(p - s) = \mathrm{sgn}(p - m)$.] Consequently, it seems reasonable to use a single set of parameter values for all of the $p \neq s$ trials in each panel, and that is why each hexagon is regular.

[5] It may be easier to first consider an observer with a more extreme case of nonstationarity. This observer either perfectly *remembers* (R) or entirely *forgets* (F) the standard *S*. In the former case, his response will be "same" for $p = s$ trials and "different" otherwise. When this observer forgets, he will respond randomly whether $p = s$ or not. The probabilities of a "different" response when $p = s$ and $p \neq s$ are then given by:

$$\Pr(\text{“different”}\ p = s) = \Pr(p = s, \mathrm{F}) \times \Pr(\text{“different”}|\mathrm{F})$$

$$\mathrm{p}(\text{“different”}\ p \neq s) = \Pr(p \neq s, \mathrm{R}) + \Pr(p \neq s, \mathrm{F}) \times \Pr(\text{“different”}|\mathrm{F})$$

$$= \Pr(p \neq s, \mathrm{R}) + \Pr(p = s, \mathrm{F}) \times \Pr(\text{“different”}|\mathrm{F})$$

$$> \Pr(\text{“different”}\ p = s).$$

[6] In these equations we use var$X$ to denote the squared *SD* of *X*.

[7] It is worth pointing out that, ultimately, from both a conceptual and computational point of view internal and external noises are (or can be made) equivalent (Ahumada, 1987; Pelli, 1990).

[8] When modelling the performance of AG and KM in the size experiment, a slightly different decision rule was required, due to the fact that $|p - s|$ could assume one of three values (see footnote 2). In this case, the decision rule was: respond "same" if and only if $C_L < p - m < C_H$. Numerical methods were used to find the negative and positive criteria (C_L and C_H, respectively) that maximised Pr(Correct).

References

Ahumada, A. J., Jr. (1987). Putting the visual system noise back in the picture. *Journal of the Optical Society of America A, 4*(12), 2372–2378.

Appelle, S. (1972). Pattern and discrimination as a function of stimulus orientation: The oblique effect in man and animals. *Psychological Bulletin, 78,* 266–278.

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and old adults. *Journal of Verbal Learning & Behavior, 6*, 325–337.

Busey, T. A., Tunncliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26–48.

Churchland, A. K., Kiani, R., Chaudhuri, R., Wang, X.-J., Pouget, A., & Shadlen, M. N. (2011). Variance as a signature of neural computations during decision making. *Neuron, 69*(4), 818–831.

Cohen, M. R., & Maunsell, J. H. R. (2010). A neuronal population measure of attention predicts behavioral performance on individual trials. *Journal of Neuroscience*, *30*(45), 15241–15453.

David, S. V., Hayden, B. Y., Mazer, J. A., & Gallant, J. L. (2008). Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron*, *59*(3), 509–21.

de Lafuente, V., & Romo, R. (2011). Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. *Proceedings of the National Academy of Sciences of the United States of America, 108*(49), 19767–19771.

Denève, S. (2012). Making decisions with unknown sensory reliability. *Frontiers in Neuroscience, 6*(6), doi: 10.3389/fnins.2012.00075.

Dwight, J. K., & Latrice, D. V. & Chris, I. B. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences, 12*(3), 114–122.

Fechner, G. (1860 & 1966) *Elements of psychophysics (E. Adler, Trans.).* In D. H. Howe & E. G. Boring (Eds.). New York: Holt, Rinehart and Winston.

Fisher, N. I. (1993). *Statistical analysis of circular data.* Cambridge: Cambridge University Press.

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

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning & Verbal Behavior, 6*, 685–691.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*, 759–764.

Koriat, A. (1993). How do we know what we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*, 311–333.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)*. Mahwah: Lawrence Erlbaum.

Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature*, *459*(7243), 89–92.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics (3rd ed.)*. New York: McGraw-Hill.

Pelli, D. G. (1990). The quantum efficiency of vision. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 3–24). New York: Cambridge University Press.

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

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiency for the statistics of size discrimination. *Journal of Vision*, *11*(12):13, 1–11, http://www.journalofvision.org/content/11/12/13, doi:10.1167/11.12.13. [PubMed] [Article]

Tanner, T. A., Rauk, J. A., & Atkinson, R. C. (1970). Signal recognition as influenced by information feedback. *Journal of Mathematical Psychology, 7*, 259.

