The same type of visual working memory limitations in humans and monkeys

Deepna T. Devkar
Department of Neurobiology & Anatomy, The University of Texas Medical School, Houston, TX, USA
Department of Psychology, New York University, New York, NY, USA

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

Wei Ji Ma
Department of Psychology, and Center for Neural Science, New York University, New York, NY, USA

Rhesus monkeys are widely used as an animal model for human memory, including visual working memory (VWM). It is, however, unknown whether the same principles govern VWM in humans and rhesus monkeys. Here, we tested both species in nearly identical change-localization paradigms and formally compared the same set of models of VWM limitations. These models include the classic item-limit model and recent noise-based (resource) models, as well as hybrid models that combine a noise-based representation with an item limit. By varying the magnitude of the change in addition to the typical set size manipulation, we were able to show large differences in goodness of fit among the five models tested. In spite of quantitative performance differences between the species, we find that the variable-precision model—a noise-based model—best describes the behavior of both species. Adding an item limit to this model does not help to account for the data. Our results suggest evolutionary continuity of VWM across primates and help establish the rhesus monkey as a model system for studying the neural substrates of multiple-item VWM.

Introduction

Understanding cognition requires understanding its limitations. While cognitive limitations have been extensively characterized in humans, a complete understanding of their neural basis requires invasive studies in nonhuman animals. It cannot, however, be blindly assumed that findings from such studies will transfer to humans. This important concern can be partially preempted by demonstrating that the cognitive behavior of humans and of nonhuman animals are best described by the same models. In other words, model-defined similarity of human and nonhuman behavior might help justify claims that invasive studies in nonhuman animals can teach us about human cognition.

Here, we pursue this goal as it pertains to visual working memory (VWM; Luck & Vogel, 2013; Ma, Husain, & Bays, 2014). VWM is limited in two aspects: time—how long memories are maintained—and content—what is remembered and how well. Monkey studies of VWM have traditionally focused on the process of maintaining a single memory item over time (Funahashi, Bruce, & Goldman-Rakic, 1989; Fuster & Alexander, 1971; Miller, Erickson, & Desimone, 1996). More recently, they have started to address issues related to VWM content, in particular the effect of the number of items in a display (Buschman, Siegel, Roy, & Miller, 2011; Elmore et al., 2011; Heyselaar, Johnston, & Paré, 2011; Lara & Wallis, 2012, 2014; Warden & Miller, 2007). However, these studies did not focus on formally comparing mathematical models of VWM limitations. Here, we quantitatively compare multiple VWM models in both humans and monkeys based on a nearly identical experimental paradigm.

VWM content limitations have traditionally been described using item-limit models (Awh, Barton, & Vogel, 2007; Cowan, 2001; Fukuda, Awh, & Vogel, 2010; Luck & Vogel, 1997; Pashler, 1988). According to these models, only a fixed number of items (the

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

doi: 10.1167/15.16.13 Received February 11, 2015; published December 31, 2015 ISSN 1534-7362

