Monkeys and humans take local uncertainty into account when localizing a change

Deepna Devkar
Department of Neurobiology & Anatomy, University of Texas Medical School, Houston, TX, USA

Anthony A. Wright
Department of Neurobiology & Anatomy, University of Texas Medical School, Houston, TX, USA

Wei Ji Ma
Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA
Present address: Center for Neural Science, New York University, New York, NY, USA

Since sensory measurements are noisy, an observer is rarely certain about the identity of a stimulus. In visual perception tasks, observers generally take their uncertainty about a stimulus into account when doing so helps task performance. Whether the same holds in visual working memory tasks is largely unknown. Ten human and two monkey subjects localized a single change in orientation between a sample display containing three ellipses and a test display containing two ellipses. To manipulate uncertainty, we varied the reliability of orientation information by making each ellipse more or less elongated (two levels); reliability was independent across the stimuli. In both species, a variable-precision encoding model equipped with an "uncertainty–indifferent" decision rule, which uses only the noisy memories, fitted the data poorly. In both species, a much better fit was provided by a model in which the observer also takes the levels of *reliability-driven* uncertainty associated with the memories into account. In particular, a measured change in a low-reliability stimulus was given lower weight than the same change in a high-reliability stimulus. We did not find strong evidence that observers took *reliability-independent* variations in uncertainty into account. Our results illustrate the importance of studying the decision stage in comparison tasks and provide further evidence for evolutionary continuity of working memory systems between monkeys and humans.

Introduction

Many paradigms used in the study of visual working memory (VWM), such as change detection (Phillips, 1974), require the observer to compare two displays (sample array and test array), separated by a delay interval. Textbook theories of VWM attribute errors in such "comparison paradigms" primarily, if not solely, to a maximum on the number of stimuli (items) that can be stored in VWM (Cowan, 2005; Pashler, 1988). By contrast, from the perspective of threshold psychophysics and signal detection theory, errors in these tasks are primarily "comparison errors," caused by noise in the encoding process (Palmer, 1990). Models adhering to the latter view account better for psychometric curves in change detection (Keshvari, van den Berg, & Ma, 2012; Keshvari, van den Berg, & Ma, 2013; Lara & Wallis, 2012), change discrimination (Bays & Husain, 2008; Lakha & Wright, 2004; Palmer, 1990), change localization (Devkar, Wright, & Ma, 2015; R. Van den Berg, Shin, Chou, George, & Ma, 2012), and VWM-based search (Mazyar, van den Berg, & Ma, 2012). Further support is provided by findings that accuracy is lower in within-category than in between-category change detection (Alvarez & Cavanagh, 2004), and that receiver operating characteristics in change detection resemble regular signal detection theory curves (Wilken & Ma, 2004). Advocates of the item-limit view have proposed variants of item-limit models that contain a noisy encoding stage (Zhang & Luck, 2008); although these models still fit poorly (van den Berg, Awh, & Ma, 2014; Van den Berg et al., 2012), at least they contribute to the consensus in the field that working memories are noisy.

When memories are noisy, the decision stage in a comparison task warrants careful study (Ma, Husain, & Bays, 2014). In this stage, the observer combines the noisy memories of the sample stimuli with the

Citation: Devkar, D., Wright, A. A., & Ma, W. J. (2017). Monkeys and humans take local uncertainty into account when localizing a change. *Journal of Vision*, *17*(11):4, 1–15, doi:10.1167/17.11.4.

doi: 10.1167/17.11.4 Received December 10, 2016; published September 6, 2017 ISSN 1534-7362

information about the test stimuli to reach a decision about the presence or location of a change. There is an optimal (accuracy-maximization) way to do so, namely maximum-a-posteriori estimation. For example, when determining which of the N stimuli changed, the optimal observer would compute from the N noisy memories and the N test stimuli the probability that any stimulus changed, and report the test stimulus with the highest probability. Human decisions in VWM-based comparison tasks are well described by an optimal decision rule acting on a variable-precision encoding stage (Devkar, Wright, & Ma, 2015; Keshvari et al., 2012; Mazyar, van den Berg, & Ma, 2012; van den Berg, Shin, Chou, George, & Ma, 2012).

However, there is more to the decision stage. An intermediate step in the optimal computation is the computation of likelihood functions over the identities of the sample stimuli. If the noisy memory of a sample stimulus is x, then the likelihood function over the hypothesized sample stimulus θ that produced x is $L(\theta) = p(x|\theta)$. The likelihood function represents degrees of belief in different hypothesized values of θ, and the width of the likelihood function represents uncertainty. If the observer is not only optimal but also Bayesian in a strong sense (engaging in what has been called "probabilistic computation" or "Bayesian transfer" (Ma, 2012; Ma & Jazayeri, 2014; Maloney & Mamassian, 2009), then they will incorporated the full likelihood function (and associated level of uncertainty) over each sample stimulus on each trial, and perform optimally even as uncertainty differs between sample stimuli and across trials. A main alternative to such an "uncertainty-incorporating" rule is an "uncertainty-indifferent" rule, which only uses the measured changes (the differences between the memories and the test stimulus) to make a comparison decision (Donkin, Nosofsky, Gold, & Shiffrin, 2013). To distinguish between uncertainty-indifferent and uncertainty-incorporating decision rules, it is imperative to manipulate the reliability of the stimulus information, as is often done in the study of cue combination (Knill & Pouget, 2004; Trommershauser, Kording, & Landy, 2011; Yuille & Bulthoff, 1996) and recently also in the study of visual search (Ma, Navalpakkam, Beck, van den Berg, & Pouget, 2011). Higher reliability means a narrower likelihood function and lower uncertainty, and the Bayes-optimal observer would give less weight to less reliable (more uncertain) evidence (Yuille & Bulthoff, 1996). In an orientation change detection task with variable reliabilities, the Bayes-optimal decision model with a variable-precision encoding model again best described human subjects' behavior (Keshvari et al., 2012), indicating that the brain treats working memories differently depending on their associated levels of uncertainty.

This finding raises the question how neural circuits incorporate uncertainty in the computation of comparing a memory with a test stimulus. In a single-stimulus change discrimination task requiring tactile working memory, the comparison computation has been studied both experimentally (Romo & Salinas, 2003) and theoretically (Machens, Romo, & Brody, 2005), but without consideration of the role of uncertainty. As a preparation for future physiological studies, we investigate here whether rhesus monkeys take into account varying reliability in a VWM-based task. It is known that monkeys take varying reliability into account in cue combination (Fetsch, Deangelis, & Angelaki, 2010; Gu, Angelaki, & DeAngelis, 2008), but that task does not require working memory. We previously showed that monkeys behave similarly to humans in two-alternative change localization (Devkar et al., 2015), but all stimuli had the same reliability. Here, we use the same task but vary reliability, allowing us to ask whether rhesus monkeys behave in accordance with the prediction of the Bayesian model. Moreover, we perform a direct cross-species comparison: We tested humans in the exact same task.

Experimental methods

Monkeys

Subjects

Two adult male rhesus monkeys (*Macaca mulatta*; weights: M1 = 16.5 kg and M2 = 13.5 kg; ages: M1 = 17.5 and M2 = 12.5 years) were tested in a change localization task for five days each week. Before daily testing, we restricted the monkeys' food and water intake. After completing the daily experimental sessions, animals were returned to their individual caging room and received a standard diet of primate chow and water. All animal procedures conformed to the National Institutes of Health guidelines, approved by the Institutional Review Board at University of Texas Health Science Center at Houston, and supervised by the Institutional Animal Care and Use Committee. The study adhered to the ARVO (Association for Research in Vision and Ophthalmology) Statement for the Use of Animals in Ophthalmic and Visual Research.