Tomassini, A., Morgan, M. J., & Solomon, J. A. (2010). Orientation uncertainty reduces perceived obliquity. *Vision Research, 50*(5), 541–547.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*(1), 68–111.

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

Vandenbussche, E., Vogels, R., & Orban, G. A. (1986). Human orientation discrimination: Changes with eccentricity in normal and amblyopic vision. *Investigative Ophthalmology & Visual Science, 27*(2): 237–245, http://www.iovs.org/content/27/2/237. [PubMed] [Article]

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis in absolute and comparative judgments. *Journal of Mathematical Psychology, 5*, 102–122.

Appendix A

We begin in this Appendix with the size experiment because the regression models are more straightforward. Orientation will be discussed below. Specifically we used standard linear regression to model constant error in the size experiment. Constant error can be thought of as an adjustment bias. It is the average matching error ($s - m$) for any given combination of standard size s and starting size m_0:

$$s - m = c_0 + c_1 s + c_2 m_0 + c_3 s m_0 + \varepsilon_c, \qquad \text{(A1)}$$

where c_0, c_1, c_2, and c_3 are arbitrary constants. The residual matching error (for which the other terms do not account) is contained in the term ε_c. Note this model contains the possibility of an interaction between the factors s and m_0. For each observer, Equation A1 was simultaneously fit to all matches. Analyses of variance (ANOVA) indicate significant effects ($P < 0.05$) of starting size in the data from all observers, significant effects of standard size in the data from three observers (KM, JAS, and FL), and significant interactions in the data from no observers.

To model the variable error, the squared residuals in Equation A1 were also subject to linear regression:

$$\varepsilon_c^2 = \upsilon_0 + \upsilon_1 s + \upsilon_2 |s - m_0| + \upsilon_3 s |s - m_0| + \varepsilon_v. \qquad \text{(A2)}$$

This model assumes that the squared variable error is linearly related to the starting error $|s - m_0|$, not to m_0. Consistent with Weber's Law, ANOVA failed to turn up a significant difference between v_1 and zero for any observer. The same thing occurred for the coefficient of interaction v_3. On the other hand, four of our five observers (JAS was the exception) had significant effects of starting error.

Figure A1 shows scatter plots of the size matching errors. The solid lines illustrate how each observer's constant error depends upon the standard size. Dashed curves show two variable errors (i.e., 2 *SD*s) about the constant error.

The model for constant error in the orientation experiment was previously used for similar purposes by Tomassini, Morgan, and Solomon (2010):

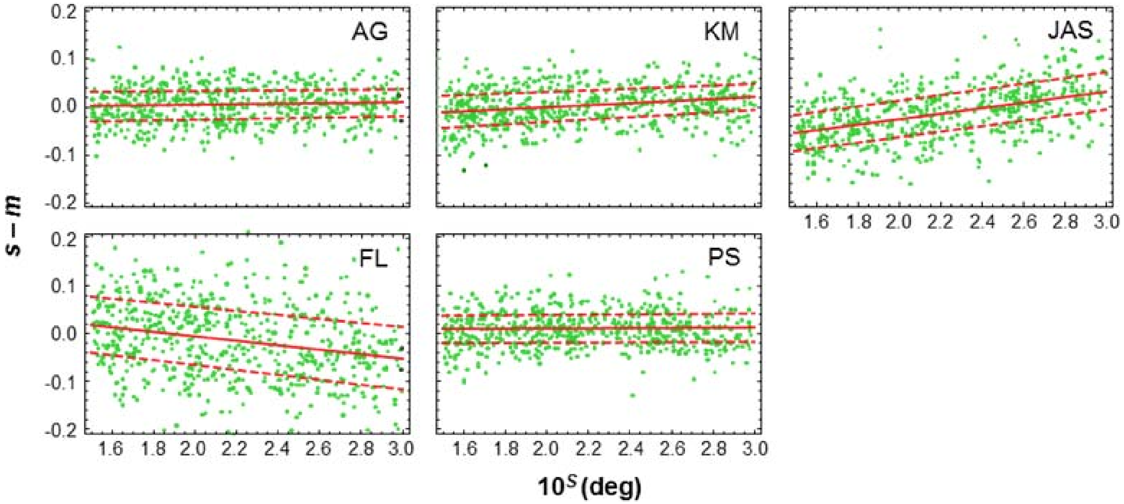

Figure A1. Scatter plots of the five observers' *size* errors ($s - m$) relative to the standard size s. Dashed curves contain two standard deviations about each observer's constant error (solid line).