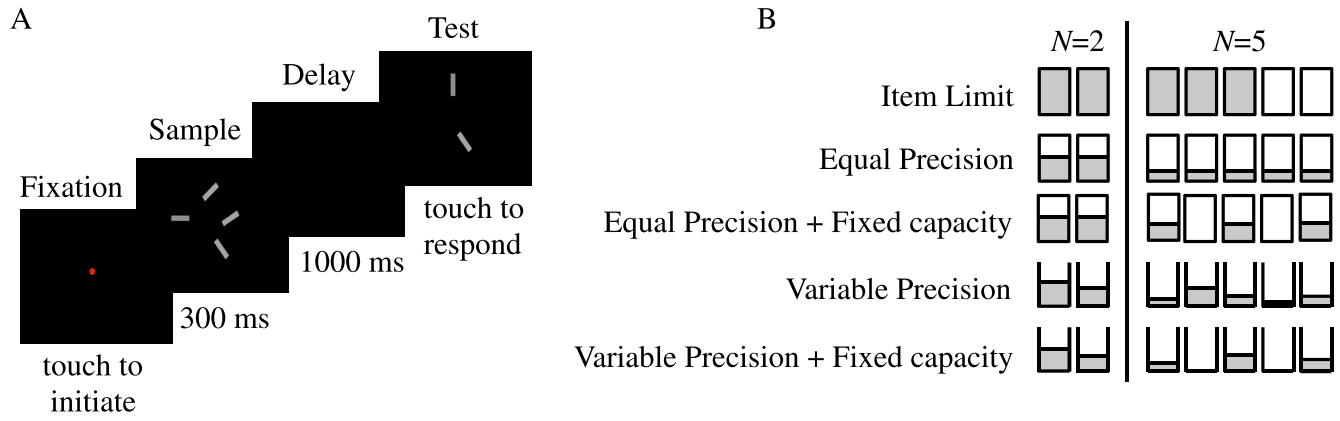

Figure 1. (A) Trial procedure in the change localization task. Subjects (monkeys and humans) were asked to report which item changed in orientation between the sample and test displays. (B) Schematic representation of precision (height of fill) of different items (boxes) in five leading models of VWM, at set sizes 2 and 5, with a hypothetical capacity limit of three for the IL, EPF, and VPF models. We use open boxes to indicate that in the models with variable precision we do not specify an upper bound to precision.

capacity) are held in memory with high quality, and no information is retained about any other items. Recently, an alternative category of models, based on human behavioral studies, has risen to prominence. In these "noise-based" or "resource" models, all items are remembered, but memories are noisy and memory precision is inversely related to the number of items (Bays & Husain, 2008; Keshvari, van den Berg, & Ma, 2013; van den Berg, Shin, Chou, George, & Ma, 2012; Wilken & Ma, 2004).

We tested fixed-capacity models against noise-based models in parallel in monkeys and in humans. We used a change localization paradigm that is similar to paradigms that have been successfully used to compare VWM models in humans (Keshvari et al., 2012, 2013; van den Berg et al., 2012). In our experimental design (Figure 1A), the subject viewed a sample array consisting of oriented bars, of which only the orientation was relevant. This display was followed by a 1-s blank screen, then by a test array consisting of two oriented bars selected from the sample array but with one having changed its orientation. The subject touched the location of the changed bar.

In addition to varying the number of sample items (set size), we varied the magnitude of the change in orientation randomly from trial to trial. While fixed-capacity and noise-based models can equally well account for observers' performance as a function of only set size, we have previously shown that the parametric variation of the magnitude of change allows one to effectively distinguish fixed capacity from noise-based models (Keshvari et al., 2013; van den Berg et al., 2012).

For each individual monkey and human subject, we compared four leading models of VWM limitations (Figure 1B). According to the item-limit (IL) model, a fixed number of items (the capacity) are kept in memory with infinite precision, while remaining items are absent from memory (Cowan, 2001; Luck & Vogel, 1997; Pashler, 1988). The equal-precision (EP) model postulates that all items are remembered with equal precision and that precision per item decreases with increasing set size (Palmer, 1990; Shaw, 1980). Decreasing precision is associated with increasing noise; that is, at a larger set size, each item is remembered in a noisier fashion. The equal-precision-with-fixed-capacity (EPF) model is a hybrid model that combines elements of the IL and EP models: Only a fixed number of items can be remembered, but a fixed precision budget is distributed across the remembered items (Zhang & Luck, 2008). For set sizes smaller than or equal to the capacity, this model predicts that precision will decrease with increasing set size. The variable-precision (VP) model is similar to the EP model in that all items are remembered with finite precision, but precision varies from item to item and trial to trial (Fougnie, Suchow, & Alvarez, 2012; Keshvari et al., 2013; van den Berg et al., 2012).

We also tested a recently proposed hybrid model—variable precision with fixed capacity (VPF)—that combines elements of the IL and VP models: Only a fixed number of items can be remembered, but precision varies randomly across items and trials (van den Berg & Ma, 2014). The four finite-precision models (EP, EPF, VP, and VPF) attribute all (EP and VP) or some (EPF and VPF) change localization errors to the difficulty of separating the signal from memory noise. For these four models, we used Bayesian inference to model the decision stage; on each trial, the observer reports the location that has the highest probability of containing the changed item (see Theory). The IL, EP, EPF, VP, and VPF models have two, two, three, three, and four free parameters, respectively.