Apparatus

The monkeys were placed unrestrained in a custom-made aluminum experimental chamber (47.5 cm wide by 53.1 cm deep by 66.3 cm high) during training and testing. An infrared touchscreen detected touch responses to a 17-in. computer monitor. A Plexiglas template was used to guide touch responses. The

template had six cutouts (diameter of each circle cutout = 2.75 cm), matching the possible locations of the stimuli arranged on an imaginary circle of 9.0 cm diameter and a cutout in the center (diameter = 2.5 cm) for directing touches to a fixation point. Experimental sessions were designed, controlled, and recorded using a custom program written in Microsoft Visual Basic 6.0. A computer-controlled relay interface (Model P10-12; Metrabyte, Taunton, MA) was used to control food reinforcement (either a banana pellet or cherry Kool-Aid) for correct responses and to operate the illumination of the chamber with a 25 W green light bulb located outside of the chamber. The offset of the green light illuminating the chamber through a small gap between the touchscreen and the monitor cued the beginning of the next trial. The monkeys were monitored with a video camera that was focused through a small glass port on the right side of the chamber.

Stimuli

Stimuli consisted of gray ellipses with luminosity of 190 cd/m^2 displayed on a black background. Two types of ellipses, of equal area, were used: "high reliability" (HR; long and narrow) and "low reliability" (LR; short and wide). Based on the average distance of the monkey from the screen (approximately 35 cm), the HR and LR stimuli subtended visual angles of 2.9° × 0.65° and 1.5° × 1.3°, respectively. Stimuli were presented in six possible locations on the screen, arranged on an imaginary circle of radius 7.4°.

Trial procedure

Each trial began with a red fixation point in the center of the screen as shown in Figure 1. The monkeys had to make a one-touch response to the fixation point, which initiated the presentation of a sample display of three ellipses for 300 ms. The reliability of each ellipse in the sample display was independently chosen to be high or low. After a delay of 1000 ms, the test display was presented, which always consisted of two stimuli, placed at the same locations as two randomly chosen stimuli in the sample display. One test stimulus had the same orientation as the corresponding stimulus in the sample display, and the other test stimulus had a different orientation. Each test stimulus always had the same reliability as the corresponding sample stimulus.

Focusing on the two test stimuli, there were four reliability conditions: (1) Both test stimuli had high reliability. (2) The changed stimulus had high reliability and the unchanged stimulus had low reliability. (3) The changed stimulus had low reliability and the unchanged stimulus had high reliability. (4) Both test stimuli had low reliability. The monkeys' task was to identify the changed stimulus and touch it. The test display remained on the screen until the touch response was made. Correct responses were rewarded. An intertrial interval of 3000 ms followed the response, during which a green light illuminated the chamber and the screen was dark.

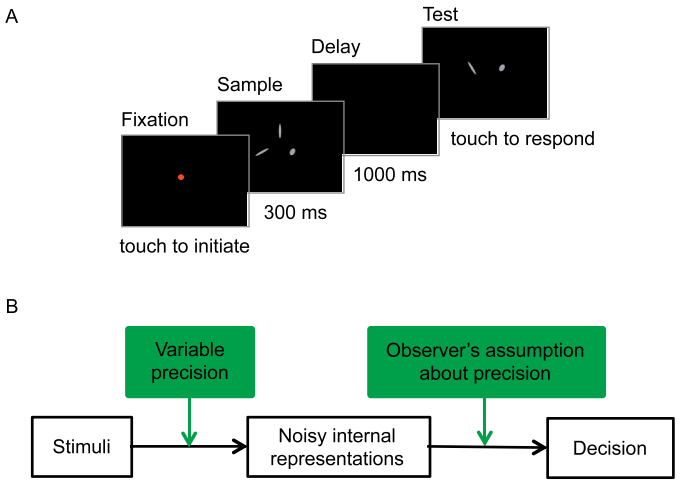

Figure 1. Experiment and model design. (A) Trial procedure for both humans and monkeys. Subjects were presented with a sample array containing three stimuli, followed by a delay, followed by a test array containing two stimuli. Subjects touched the stimulus in the test array that changed orientation from the sample array. Stimuli were ellipses of either high or low elongation; the elongation of any one stimulus was the same between sample and test array. (B) Model structure. We model the encoding stage using the variable-precision model. The three models differ in the decision stage, specifically, in the assumption the observer makes about encoding precision.

Training

Both monkeys had been previously trained in a change localization task with oriented bars (Devkar et al., 2015). To train them on this task, we first intermixed trials of oriented bars (old stimuli) with trials of oriented ellipses for initial task acquisition. Once the monkeys' average performance on ellipse trials was similar to their average performance with oriented bars, we began training them with only ellipses. Both monkeys were first trained with only high-reliability ellipses at a set size of 2 and a sample viewing time of 300 ms. Both monkeys reached criterion performance (overall accuracy of 70%) after six sessions. Then, we began training them with a set size of three and gradually intermixed the low-reliability trials. Once the monkeys' performance on these trials reached approximately 60%, they were ready for testing. For M1 and M2, the total training required 28 and 32 sessions, respectively.

Testing

The sample display was shown for 300 ms and set size was fixed at three. On every trial, each stimulus had an equal probability of being a high- or low-reliability ellipse. The locations of the ellipses were chosen randomly from six possibilities. The orientation of each sample stimulus, θ, was drawn independently from a uniform distribution over 18 possible orientations (–90° to 80° in increments of 10°). The orientation of the changed stimulus in the test display was drawn from the same uniform distribution. Testing consisted of 60 sessions, with 192-trial blocks per session, for a total of 11,520 trials per monkey.

Humans

Subjects

Ten human subjects (eight females) aged 21–35 years (mean age = 29.1 years) participated. Each subject was compensated $20 for two 1.5-hr sessions. Study procedures were approved by the University of Texas Health Science Center at Houston Institutional Review Board. The study conformed to the Declaration of Helsinki.

Apparatus and stimuli

Subjects were seated in a chair in a small room equipped with a computer monitor and touchscreen that were identical to those used for monkeys. The distance between the chair and the screen was adjusted before testing so that the stimuli and display would subtend approximately the same visual angles as for the monkeys. Subjects were asked to maintain approximately the same distance.

Trial procedure

The trial procedure was identical to that for the monkeys, except for the feedback. Subjects received trial-to-trial feedback, which consisted of a green light that was illuminated for 1 s and accompanied by a tone for correct responses, or a red light illuminated for 1 s for incorrect responses. At the beginning of each session, subjects were given instructions about the task and the feedback.

Training and testing

Each subject completed eight practice trials at the beginning of the first session for training. Each testing session consisted of three 192-trial blocks with a 10-min break time in between blocks. Each subject completed two such testing sessions for a total of 1,152 trials per subject.

Models

We model behavior in the change localization task using an encoding stage and a decision stage. The three models only differ in their decision stage.

Encoding stage

The encoding stage consists of the processes that produce the measurements of the sample and test stimuli. Since the test stimuli were available on the screen until response, we assumed that the measurements of these stimuli were noiseless. By contrast, the memories (measurements) of the sample stimuli are subject to noise.

Orientation convention

We mapped the orientation space to the interval $[0,2\pi]$ by multiplying all orientations and orientation change magnitudes by 2 before analysis. We consistently follow this convention in all equations as follow; however, for the figures only, we mapped change magnitudes back to actual orientation space.