$$s - m = c_1 \mathrm{sgn}(s)\left(\sin[4|s| - \sin - 1(c_2)] + c_2\right) + \varepsilon_c. \tag{A3}$$

In this expression, the parameter c_1 determines the overall error size and the parameter c_2 determines the (near intercardinal) orientations at which the tendency for clockwise errors equals that for anti-clockwise errors. To model the variable error, the residuals in Equation A3 were fit with the following model:

$$|\varepsilon_c| = v_0 + v_1\left(\sin\left[2|s| - \sin^{-1}(v_2)\right] + v_2\right) + v_3\left(\sin\left[2|s - m_0| - \sin^{-1}(v_4)\right] + v_4\right). \tag{A4}$$

Note that there really is no firm theory behind either of these equations. They are provided merely to produce curves that illustrate the effects of standard orientation and starting error. For example, in Equation A4, the right-hand side is the sum of two full-wave rectified sine functions, which has been elevated so that its minimum is greater than zero. The fancy bit with the arcsine allows each effect to reach its maximum at an arbitrary orientation without moving the local minima away from −90, 0, and 90°. Large values of v_1 correspond to large oblique effects. For observers AG, JAS, KM, and HLW, the best-fitting values for this parameter were 4°, 6°, 4°, and 1°, respectively. That is, some observers (especially JAS) exhibited stronger oblique effects than others (especially HLW).

Figure A2 contains scatter plots for orientation. As in Figure A1, here the dashed lines contain two standard deviations about the constant error. In this case, both the expected error and its standard deviation were modeled as (two-parameter) lines.

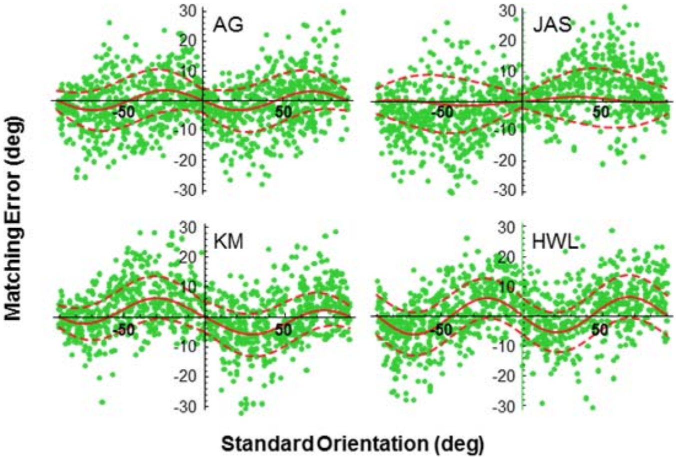

Figure A2. Scatter plots of the four observers' *orientation* errors ($s - m$) relative to the standard orientation s. Dashed curves contain two standard deviations about each observer's constant error (solid sinusoid).

Appendix B

In all our modeling, we assume that each matching error $s - m$ is drawn from a zero-mean Gaussian distribution having variance Y, i.e., $s - m \sim N(0, Y)$. Furthermore, we assume that observers respond "different" if and only if $|p - m| > C > 0$, where the C is known as the criterion.[8] In this Appendix we describe how to calculate the best possible criterion c_y for any value of variance y (i.e., regardless whether or not that variance itself is a random variable, as in the variable precision model).

On half the trials, in which probe and standard are identical, the probability density function of $p - m$ is

$f_{p=s}(p-m) = 1/\sqrt{y}\phi[(p-m)/\sqrt{y}]$, where ϕ is the Normal probability density function. On the other half of the trials, in which $p = s \pm \Delta s$, the density is $f_{p\neq s}.(p-m) = 1/2\sqrt{y}\{\phi[(p-m-\Delta s)/\sqrt{y}] + \phi[(p-m+\Delta s)/\sqrt{y}\}$.