Methods: Experiments

Monkeys

Subjects

Three adult male rhesus monkeys (*Macaca mulatta*; weights: M1 = 16.5 kg, M2 = 14.5 kg, M3 = 13.5 kg; ages: M1 = 17.5 years, M2 = 16.5 years, M3 = 12.5 years) were tested in a change localization experiment for five days a week. Food and water were regulated prior to experimental sessions. After completing daily testing, animals were returned to their caging room, where they were housed individually and received primate chow and water to maintain their normal body weight. All animal procedures were performed in accordance with the National Institutes of Health guidelines, approved by the institutional review board of the University of Texas Health Science Center at Houston, and supervised by the Institutional Animal Care and Use Committee.

Apparatus

During experimental sessions, the monkeys were placed unrestrained in a custom-made aluminum experimental chamber (47.5 cm wide × 53.1 cm deep × 66.3 cm high). An infrared touch screen detected touch responses to a 17-in. computer monitor. The touch responses were guided using a Plexiglas template with six cutouts (each a circle with a diameter of 2.5 cm) that were arranged on an imaginary circle with a diameter of 9.0 cm, matching the six possible locations of the stimuli, and a cutout in the center for touches to a fixation point. Using a computer-controlled relay interface (Model P10-12; Metrabyte, Taunton, MA), correct responses were rewarded with either a banana pellet or Tang orange drink (M1) or a banana pellet or cherry Kool-Aid (M2 and M3). The relay interface controlled the illumination of the chamber using 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 touch screen and the monitor marked the start of the next trial. Throughout testing, the monkeys were monitored with a video camera outside the chamber and focused through a small glass-covered port on the right side of the chamber. Experimental sessions were designed, operated, and recorded using a custom program written in Microsoft Visual Basic 6.0.

Stimuli

Stimuli consisted of 1.8 cm × 0.4 cm gray bars with luminosity of 190 cd/m^2 displayed on a black background. Based on the average distance of the monkey from the screen (approximately 35 cm), the stimuli subtended a visual angle of approximately 2.9° × 0.65°. Stimuli were presented in six possible locations on the screen, arranged on an imaginary circle of radius 7.4° (see Apparatus).

Trial procedure

Each trial began with a red fixation point in the center of the screen. The monkey had to make a one-touch response to the fixation point, which initiated the presentation of a sample display. This display contained two or more items (see later), and had a duration that differed between monkeys and between training and testing (see later). After a delay of 1000 ms, the test display was presented, which always consisted of two items placed at the same locations as two items from the sample display. One test item had the same orientation as the corresponding item in the sample display, and the other test item had a different orientation. The monkey's task was to identify which item had changed and to touch that item. The test display remained on the screen until response. Correct responses were rewarded. An intertrial interval of 3 s followed the response, during which a green light illuminated the chamber and the screen was dark.

Training

Two of the monkeys that participated in this study (M2 and M3) had been previously trained in a change-localization task using clip art images and colored squares (Elmore et al., 2011). For these two monkeys, we intermixed trials of oriented bars (new stimuli) with trials of colored squares for initial task acquisition. Once the monkeys' performance on these orientation trials was similar to their baseline color-trial performance, we began training them with only orientation trials. Since M1 had not been previously trained on this task, we directly trained him with oriented bars. All three monkeys were first trained at set sizes 2 and 3, change magnitudes of 22.5°, 45°, 67.5°, and 90°, and a sample viewing time of 1000 ms. Once overall accuracy reached approximately 70%, set sizes 4 and 5 and finer change magnitudes (10° to 90° in 10° increments) were gradually introduced. Finally, we gradually reduced sample-viewing times while maintaining approximately 70% accuracy on trials with set size 2. For M1 and M3, this led to a viewing time of 300 ms, and for M2 to a viewing time of 600 ms. Total training lasted approximately 8 months.

Testing

The sample display was shown for 300 ms for M1 and M3, and 600 ms for M2. Set size was 2, 3, 4, or 5.

Set sizes were pseudorandomized within each 192-trial block (48 trials per set size). The orientations of the sample items were drawn independently from a discrete uniform distribution over 18 possible orientations (–90° to 80° in increments of 10°). The orientation of the changed item in the test display was drawn from the same distribution, except that the orientations of the other sample stimuli were excluded. (This exception is unnecessary and potentially problematic because it slightly changed the statistics of the task. However, since at most four out of 18 orientations were excluded, and observers probably did not notice it, we expect the impact to be small and we did not model it.) 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 women, two men) aged 21–33 years (mean age = 27.1 years) participated. Each subject visited the lab for two 1.5-hr sessions and was compensated $10 per session. Study procedures were approved by the institutional review board of the University of Texas Health Science Center at Houston.