Noisy memories

We assume that the memory x_i of the i^{th} sample orientation, θ_i, follows a von Mises distribution centered at θ_i with concentration parameter κ_i:

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

where κ_i is the concentration parameter that controls the width of the distribution and I_0 is the modified Bessel functions of the first kind of order 0. Following our previous work, we express encoding precision in terms of Fisher information, denoted by J_i (Keshvari et al., 2013; van den Berg et al., 2012). Fisher information measures the best possible performance of any unbiased decoder through the Cramér-Rao bound (Cover & Thomas, 1991). If x_i were normally distributed, Fisher information would be equal to the inverse of the variance of the Gaussian distribution. Moreover, Fisher information has a neural interpretation: in Poisson-like neural populations, it is proportional to the amplitude (gain) of activity in the population (W. J. Ma, 2010; Seung & Sompolinsky, 1993), which, in turn, can be thought of as the amount of neural resource devoted to the encoding of the memory (Bays, 2014; Ma et al., 2014; Van den Berg et al., 2012). For a von Mises distribution (Equation 1), Fisher information is directly related to the concentration parameter κ_i through

$$J_i = \kappa_i \frac{I_1(\kappa_i)}{I_0(\kappa_i)} \quad (2)$$

where I_1 is the modified Bessel function of the first kind of order 1.

We first distinguish between two sources of variability in precision: due to the random assignment of high and low reliability to the physical stimulus ("reliability-driven"), and due to internal fluctuations in precision for a given physical reliability ("reliability-independent").

Reliability-driven variability in precision

The experiment contained two levels of stimulus reliability: narrower and wider ellipses. As in a previous study (Keshvari et al., 2012), we allow mean precision $\bar{J}$ to differ between the high- and low-reliability stimuli. We denote mean precision for high-reliability stimuli by $\bar{J}_{\text{high}}$ and that for low-reliability stimuli by $\bar{J}_{\text{low}}$.

Reliability-independent variability in precision

Even for a given physical reliability, encoding precision might vary across stimuli and trials, as postulated by the variable-precision (VP) model (Fougnie, Suchow, & Alvarez, 2012; van den Berg et al., 2012). Such variability could be caused by a variety of possible factors, including stimulus-related differences (Bae, Olkkonen, Allred, Wilson, & Flombaum, 2014; Girshick, Landy, & Simoncelli, 2011; Pratte, Park, Rademaker, & Tong, 2017), configural effects (Brady & Tenenbaum, 2013), attentional fluctuations (Cohen & Maunsell, 2010; Goris, Simoncelli, & Movshon, 2012), and differential decay (Fougnie et al., 2012). In previous work, the variable-precision model has been successfully used to describe VWM limitations in monkeys (Devkar et al., 2015) and humans (Fougnie et al., 2012; Keshvari, van den Berg, & Ma, 2012; Keshvari et al., 2013; van den Berg & Ma, 2013; Van den Berg et al., 2012).

We implement reliability-independent variability in precision as follows. For each trial, we generate noisy measurements of the three sample stimuli, denoted by x_1, x_2, and x_3 through a doubly stochastic process, where each x_i is drawn from a von Mises distribution (a circular analog of a Gaussian distribution, used because orientation space is periodic) with a precision value that is itself randomly drawn from a gamma distribution with mean $\bar{J}$ and scale parameter τ, denoted by $p(J|\bar{J};\tau)$ with mean $\bar{J}$ and variance $\bar{J}\tau$.

Uncertainty

Given the noisy memories x_1 and x_2, the optimal observer would build beliefs about the underlying sample stimuli, θ_1 and θ_2, by "inverting" the generative model. Specifically, based on Equation 1, the likelihood function over θ_i is a von Mises distribution over θ_i centered at x_i and concentration parameter κ_i. The width of this likelihood function is a measure of the observer's uncertainty about θ_i: the higher encoding precision, the higher κ_i, and the lower the optimal observer's uncertainty. Although uncertainty is determined by the sources of variability in precision in the encoding stage, both reliability-driven and reliability-independent, we use the term "uncertainty" only to describe a property of the observer's beliefs about the stimuli, not to describe the encoding stage.

The three models we consider only differ in their decision stage. In the VP–VP model, the observer takes the correct likelihood functions over the sample orientations into account in their change localization decision. In the other two models, the observer behaves as if the likelihood functions they use in their decisions were based on incorrect assumptions about encoding precision. In the VP–FP (variable precision–fixed precision) model, the likelihood functions used are only based on only reliability-driven variations in encoding precision; reliability-independent variation in precision is ignored. In the VP–SP (variable precision–single precision) model, the likelihood functions used are based on the assumption that encoding precision in the same for both memories. In our notation, the "VP" before the hyphen refers to the common variable-precision encoding stage, and the part after the hyphen refers to the observer's assumption in the decision stage (variable precision, fixed precision, or single precision).

Decision rules

We now describe formally how in each model, the observer decides on the location of the change, L (1 or 2), given the noisy memories, x_1 and x_2, the associated concentration parameters, κ_1 and κ_2 (corresponding with uncertainty levels), and the two test orientations, φ_1 and φ_2. We will use the term "measured changes" for the differences $x_1-\varphi_1$ and $x_2-\varphi_2$.

VP–VP model

In the *Variable Precision-Variable Precision (optimal)* model, the observer takes both reliability-driven and reliability-independent variations in uncertainty into account. Specifically, the observer responds that the change occurred at location 1 if

$$\log I_0(\kappa_1) - \kappa_1 \cos(x_1 - \varphi_1) > \log I_0(\kappa_2) - \kappa_2 \cos(x_2 - \varphi_2) \quad (3)$$

(see Appendix A). Critically, this rule depends not only on the two measured changes, but also on the

concentration parameters κ_1 and κ_2, which are determined by the respective values of precision, J_1 and J_2, which are in turn determined both by the physical reliability and by random fluctuations in precision.

One can interpret the negative cosine of a measured change, $-\cos(x_i-\varphi_i)$, as a measure of dissimilarity between memory and test, and of the corresponding prefactors κ_1 and κ_2 as weights on dissimilarity, analogous to the $\frac{1}{\sigma^2}$ weights in cue combination (Trommershauser et al., 2011). Two special cases can be interpreted intuitively. One is in which the measured change is maximal at both locations, i.e. $x_1-\varphi_1=x_2-\varphi_2=\pi$. Then, the decision rule becomes $\log I_0(\kappa_1)+\kappa_1 > \log I_0(\kappa_2)+\kappa_2$, which simplifies to $\kappa_1>\kappa_2$, since the function $\log I_0(y)+y$ is a monotonically increasing function of y. In other words, the observer reports the location where uncertainty is lower; this makes sense, since at this location, the measured change is less likely to have been caused by noise. The other special case is in which the measured change is 0 at both locations, i.e. $x_1-\varphi_1=x_2-\varphi_2=0$. Then, the decision rule becomes $\log I_0(\kappa_1)-\kappa_1 > \log I_0(\kappa_2)-\kappa_2$, which simplifies to $\kappa_1<\kappa_2$, since the function $\log I_0(y)-y$ is a monotonically decreasing function of y. In other words, the observer now reports the location where uncertainty is *higher*; this makes sense, since at this location, it is more likely that the measured change of 0 was caused by a true change greater than 0.

VP–FP model

In the *Variable Precision-Fixed Precision* model, a suboptimal Bayesian model, the observer takes reliability-driven, but not reliability-independent variations in uncertainty into account: they behave as if κ_1 and κ_2 are completely determined by the physical reliabilities of the stimuli, and ignores any additional internal variability in encoding precision. This model is the same as the Variable Precision-Equal Precision (VEO) model in (Keshvari et al., 2012), but we found "fixed" a more intuitive description for "lack of variability across trials" than "equal." The observer thus uses only two levels of assumed precision: $\bar{J}_{\text{high}}$ for a high-reliability stimulus, and $\bar{J}_{\text{low}}$ for a low-reliability stimulus; these correspond to concentration parameters κ_{high} and κ_{low}. The decision rule, then, is identical to Equation 3 but with κ_1 and κ_2 each taking on one of only two possible values, κ_{high} and κ_{low}, depending on the reliability of that stimulus.