An observer who adopts some arbitrary criterion c, will be correct with probability $\frac{1}{2}[\int_{-c}^{c} f_{p=s}(x)dx] + \frac{1}{2}[1 - \int_{-c}^{c} f_{p\neq s}(x)dx]$.

Analytical methods (i.e., Mathematica) were used to find that this function has its maximum at the value

$$c_y = \frac{y \ln\left[e^{\frac{(\Delta s)^2}{2y}} + \sqrt{-1 + e^{\frac{(\Delta s)^2}{y}}}\right]}{\Delta s} \qquad \text{(B1)}$$

Appendix C

Figures C1 (and C2) show cumulative Normal (Equation 1) fits to the responses of each observer in the orientation (and size) experiment when segregated according to whether the matching error $|s - m|$ was smaller or larger than the median matching error (solid and dashed curves, respectively). With two exceptions (AG and KM with the two smallest Δs values) out of the 20 cases (5 Δ × 4 Obs) for orientation and two exceptions (JAS medium Δs; PS small Δs), dashed curves are shifted to the right of the solid ones, indicating that observers adopt higher criteria for larger matching errors.

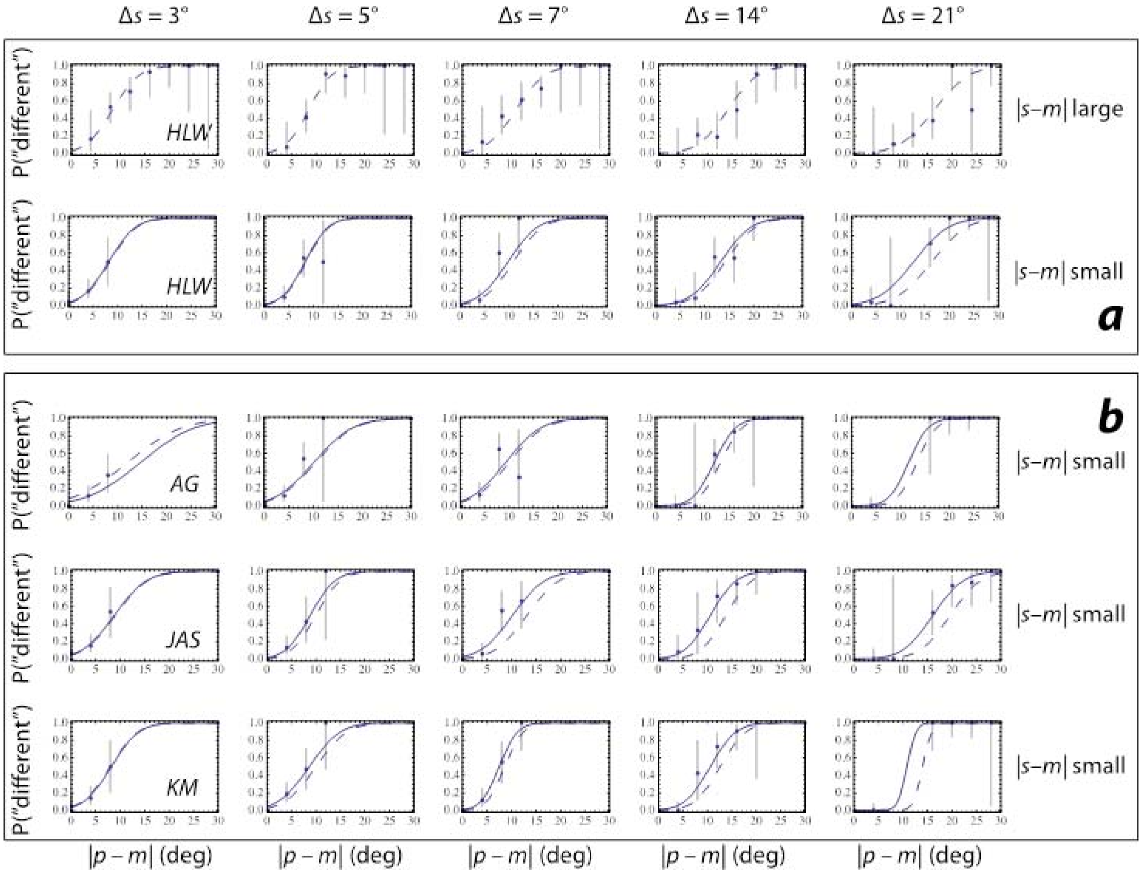