Apparatus and stimuli

Subjects were seated in a chair in a small room equipped with a computer. At the beginning of the experiment, the distance between the chair and the screen was adjusted 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. The monitor and touch screen were identical to those used for monkeys. Two 25-W light bulbs were mounted on the wall behind the subjects to provide feedback. Stimuli were identical to those used for monkeys.

Trial procedure

The trial procedure was identical to that for the monkeys, except for the feedback. Feedback 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.

Training and testing

Each subject completed two testing sessions, each consisting of three 192-trial blocks, for a total of 1,152 trials per subject. Subjects were given a 10-min break in between blocks. Each subject completed eight practice trials at the beginning of the first session.

Theory

We compared five models of behavior in this task. In the IL model, noise does not play a role. In the other four models, noise does play a role, which will require a model for how subjects integrate information from noisy measurements.

For simplicity, we mapped orientation space to the interval $[0, 2\pi)$ by multiplying all orientations and orientation-change magnitudes by 2 before analysis. All equations in this article are consistent with this convention, but orientations and orientation changes in the figures are back in actual orientation space.

IL model

In the IL model (Cowan, 2001; Luck & Vogel, 1997; Pashler, 1988), observers cannot store more than K items. When $N \leq K$, all items are stored. The probability of being correct is then $1 - \varepsilon$, where ε accounts for lapses of attention and unintended responses. When $N > K$, K randomly selected items from the sample display are stored. When the test display appears, there are three scenarios to consider:

- Both test items correspond to stored sample items. This happens with probability $\frac{K(K-1)}{N(N-1)}$. The probability of being correct is then $1 - \varepsilon$.
- One test item corresponds to a stored sample item and the other does not. This happens with probability $2\frac{K(K-1)}{N(N-1)}$. The probability of being correct is then $1 - \varepsilon$.
- Neither test item corresponds to a stored sample item. This happens with probability $\frac{(N-K)(N-K-1)}{N(N-1)}$. The observer then has to guess about which item changed, and the probability of being correct is 0.5.

The overall proportion correct is then

$$\begin{aligned} p(\text{correct})(N,K) = & \\ & \frac{K(K-1)}{N(N-1)}(1-\varepsilon) + 2\frac{K(N-K)}{N(N-1)}(1-\varepsilon) \\ & + \frac{(N-K)(N-K-1)}{N(N-1)} \cdot 0.5 \\ & = 1 - \varepsilon - \frac{(N-K)(N-K-1)}{N(N-1)} \cdot (0.5-\varepsilon). \end{aligned}$$

Storing all N items ($K = N$) yields the same proportion correct, namely $1 - \varepsilon$, as storing only $N - 1$ items, since even if one test item is not stored, the trial can be answered correctly by using the other test item. As can be seen from the equation, in the IL model the

proportion correct depends on set size but not on change magnitude.

Noise-based models

We now turn to models in which VWM is noisy (Ma et al., 2014; Wilken & Ma, 2004). We assume that both orientations in the test display, which we denote φ_1 and φ_2, are known noiselessly to the observer, because they remain on the screen until the subject responds. We model the memories of the orientations in the sample display as noisy. Noise can stem from encoding (presentation time was limited) or maintenance of memories; we do not distinguish between these sources. We model the noisy memory of the ith item in the sample display, denoted x_i ($i = 1, \ldots, N$), as following a von Mises distribution (a circular analog of a Gaussian distribution, used because orientation space is periodic) centered at the true stimulus θ_i with concentration parameter κ_i:

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

where I_0 is the modified Bessel function of the first kind of order 0 (Mardia & Jupp, 1999). The concentration parameter controls the width of the noise distribution, and the Bessel function serves as a normalization. We have postulated previously that the role of precision is played by the Fisher information in this memory representation, denoted J_i (Keshvari et al., 2013; van den Berg et al., 2012). Fisher information determines the best possible performance of any unbiased estimator through the Cramér–Rao bound (Cover & Thomas, 1991). When the measurement x follows a Gaussian distribution, Fisher information is equal to inverse variance, $J = \frac{1}{\sigma^2}$. When neural variability is Poisson-like, Fisher information is proportional to the gain of a population (Seung & Sompolinsky, 1993). Thus, our choice of using Fisher information for precision is consistent with an interpretation of neural activity as "memory resource" (Bays, 2014; Ma et al., 2014; van den Berg et al., 2012). For Equation 1, Fisher information is related to the concentration parameter through

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

where I_1 is the modified Bessel function of the first kind of order 1. The relationship between precision and the concentration parameter is nearly the identity mapping, and none of our results would qualitatively change if we were to replace J_i with κ_i.