VP-SP model

In the *Variable Precision-Single Precision* model, another suboptimal Bayesian model, the observer completely disregards variations in uncertainty and behaves as if $\kappa_1=\kappa_2$ on every trial. Then, the observer reports location 1 when

$$\cos(x_2-\varphi_2) > \cos(x_1-\varphi_1), \quad (4)$$

or equivalently, when the measured change at location 1 is greater than at location 2. This model can be thought of as a "naïve signal detection theory model:" in signal detection theory, decision rules often only depend on point estimates (Macmillan & Creelman, 2005). The VP–SP model is distinct from both other models in that a measured change in a low-reliability stimulus is treated the same as a measured change of the same magnitude in a high-reliability stimulus.

Model predictions

In each model, we can compute the probability of a correct response in each stimulus condition, given a set of model parameters. Under the assumptions in the generative model, the stimulus condition is uniquely determined by change magnitude Δ and the reliability condition. We proceed as follows:

1. Without loss of generality, we assume $\theta 1=\theta 2=0$ and $\mathrm{L}=1$ (the changing stimulus is always the first one), so that $\varphi 1=\Delta$ and $\varphi 2=0$.
2. We drew 10,000 random values of J1 and J2 from a gamma distribution with scale parameter τ. Depending on the reliability condition, the means of the gamma distributions are $(\bar{J}_{\text{high}}, \bar{J}_{\text{high}})$, $(\bar{J}_{\text{high}}, \bar{J}_{\text{low}})$, $(\bar{J}_{\text{low}}, \bar{J}_{\text{high}})$, or $(\bar{J}_{\text{low}}, \bar{J}_{\text{low}})$.
3. For each combination of J_1 and J_2, we computed the corresponding κ_1 and κ_2 through Equation 2, then drew of x_1 and x_2 from a von Mises distribution with mean 0 and those concentration parameters.
4. We evaluated the decision rule for each of the 10,000 draws, and then computed the proportion of correct responses across all draws. This is our estimate of the probability correct according to the model for a given change magnitude and reliability condition.

Thus, the three models only differ in Step 4, where the decision rule is incorporated. Importantly, stimulus reliability affects the encoding stage of all models in the same way (through Step 2); in particular, all models predict performance difference among the four reliability conditions. However, in the VP–FP and VP–VP models, stimulus reliability also plays a role in the decision stage (Step 4).

In each model, we also included a lapse rate parameter, which accounts for errors due to lapses in attention, blinking or eye movements during stimulus presentation, or motor errors when making a response.

Each model has four free parameters: $\bar{J}_{\text{high}}$, $\bar{J}_{\text{low}}$, τ, and lapse rate. For each model, we finely discretized the

Model		IC* (model)- IC* (VP–FP)		LML(model)-LML(VP–FP)	
		Mean	*SEM*	Mean	*SEM*
VP-SP	M1	−102.6		−102.5	
	M1 − bootstrapped	−107	14	−107	14
	M2	−195.1		−201.6	
	M2 − bootstrapped	−199	23	−204	24
	Humans	−57.5	5.9	−39	4.5
VP−VP	M1	−12.6		−15.8	
	M1 – bootstrapped	−16	11	−19	11
	M2	−100.7		−106.9	
	M2 − bootstrapped	−102	17	−107	17
	Humans	−12.1	2.4	−2.4	1.4

Table 1. Model comparison. IC* stands for any of the information criteria AIC (Akaike information criterion), AICc (corrected AIC), or BIC (Bayesian information criterion), divided by −2 to make them comparable to log likelihoods; these metrics produce identical results, because all models have the same number of parameters. LML = log marginal likelihood. Values are for the VP–SP and VP–VP models relative to the VP–FP model. Negative values indicate worse fits. The VP–FP model outperforms both other models according to all metrics.

parameter space (see Table 1) and calculated a look-up table for the predicted probability of a correct response for each condition (defined by change magnitude Δ and reliability condition c) and for each parameter combination.

Model fitting

To fit model parameters, we used maximum-likelihood estimation. For a given model, we denote the model parameters collectively by a vector $\mathbf{t}$. The likelihood of $\mathbf{t}$ is the probability of the trial-to-trial subject responses given the trial-to-trial stimuli and $\mathbf{t}$. In our task, each response is uniquely defined by the correctness of the response. Moreover, the model's prediction for the probability of a correct response only depends on reliability condition c and change magnitude Δ. Finally, we assume that all trials are independent. Then, the log likelihood of $\mathbf{t}$ is

$$\begin{aligned}\mathrm{LL}(\mathbf{t}) &= \log p(\text{data}|\text{model}, \mathbf{t}) \\ &= \log \prod_{i=1}^{n_{\text{trials}}} p(\text{correctness}_i|c_i, \Delta_i, \mathbf{t})\end{aligned}$$

where the product is over trials (from 1 to n_{trials}) and correctness_i is 1 if the subject was correct on the i^{th} trial and 0 if not. We can rewrite this as

$$\begin{aligned}\mathrm{LL}(\mathbf{t}) &= \sum_{i=1}^{n_{\text{trials}}} p(\text{correctness}_i|c_i, \Delta_i, \mathbf{t}) \\ &= \sum_c \sum_\Delta n(c, \Delta, \text{correct}) \log p(\text{correct}|c, \Delta, \mathbf{t}) \\ &\quad + \sum_c \sum_\Delta n(c, \Delta, \text{incorrect}) \log p(\text{incorrect}|c, \Delta, \mathbf{t}),\end{aligned}$$

where trials are grouped by reliability condition c, change magnitude Δ, and by whether the observer was correct or incorrect, and n (c, Δ, correct) is the number of trials with a particular c, Δ, and correctness.

For each subject's data set, we used Equation 5 and the precomputed look-up table of model predictions mentioned already to find the log likelihood of each parameter combination. The parameter combination on this grid that maximized the log likelihood gave the parameter estimates. The model predictions corresponding to that parameter combination were then used to compute the model fits to the psychometric curves.

Model comparison

We compared models using the Akaike/Bayesian Information Criterion (Akaike, 1974; Schwartz, 1978), which are equivalent because all models have the same number of parameters. We also compare models using log marginal likelihood (LML; MacKay, 2003), as estimated with the same parameter grid as used for fitting. To make the information criteria comparable with the log marginal likelihood, we divide them by −0.5; we denote the resulting quantity by "IC*".[1]

Bootstrapping

The original data set for each of the two monkeys consisted of 11,520 trials. A random sample of 11,520 trials (a combination of condition, change magnitude, and correctness) was selected with replacement from the original dataset to create each of the 100 bootstrapped data sets. The parameter estimates, psychometric curves, and model comparisons were generated for each bootstrapped data set separately.

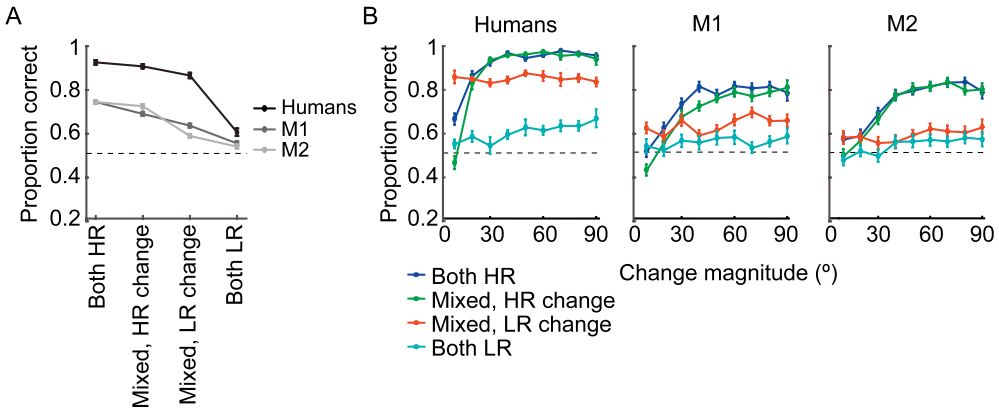

Figure 2. Psychometric curves. (A) Proportion correct as a function of reliability condition. Error bars are *SEM* for humans, and *SD* across 100 bootstrapped data sets for each monkey. (B) Proportion correct as a function of change magnitude and reliability condition (HR: high reliability; LR: low reliability), for humans, Monkey 1, and Monkey 2.