Figure C1. Cumulative Normal fits to the proportion of "different" responses vs. the $|p - m|$ difference. Each panel in (a) shows the maximum-likelihood fit (dashed curve) to half of observer HLW's trials at a given level of difficulty (Δs). These trials are those producing matching errors larger than the median $|s - m|$. For the purposes of illustration, the data have been pooled within 4° bins. Error bars contain 95% confidence intervals based on the binomial distribution. Each panel in the lower half of (a) replots the dashed curve from above, along with the cumulative Normal fit (solid curve) to the other half of HLW's data. Allowing these two fits to have different slopes (not shown) did not significantly increase their joint likelihood. In (b) the corresponding fits are shown for the remaining three observers, along with the binned data from trials with the most accurate matches (i.e., $|s - m|$ small).

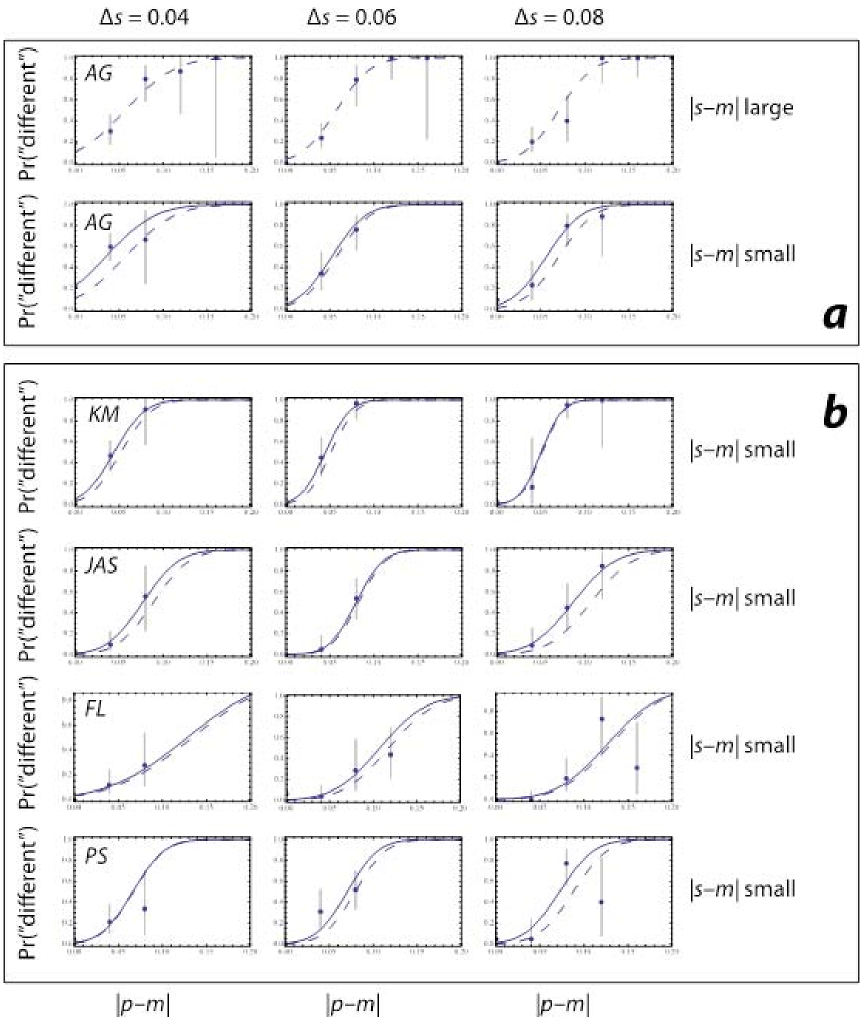

Figure C2. Cumulative Normal fits to the proportion of "different" responses vs. $|p - m|$. Each panel in (a) shows the maximum-likelihood fit (dashed curve) to half of observer AG's trials at a given level of difficulty ($\Delta \underline{s}$). For the purposes of illustration, the data have been pooled within 0.04 log unit bins. Everything else as in Figure C1.