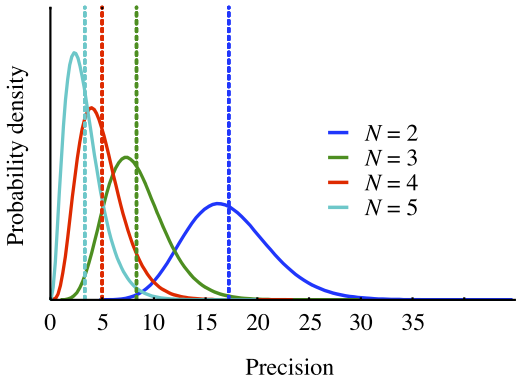

Figure 2. Illustration of probability density functions over precision in the VP model for four set sizes. Mean precision, marked by dashed lines, is inversely related to set size. For the parameters $\bar{J}_{N=1}$, τ, and α, we used mean parameter estimates from humans (Table A1).

In the EP model (Bays & Husain, 2008; Palmer, 1990), the precision of each item is inversely related to set size through a power law:

$$J_i = \frac{J_{N=1}}{N^\alpha},$$

where $J_{N=1}$ is the precision with which a single item is stored. The precision of all items in a display is equal.

In the EPF model (also known as slots-plus-resources; Zhang & Luck, 2008), no more than K items can be stored. Thus the number of stored items is $\min(N, K)$. The precision of a stored item is inversely related to the number of stored items through a power law:

$$J_i = \frac{J_{N=1}}{\min(N, K)^\alpha}.$$

The precision associated with a nonstored item is zero. When $N \leq K$, the EPF model is equal to the EP model. The slots-plus-averaging model is very similar to this model (as was quantitatively shown by van den Berg, Awh, & Ma, 2014).

In the VP model (Fougnie et al., 2012; Keshvari et al., 2013; van den Berg et al., 2012), precision exhibits fluctuations across both space and time. To be concrete, we assume that the precision values associated with the N items are drawn independently from a gamma distribution with mean $\bar{J}$ and scale parameter τ (a flexible family of distributions on the positive real line). We further assume that its mean is inversely related to set size through a power law:

$$\bar{J} = \frac{\bar{J}_{N=1}}{N^\alpha},$$

where $\bar{J}_{N=1}$ is the mean precision of a single item (Figure 2).

The VPF model (van den Berg & Ma, 2014) is equal to the VP model, but the number of stored items is

$\min(N, K)$; thus, no more than K items can be stored. The precision of a stored item is again drawn from a gamma distribution with mean $\bar{J}$ and scale parameter τ, and the mean is inversely related to the number of stored items through a power law:

$$\bar{J} = \frac{\bar{J}_{N=1}}{\min(N, K)^{\alpha}}.$$

The precision associated with a nonstored item is zero.

The IL, EP, EPF, VP, and VPF models have two, two, three, three, and four free parameters, respectively.

Decision rules

So far, we have described the encoding stage: how stimuli give rise to noisy memories. What is also needed in each of the noise-based models is a description of how the observer makes the two-alternative change-localization decision based on the noisy memories and the test display. We use an ideal (Bayesian) observer to describe this process. The resulting decision rule is similar to the ideal-observer models of related N-alternative change localization and change-detection tasks (Keshvari et al., 2012, 2013; van den Berg et al., 2012), but differs in the details.

We begin by describing the decision process for the EP and VP models. The relevant variables are the location L of the change (1 or 2), the magnitude Δ of the change, the relevant sample orientations θ_1 and θ_2 (all other sample items are irrelevant to the decision), their noisy memories x_1 and x_2, and the two test orientations φ_1 and φ_2. The ideal observer responds that the change occurred at location 1 when the log posterior ratio is positive:

$$\log\left(\frac{I_0(\kappa_1)}{I_0(\kappa_2)}\right) + \kappa_2\cos(x_2 - \varphi_2) - \kappa_1\cos(x_1 - \varphi_1) > 0. \quad (3)$$

The derivation of this decision rule can be found in Appendix A; we have assumed that the observer knows the values of κ_1 and κ_2 on each trial. The decision rule is valid for both the VP and EP models. In the VP model, precision per item is a random variable, and therefore κ_1 and κ_2 will generally not be equal to each other. However, in the case of the EP model, we have $\kappa_1 = \kappa_2$ and the inequality simplifies to

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

This rule is intuitive: The observer reports that the change occurred at location 1 when the angular distance between the noisy memory at location 2 and the test orientation at location 2 is smaller than the corresponding distance at location 1 (and thus the cosine is larger). There is then more evidence that the change occurred at location 1. One can think of Equation 3 as a precision-weighted version of Equation 4.