The means for each of these were computed by averaging across all bootstrapped data sets from the same monkey, and the standard deviations served as estimates of the standard errors of the means.

Results

Data

Overall percentage correct was 82.5 ± 1.0 for humans (mean ± *SEM*), 65.4 ± 0.5 for Monkey 1, and 64.7 ± 0.4 for Monkey 2 (both mean ± *SD* across bootstrapped data sets, which is an estimate of the *SEM*; nonbootstrapped: 65.4 for Monkey 1 and 64.6 for Monkey 2). A logistic regression of proportion correct against species (with the data from all individuals aggregated) and condition (arranged as both HR, mixed reliability with the HR item changing, mixed reliability with the LR item changing, and both LR) shows significant effects of species ($\beta = 0.175 \pm 0.005$; $p<10^{-10}$) and of condition ($\beta = -0.0788 \pm 0.0021$; $p<10^{-10}$) (Figure 2A). Yet, proportion correct was above chance even in the both-LR condition (binomial tests: aggregated human data: $p < 10^{-10}$; Monkey 1: $p = 6.8 \times 10^{-9}$; Monkey 2: $p = 6.8 \times 10^{-5}$). Thus, the low-reliability stimulus was not completely ignored.

A more detailed representation of the data is provided by proportion correct as a function of change magnitude for each of the four conditions (Figure 2B). We observe an interaction between reliability condition and change magnitude: when a high-reliability stimulus changes, change magnitude has a large effect, whereas when a low-reliability stimulus changes, change magnitude has a much weaker effect. This seems inconsistent with the premise of the VP–SP model that the observer treats a change the same whether it occurs in a low-reliability or in a high-reliability stimulus; however, detailed model comparison is needed.

Models

We fitted all models using maximum-likelihood estimation on a parameter grid; we individually fitted the data from each human subject, each monkey subject, as well as each data set bootstrapped from a monkey's data (see Model fitting). Parameter estimates in each model are given in Table B1.

According to both model comparison metrics and for both species, the "uncertainty–indifferent model", VP-SP, fits worst (Table 1). For example, IC* values of VP–SP are lower than those of VP-FP by 57.5 ± 5.9 for humans (with identical signs for each human), 102.6 for M1 (bootstrapped data: 107 ± 14), and 195.1 for M2 (bootstrapped data: 199 ± 23). We conclude that both monkeys and humans take local uncertainty associated with the memories into account when localizing a change.

We now compare the two uncertainty-incorporating models, VP–FP and VP–VP. Applied to human data, the VP–FP model is indistinguishable from the VP–VP model according to LML (better by 2.4 ± 1.4, inconsistent signs across subjects), but better according to IC* (by 12.1 ± 2.4). For M1, the models are indistinguishable according to either metric (see Table 1), while for M2, VP–FP fits much better according to both metrics (IC difference: 102 ± 17; LML difference: 107 ± 17). Overall, there is no strong evidence that either species consistently takes reliability-independent variations in uncertainty into account in the change localization decision; the balance of evidence points to only the main effect of stimulus reliability being taken into account.

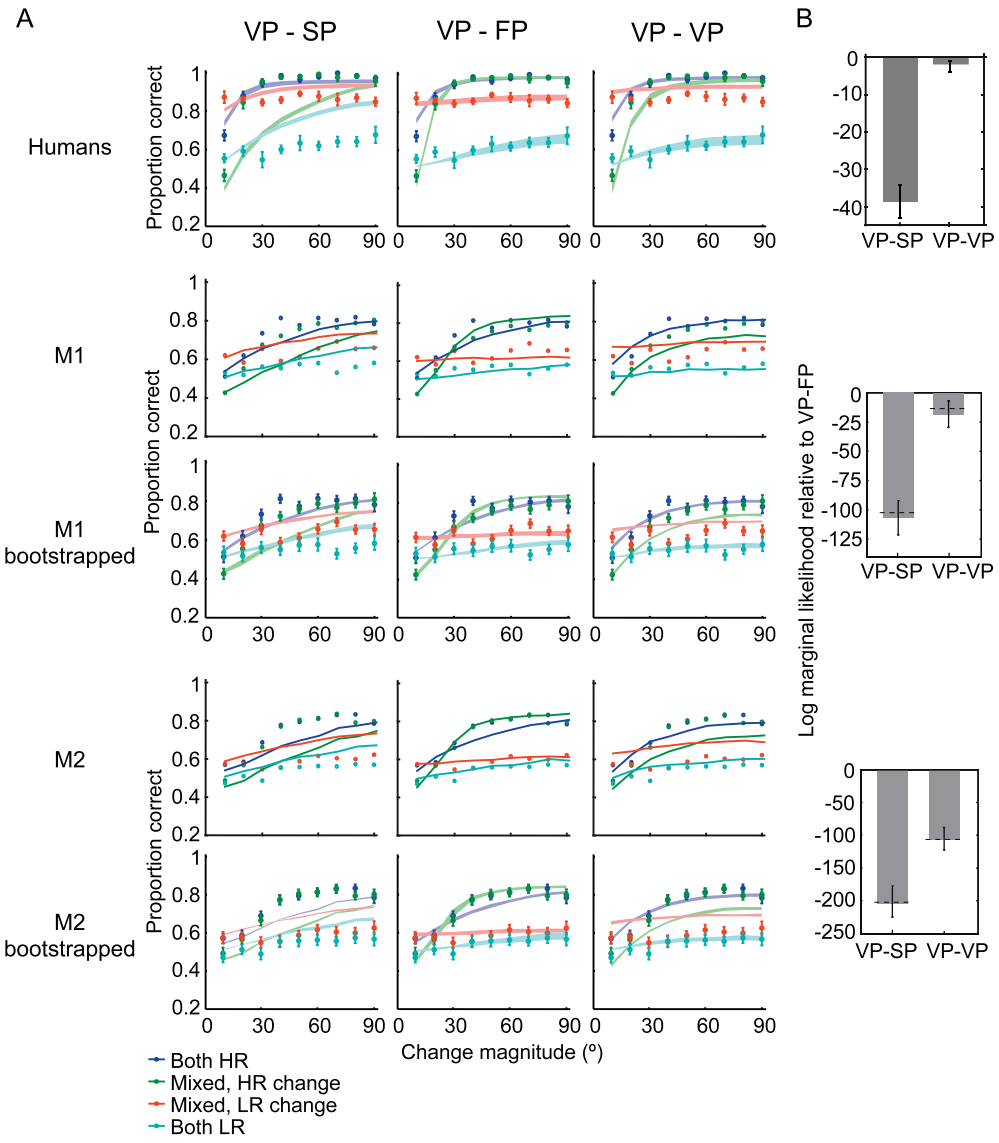

Figure 3. Model fits. (A) Fits of the three models (columns) to the psychometric curves for humans (top row), Monkey 1 (second row), Monkey 2's bootstrapped data (third row), Monkey 2 (fourth row), and Monkey 2's bootstrapped data (bottom row). Shaded areas: mean and *SEM* of model fit. VP–SP fits worst overall. The differences between VP–FP and VP–VP are subtle for humans and Monkey 1, but VP–FP fits better for Monkey 2. (B) Log marginal likelihoods of the VP–SP and VP–VP models relative to the VP–FP model. AIC/BIC results are consistent (see Table 1).