The EPF model is very similar to the EP model, but with one difference when $N > K$. Then, a noisy measurement has a probability of not being stored. This is equivalent to setting the concentration parameter of the corresponding memory to 0. Thus, we can immediately obtain the decision rule from the EPF model by taking special cases of Equation 3:

$$\begin{cases} \text{Report location 1 when...} \\ \cos(x_2 - \varphi_2) > \cos(x_1 - \varphi_1) \\ \qquad\qquad \text{if both items were stored;} \\ \kappa\cos(x_1 - \varphi_1) < \log I_0(\kappa) \quad \text{if only item 1 was stored;} \\ \kappa\cos(x_2 - \varphi_2) < \log I_0(\kappa) \quad \text{if only item 2 was stored;} \\ \text{...and guess randomly when neither item was stored.} \end{cases} \quad (5)$$

The second and third inequalities may seem counterintuitive, since they only involve one memory. However, they make sense: Even when the observer has only the memory corresponding to one of the test items, the discrepancy between the memory and the test is still informative about whether or not the change occurred in that one item.

The VPF model is identical to the VP model when $N \leq K$. When $N > K$, just as in the EPF model, a noisy measurement has a probability of not being stored (precision = 0). But unlike in the EPF model, the concentration parameters κ_1 and κ_2 in the VPF model are independent. With these modifications, we can again take the special cases of Equation 3 and obtain the decision rules for the VPF model:

$$\begin{cases} \text{Report location 1 when...} \\ \log\left(\frac{I_0(\kappa_1)}{I_0(\kappa_2)}\right) + \kappa_2\cos(x_2 - \varphi_2) > \kappa_1\cos(x_1 - \varphi_1) \\ \qquad\qquad \text{if both items were stored;} \\ \kappa_1\cos(x_1 - \varphi_1) < \log I_0(\kappa_1) \quad \text{if only item 1 was stored;} \\ \kappa_2\cos(x_1 - \varphi_2) > \log I_0(\kappa_2) \quad \text{if only item 2 was stored;} \\ \text{...and guess randomly when neither item was stored.} \end{cases} \quad (6)$$

Model predictions

If we had access to the observer's noisy memories x_1 and x_2 on each trial, the models would predict the observer's response exactly. Since we do not know x_1 and x_2, the best we can do is to compute the *probability* of being correct for a given stimulus condition. Under the assumptions in our generative model, the stimulus

condition is determined completely by set size N and change magnitude Δ, and the values of θ_1 and θ_2 are irrelevant. Thus, we are interested in the probability that the decision rule (Equation 3 for VP, Equation 4 for EP, Equation 5 for EPF, and Equation 6 for VPF) returns the correct location when the memories x_1 and x_2 follow their model-specific distributions given N and Δ. Without loss of generality, we compute the proportion correct by taking $\theta_1 = \theta_2 = 0$ and $L = 1$, so that $\varphi_1 = 0$ and $\varphi_2 = \Delta$.

For the EP model, then,

$$p(\text{correct})(N,\Delta) = \Pr\Big(\cos x_2 > \cos(x_1 - \Delta); x_1, x_2 \sim \text{VM}(0,\kappa)\Big),$$

where $\text{VM}(\mu, \kappa)$ denotes the von Mises distribution with mean μ and concentration parameter κ, and we use the notation Pr(statement involving X; $X \sim$ distribution) to indicate the probability that the statement is true when X follows the given distribution.

For the VP model, both the decision rule and the distributions of x_1 and x_2 are different:

$$p(\text{correct})(N,\Delta) = \Pr\left(\log\left(\frac{I_0(\kappa_1)}{I_0(\kappa_2)}\right) + \kappa_2\cos x_2 > \kappa_1\cos(x_1 = \Delta); \; \begin{array}{r} x_1 \sim \text{VM}(0,\kappa_1), \\ x_2 \sim \text{VM}(0,\kappa_2), \\ J_i \sim \text{Gamma}(\bar{J},\tau) \end{array}\right),$$

where κ_i is related to J_i through Equation 2.