Model fits to the psychometric curves qualitatively confirm the model ranking (Figure 3). In earlier papers we also used R^2 computed from the fit to the psychometric curves for model comparison, but R^2 is an unprincipled measure for binary data (it is different from the log likelihood, which we optimize) and should not be used.

A *best*-fitting model does not necessarily fit the data *well*, but here it does. The VP–FP model, and to a lesser extent the VP–VP model, captures the features of the psychometric curves both qualitatively and quantitatively. In particular, the VP–FP and VP–VP models both correctly predict that in the mixed-reliability condition where the low-reliability stimulus changes, the magnitude of the change does not affect proportion correct. This makes sense: If the observer is confident that the high-reliability stimulus did not change, then the other stimulus must have changed; this type of decision by elimination does not take the magnitude of the change into account.

Discussion

In an orientation change localization task performed both by humans and rhesus monkeys, we varied reliability (ellipse elongation) between stimuli and

across trials. A model with a variable-precision encoding stage and a decision stage in which the observer takes reliability-driven variations of uncertainty into account (VP–FP) provided the best fits. We ruled out a "naïve signal detection theory" model (VP–SP) in which the observer only uses the measured orientation changes but does not take uncertainty into account. A qualitative feature of both the data and the VP–FP model is that when a change happens at a low-reliability location, the magnitude of the change has little effect on accuracy. The similarities between the species provide further support for evolutionary continuity of visual working memory systems (Devkar et al., 2015). We now point out some caveats and limitations of our work.

Encoding model

Our conclusions are conditioned on the variable-precision model being the correct encoding model. From previous work, evidence for this assumption is strong (van den Berg et al., 2014). In that study, we also found that a cap on the number of remembered stimuli cannot be ruled out in a variable-precision model, but the value of this cap was estimated at 5.66 ± 0.16. Therefore, if a cap exists, it would be unlikely to affect the current study, in which set size was 3.

Working memory representation of uncertainty

The reliabilities of the two test stimuli were always identical to the reliabilities of the two corresponding sample stimuli. Thus, there was no need for the subject to remember the uncertainty associated with the memories: they could simply get this information from the test array. Therefore, the present study cannot be used to answer the question of whether uncertainty is accurately stored in memory on a trial-to-trial basis for later use in decision-making. Evidence for such storage is provided by the finding that confidence ratings in a delayed-estimation task are correlated with error (Rademaker, Tredway, & Tong, 2012), modeled in (van den Berg, Yoo, & Ma, 2017). To address the same question within the current paradigm, one could give both test stimuli maximal reliability, for example, by using oriented line segments instead of ellipses. In this way, the test array would contain no information about the uncertainty of the memories of the sample array. In order to use a probabilistic decision rule (as in the VP–VP and VP–FP models), the observer would then have to remember the uncertainty values from the first array.

Suboptimality

Even though we showed that humans and monkeys take into account uncertainty, our results also suggest that monkeys use only two possible values of uncertainty, which are determined by the two possible reliabilities of the stimulus and not by additional variability. This finding is somewhat surprising given that in a very similar change detection task in humans (Keshvari et al., 2012), the VP–VP model (there called VVO) clearly won over the VP–FP model (there called VEO). We can think of two possible causes of this difference. First, in the Keshvari study, the test array was presented for 100 ms, whereas in the present study, it remained on the screen until response. This means that the observer had perfect knowledge of the elongations of the test ellipses, and therefore also of the elongations in the sample array. This could have encouraged subjects to use physical reliability in the decision rule. In the Keshvari study, this strategy (which, though suboptimal, might be easier) was not available to the subject. Second (and perhaps interacting with the first point), subjects received feedback in the present study, and no feedback in the Keshvari study. Feedback would allow subjects to learn a mapping from memories and reliability levels to decision. If reliability-independent variations in uncertainty are not part of this mapping, the VP–VP model would not be an accurate description of the learned mapping. The Keshvari study was different in other ways as well: The sample array was presented for 100 rather than 300 ms, set size was 4 rather than 3, and the task was *N*-stimulus change detection rather than two-stimulus change localization. However, it is not clear to us how these differences could have caused the VP–VP model to win.

Conclusion

Change localization has previously primarily been used to understand the encoding limitations of VWM processing (Buschman & Miller, 2009; Heyselaar, Johnston, & Pare, 2011; van den Berg et al., 2012). Here, we instead use the paradigm to probe decision-making strategies. Our findings provide cross-species evidence that while humans and monkeys seem to take uncertainty into account, they do not seem to do so optimally. In addition, the fact that this qualitative conclusion is the same in both species makes the case that rhesus monkeys are a good model system for the studying the role of uncertainty in working memory-based decisions.

Keywords: Bayesian inference, change detection, monkey, uncertainty, working memory

Acknowledgments

Commercial relationships: none.
Corresponding author: Wei Ji Ma.
Email: weijima@nyu.edu.
Address: Center for Neural Science, New York University, New York, NY, USA.

Footnote

[1] Historically, the factor of –2 was introduced into the information criteria so that in a Gaussian model, they would be interpretable as corrected sums of squared errors. However, log likelihoods are much more general than squared errors.

References

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology*, *8*(11), e1002771.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106–111.

Bae, G.–Y., Olkkonen, M., Allred, S. R., Wilson, C., & Flombaum, J. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision*, *14*(4):7, 1–23, doi:10.1167/14.4.7. [PubMed] [Article]

Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, *34*(10), 3632–3645.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher-order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109.

Buschman, T. J., & Miller, E. K. (2009). Serial, covert shifts of attention during visual search are reflected by the frontal eye fields and correlated with population oscillations. *Neuron*, *63*, 386–396.

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), 15241–15253.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: John Wiley & Sons.

Cowan, N. (2005). *Working memory capacity*. New York: Psychology Press.

Devkar, D., Wright, A. A., & Ma, W. J. (2015). The same type of visual working memory limitations in humans and monkeys. *Joournal of Vision*, *15*(16): 13, 1–18, doi:10.1167/15.16.13. [PubMed] [Article]

Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slots models of visual working-memory response times. Psychological Review, *120*(4), 873–902.

Fetsch, C. R., Deangelis, G. C., & Angelaki, D. E. (2010). *Neural corrrelates of dynamic sensory cue reweighting in macaque area MSTd*. Paper presented at Computational and Systems Neuroscience.

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

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

Goris, R. L. T., Simoncelli, E. P., & Movshon, J. A. (2012). *Using a doubly-stochastic model to analyze neuronal activity in the visual cortex*. Paper presented at the Cosyne Abstracts, Salt Lake City.

Gu, Y., Angelaki, D. E., & DeAngelis, G. C. (2008). Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience*, *11*(10), 1201–1210.

Heyselaar, E., Johnston, K., & Pare, M. (2011). A change detection approach to study visual working memory of the macaque monkey. *Journal of Vision*, *11*(3):11, 1–10, doi:10.1167/11.3.11. [PubMed] [Article]

Keshvari, S., van den Berg, R., & Ma, W. J. (2012). Probabilistic computation in human perception under variability in encoding precision. *PLoS ONE*, *7*(6), e40216. doi:10.1371/journal.pone.0040216

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology*, *9*(2), e1002927.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience*, *27*(12), 712–719.

Lakha, L., & Wright, M. J. (2004). Capacity limitations of visual memory in two-interval comparison of Gabor arrays. *Vision Research*, *44*(14), 1707–1716.

Lara, A. H., & Wallis, J. D. (2012). Capacity and precision in an animal model of short-term memory. *Journal of Vision*, *12*(3):13, 1–12, doi:10.1167/12.3.13. [PubMed] [Article]

Ma, W. J. (2010). Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research*, *50*, 2308–2319.

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in Cognitive Sciences*, *16*(10), 511–518.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*, 347–356.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*, 205–220.

Ma, W. J., Navalpakkam, V., Beck, J. M., van den Berg, R., & Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. *Nature Neuroscience*, *14*, 783–790, doi:10.1038/nn.2814.

Machens, C. K., Romo, R., & Brody, C. D. (2005). Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science*, *307*(5712), 1121–1124.

MacKay, D. (2003). *Information theory, inference and learning algorithms*. Cambridge, UK: Cambridge University Press.

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

Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*(1), 147–155.

Mazyar, H., van den Berg, R., & Ma, W. J. (2012). Does precision decrease with set size? *Journal of Vision*, *12*(6):10, 1–16, doi:10.1167/12.6.10. [PubMed] [Article]

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

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*(4), 369–378.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, *16*(2), 283–290.

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception & Performance*, *43*(1), 6–17.

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13):21, 1–13, doi:10.1167/12.13.21. [PubMed] [Article]

Romo, R., & Salinas, E. (2003). Flutter discrimination: neural codes, perception, memory and decision making. *Nature Reviews Neuroscience*, *4*(3), 203–218.

Schwartz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Seung, H., & Sompolinsky, H. (1993). Simple model for reading neuronal population codes. *Proceedings of the National Academy of Sciences, USA*, *90*, 10749–10753.

Trommershauser, J., Kording, K., & Landy, M. S. (Eds.). (2011). *Sensory cue integration*. New York: Oxford University Press.

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

van den Berg, R., & Ma, W. (2013). Plateau-related summary statistics are uninformative for comparing working memory models. *Attention, Perception, & Psychophysics*. *76*(7), 2117–2135.

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, USA*, *109*(22), 8780–8785.

van den Berg, R., Yoo, A., & Ma, W. J. (2017). Fechner's Law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review*, *124*(2), 197–214.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12):11, 1120–1135, doi:10.1167/4.12.11. [PubMed] [Article]

Yuille, A. L., & Bulthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference*. (pp. 123–161). New York: Cambridge University Press.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

Appendix A: Derivation of decision rule

Step 1: Generative model

Figure A1 shows the relevant variables: the location of the change, L (1 or 2), the magnitude of the change, Δ, the relevant sample orientations, θ_1 and θ_2 (all other sample stimuli are irrelevant to the decision), their noisy memories, x_1, and x_2, and the two test orientations, φ_1 and φ_2. Each variable has an associated probability distribution.

- Since both test locations are equally likely to contain the change, we have $p(L) = 0.5$.
- In the experiment, both sample orientations (θ_1 and θ_2) follow a discrete uniform distribution with 18 possible values. The subject may or may not have learned these values (see next point). However, the computation as follows will hold also if the observer assumes a different discrete uniform distribution or a continuous uniform distribution. Therefore, we simply write $p(\theta_1, \theta_2) = \frac{1}{k}$, with k a constant.
- In the experiment, change magnitude Δ also follows a discrete uniform distribution with 18 possible values, but we approximate it by a continuous uniform distributions, $p(\Delta) = \frac{1}{2\pi}$. There are three reasons for this choice:
  - We consider it unlikely that an observer learns those exact 18 change magnitudes. Albeit in a different domain, a study that used a discrete stimulus distribution consisting of six values showed that subjects did not learn those values. Specifically, a single Gaussian or a mixture of two Gaussians accounted better for the data than a mixture of six Gaussians (with free standard deviations, allowing for the value 0) centered on the true stimuli (Acerbi, Wolpert, & Vijayakumar, 2012).
  - The "true ideal-observer" model in which the subject does learn the 18 change magnitudes makes very similar trial-to-trial predictions. Specifically, when we use the parameters estimated from the data of any individual human or monkey subject, and simulate 10,000 simulated pairs of x_1 and x_2 per subject, reliability, and change magnitude, the trial-to-trial agreement between the decisions made by the exact ideal observer and our approximated ideal observer was greater than 99.4%. Thus, the models are essentially identical in the relevant range.
  - The choice of continuous uniform distributions allows for a decision rule that not only has closed form but is also easily interpretable (as we show in the subsection "Models and model fitting—Decision rules").
- We assume that the noisy memories x_1 and x_2 are conditionally independent given the sample orientations θ_1 and θ_2. Formally, $p(x_1,x_2|\theta_1,\theta_2) = p(x_1|\theta_1)\, p(x_2|\theta_2)$.
- We assume that $p(x_i|\theta_i)$ is a von Mises distribution.
- When the change happens in the first location (L=1), then $\varphi_1 = \theta_1 + \Delta$ and $\varphi_2 = \theta_2$. When the change happens in the second location (L=2), then $\varphi_1 = \theta_1$ and $\varphi_2 = \theta_2 + \Delta$.

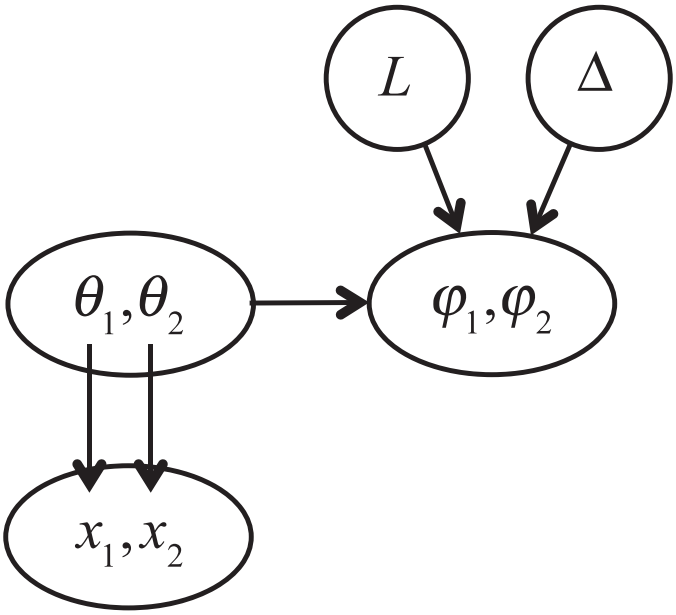

Figure A1. Graphical depiction of the generative model on which the decision rule is based.

Step 2: Inference

Now that we have specified the generative model, we can do inference. The observer infers L based on the noisy memories x_1 and x_2 and the test orientations φ_1 and φ_2. An ideal observer does this by computing the posterior distribution over L, $p(L|x_1,x_2,\varphi_1,\varphi_2)$. Since L is binary, all information about the posterior is contained in the log posterior ratio, which can be rewritten using Bayes' rule:

$$\begin{aligned}
&\log\frac{p(L=1|x_1,x_2,\varphi_1,\varphi_2)}{p(L=2|x_1,x_2,\varphi_1,\varphi_2)}\\
&\quad=\log\frac{p(L=1)}{p(L=2)}+\log\frac{p(x_1,x_2,\varphi_1,\varphi_2|L=1)}{p(x_1,x_2,\varphi_1,\varphi_2|L=2)}\\
&\quad=\log\frac{p(x_1,x_2,\varphi_1,\varphi_2|L=1)}{p(x_1,x_2,\varphi_1,\varphi_2|L=2)},
\end{aligned}$$

since $p(L=1) = p(L=2)$. We evaluate the likelihood of

$L = 1$ (the probability of the memories x_1 and x_2 if the change happened at the first location):