For the EPF model, the proportion correct is computed as a sum across the four possibilities for which items were stored (see Equation 5):

$$\begin{aligned} p(\text{correct})(N,\Delta) &= \frac{K(K-1)}{N(N-1)} \cdot \Pr\Big(\cos x_2 > \cos(x_1 - \Delta); \; x_1, x_2 \sim \text{VM}(0,\kappa)\Big) \\ &+ \frac{K(N-K)}{N(N-1)} \cdot \Pr\Big(\kappa\cos(x_1 - \Delta) > \log I_0(\kappa); \; x_1 \sim \text{VM}(0,\kappa)\Big) \\ &+ \frac{K(N-K)}{N(N-1)} \cdot \Pr\Big(\kappa\cos x_2 < \log I_0(\kappa); \; x_2 \sim \text{VM}(0,\kappa)\Big) \\ &+ \frac{(N-K)(N-K-1)}{N(N-1)} \cdot 0.5. \end{aligned}$$

For the VPF model, the proportion correct is computed as a sum across the four possibilities for which items were stored (see Equation 6):

$$\begin{aligned} p(\text{correct})(N,\Delta) &= \frac{K(K-1)}{N(N-1)} \cdot \Pr\left(\log\left(\frac{I_0(\kappa_1)}{I_0(\kappa_2)}\right) + \kappa_2\cos x_2 > \kappa_1\cos(x_1 - \Delta); \; \begin{array}{r} x_1 \sim \text{VM}(0,\kappa_1), \\ x_2 \sim \text{VM}(0,\kappa_2), \\ J_i \sim \text{Gamma}(\bar{J},\tau) \end{array}\right) \\ &+ \frac{K(N-K)}{N(N-1)} \cdot \Pr\Big(\kappa\cos(x_1 - \Delta) > \log I_0(\kappa); \; x_1 \sim \text{VM}(0,\kappa_1), J_i \sim \text{Gamma}(\bar{J},\tau)\Big) \\ &+ \frac{K(N-K)}{N(N-1)} \cdot \Pr\Big(\kappa\cos x_2 < \log I_0(\kappa); \; x_2 \sim \text{VM}(0,\kappa_2), J_i \sim \text{Gamma}(\bar{J},\tau)\Big) \\ &+ \frac{(N-K)(N-K-1)}{N(N-1)} \cdot 0.5. \end{aligned}$$

Each of these proportions correct was determined through Monte Carlo simulation. For each (N, Δ) combination, we drew 10,000 random samples of x_1 and x_2 (and in the case of the VP and VPF models, of J_1 and J_2 first). For each sample, we evaluated the decision rule and then computed the proportion of correct responses across all samples.

Finally, for each model, we discretized parameter space (Table A1) and calculated a lookup table in which each entry gave the predicted probability of a correct response at one (N, Δ) combination for one parameter combination.

Methods: Model fitting and model comparison

Model fitting

Denoting all parameters of a model by a vector $\mathbf{t}$, the log likelihood of $\mathbf{t}$ (the parameter log likelihood) is

$$\begin{aligned} \text{LL}(\mathbf{t}) &= \log p(\text{data}|\text{model},\mathbf{t}) \\ &= \log \prod_{i=1}^{n_{\text{trials}}} p(\text{correctness}_i|N_i,\Delta_i,\mathbf{t}), \end{aligned} \quad (7)$$

where the product is over trials (from 1 to n_{trials}) and correctness$_i$ is 1 if the subject was correct on the ith trial and 0 if not. We can rewrite this as

$$\begin{aligned}\text{LL}(\mathbf{t}) &= \sum_{i=1}^{n_{\text{trials}}} \log p(\text{correctness}_i | N_i, \Delta_i, \mathbf{t}) \\ &= \sum_N \sum_\Delta \Big[n(N, \Delta, \text{correct}) \cdot \log p(\text{correct}|N, \Delta, \mathbf{t}) \\ &\quad + n(N, \Delta, \text{incorrect}) \cdot \log\Big(1 - p(\text{correct}|N, \Delta, \mathbf{t})\Big)\Big] \end{aligned} \tag{8}$$

where we grouped trials by set size N, change magnitude Δ, and whether the observer was correct or incorrect, and $n(N, \Delta, \text{correct})$ is the number of trials with a particular N, Δ, and correctness.