$$\begin{aligned}
&p(x_1, x_2, \varphi_1, \varphi_2 | L = 1) \\
&= \iiint p(x_1|\theta_1) p(x_2|\theta_2) \\
&\quad \times\ p(\varphi_1, \varphi_2 | \theta_1, \theta_2, \Delta, L = 1) \\
&\quad \times\ p(\theta_1, \theta_2) p(\Delta) d\theta_1 d\theta_2 d\Delta \\
&= \iiint p(x_1|\theta_1) p(x_2|\theta_2) \delta(\varphi_1 - \theta_1 - \Delta) \\
&\quad \times\ \delta(\varphi_2 - \theta_2) \frac{1}{2\pi k} d\theta_1 d\theta_2 d\Delta \\
&= \frac{1}{2\pi k} \int p(x_1|\theta_1 = \varphi_1 - \Delta) p(x_2|\theta_2 = \varphi_2) d\Delta \\
&= \frac{1}{2\pi k} \frac{1}{2\pi I_0(\kappa_2)} e^{\kappa_2 \cos(x_2 - \varphi_2)} \\
&\quad \times \int \frac{1}{2\pi I_0(\kappa_1)} e^{\kappa_1 \cos(x_1 - \varphi_1 + \Delta)} \\
&\quad \times\ d\Delta = \frac{1}{2\pi k} \frac{1}{2\pi I_0(\kappa_2)} e^{\kappa_2 \cos(x_2 - \varphi_2)}
\end{aligned}$$

Similarly, the likelihood of $L = 2$ (the probability of the memories if the change happened at the second location) is

$$p(x_1, x_2, \varphi_1, \varphi_2 | L = 2) = \frac{1}{2\pi k} \frac{1}{2\pi I_0(\kappa_1)} e^{\kappa_1 \cos(x_1 - \varphi_1)}$$

Combining both expressions, we find for the log posterior ratio

$$\begin{aligned}
&\log \frac{p(L = 1 | x_1, x_2, \varphi_1, \varphi_2)}{p(L = 2 | x_1, x_2, \varphi_1, \varphi_2)} \\
&= \log \frac{I_0(\kappa_1)}{I_0(\kappa_2)} + \kappa_2 \cos(x_2 - \varphi_2) \\
&\quad - \kappa_1 \cos(x_1 - \varphi_1)
\end{aligned}$$

The ideal observer responds that the change occurred at location 1 when the log posterior ratio is positive, i.e., when

$$\begin{aligned}
&\log I_0(\kappa_1) - \kappa_1 \cos(x_1 - \varphi_1) > \log I_0(\kappa_2) \\
&\quad - \kappa_2 \cos(x_2 - \varphi_2).
\end{aligned}$$

This is Equation 3 in the main text.

Appendix B: Parameter estimates

		Tested range			Monkeys		Humans
					M1	M2	
Model	Parameter	Min	Step	Max	Mean ± *SEM*	Mean ± *SEM*	Mean ± *SEM*
VP−VP	$\bar{J}_{high}$	0	1.01	100	7.0 ± 1.9	6.8 ± 1.2	30.3 ± 3.6
	$\bar{J}_{low}$	0	1.01	100	1.08 ± 0.26	1.07 ± 0.24	1.72 ± 0.34
	τ	0.1	1.1	100	61 ± 18	63 ± 17	45 ± 11
	lapse	0	0.002	0.2	0.041 ± 0.038	0.044 ± 0.034	0.046 ± 0.018
VP−FP	$\bar{J}_{high}$	0	1.01	100	6.34 ± 0.92	9.2 ± 1.9	38.4 ± 4.4
	$\bar{J}_{low}$	0	1.01	100	1.09 ± 0.28	1.46 ± 0.54	2.32 ± 0.45
	τ	0.1	1.1	100	22.7 ± 4.7	41 ± 10	42.9 ± 7.8
	lapse	0	0.002	0.2	0.210 ± 0.002	0.19 ± 0.026	0.031 ± 0.013
VP−SP	$\bar{J}_{high}$	0	1.01	100	9.8 ± 3.7	7.85 ± 0.94	29.0 ± 3.1
	$\bar{J}_{low}$	0	1.01	100	4.1 ± 1.8	4.0 ± 0.47	5.35 ± 0.77
	τ	0.1	1.1	100	52 ± 25	70 ± 12	22.4 ± 5.4
	lapse	0	0.002	0.2	0.17 ± 0.061	0.011 ± 0.031	0.088 ± 0.018

Table B1. Parameter ranges and parameter estimates in the three models. For monkeys, means and standard errors were computed across 100 bootstrapped data sets. For humans, means and standard errors were computed across subjects. Disclaimer: Parameter estimates of poorly fitting models should not be taken seriously.

Appendix C: Model recovery

To reduce the probability that the implementation of our model contains mistakes, and to find out to what extent the models are in principle distinguishable, we generated synthetic data from each of the three models and fitted all three models to each synthetic data set. Figure C1 shows that the models are correctly recovered, both in terms of summary statistics and in terms of log marginal likelihood.

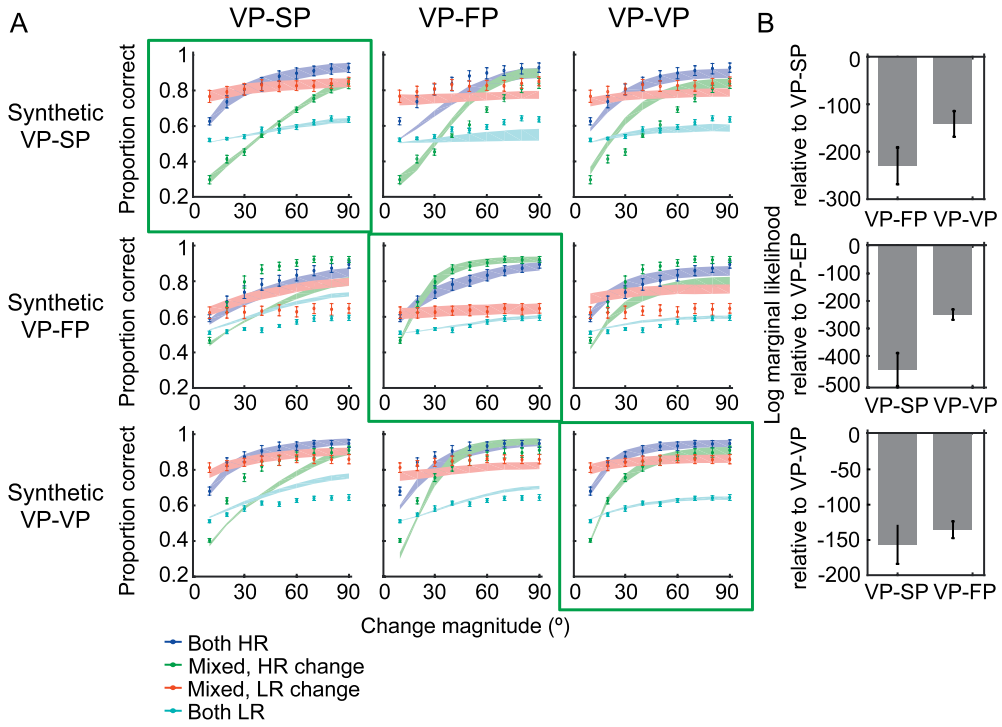

Figure C1. Model recovery. Each row represents synthetic data generated from one of the three models. The first three columns represent the fits of the same three models. Visually, the best fits are along the diagonal (green boxes): The best-fitting model is the one that generated the data. The right column shows the log marginal likelihood relative to the true generating model: Indeed, in each case, the log marginal likelihood of the two alternative models was lower. This shows that the models are distinguishable and provides some level of confidence that our implementation of the three models is correct.