For each subject data set, we used Equation 8 and the precomputed lookup table of model predictions mentioned before to find the log likelihood of each parameter combination. The parameter combination on this grid that maximized the log likelihood gave the estimates of the parameters. The model predictions corresponding to that parameter combination were then used to compute the model fits to the psychometric curves. We denote the maximum of the parameter log likelihood LL($\mathbf{t}$) by $\text{LL}_{\max}$.

Bayesian model comparison

To compare models, we used Bayesian model comparison (MacKay, 2003), which should not be confused with the Bayesian observer model that we used earlier. Bayesian model comparison is based on the *log marginal likelihood* of a model m given the data: LML(model) = log p(data|model). The attribute "marginal" refers to an integration (marginalization) over the parameters:

$$\begin{aligned}\text{LML}(\text{model}) &= \log p(\text{data}|\text{model}) \\ &= \log \int p(\text{data}|\text{model}, \mathbf{t}) p(\mathbf{t}|\text{model})\, d\mathbf{t} \\ &= \log \int e^{\text{LL}(\mathbf{t})} p(\mathbf{t}|\text{model})\, d\mathbf{t}.\end{aligned}$$

For the parameter prior $p(\mathbf{t}|\text{model})$, we chose a product of uniform distributions (one for each parameter), with their domains just covering the grid used for model predictions and parameter estimation (see before). We denote the size of the range of the jth parameter by R_j, where $j = 1, \ldots, k$. We also peak-normalize the exponential term so as to avoid highly negative numbers in the exponent, which could cause numerical underflow; we add a correction to compensate for this. This gives

$$\begin{aligned}&\text{LML}(\text{model}) = \\ &\text{LL}_{\max} + \log\left(\left(\prod_{j=1}^{k} \frac{1}{R_j}\right) \int e^{\text{LL}(t) - \text{LL}_{\max}} d\mathbf{t}\right) \\ &= \text{LL}_{\max} + \log\left(\int e^{\text{LL}(t) - \text{LL}_{\max}} d\mathbf{t}\right) - \sum_{j=1}^{k} \log R_j.\end{aligned}$$

Finally, we approximate the integral through a Riemann sum (grid sum) over the same grid as used for model predictions and parameter estimation (see earlier). We denote the grid spacing of the jth parameter by δt_j. This leads to the equation we actually implemented:

$$\begin{aligned}&\text{LML}(\text{model}) = \\ &\text{LL}_{\max} + \log\left(\left(\prod_{j=1}^{k} \delta t_j\right) \sum_{t \text{ on grid}} e^{\text{LL}(t) - \text{LL}_{\max}}\right) \\ &\qquad - \sum_{j=1}^{k} \log R_j.\end{aligned} \tag{9}$$

The difference of the log marginal likelihood between two models is also called the log Bayes factor of those two models (Kass & Raftery, 1995).

Numerical values of the ranges R_j are specified in Table A1. Our choices for these ranges were initially guided by parameter estimates from previous publications (Keshvari et al., 2013; van den Berg et al., 2012). These ranges worked well for our human data. In the monkey data, however, we noticed that the parameter estimates of $\bar{J}_{N=1}$ and τ tended to be much smaller that the upper limits of these ranges. Since the computational time required for numerically evaluating the parameter likelihood in the Riemann sum is determined by the number of grid points, we reduced the ranges for those parameters so that—keeping the number of grid values within each range constant—we could obtain a finer resolution for our parameter estimates. This more efficient use of computational resources comes at the cost of no longer being able to interpret the uniform distribution $p(\mathbf{t}|\text{model})$ as a prior, because it is now (albeit weakly) informed by the data. We will comment on the consequences of this choice in the Results.

Parameter recovery and model recovery

To validate our methods, we applied them to data sets for which we knew the ground truth, namely synthetic data sets generated using one of the models. For each of the five models, we generated 10 synthetic data sets by independently drawing the parameter values from uniform distributions on the ranges

specified in the "Humans" column of Table A1. For each of these 50 data sets, we fitted all five models and computed their LMLs. We found that in each of the 50 data sets and for each of the three metrics, the model that was used to generate the data had the highest LML. In addition, for the correct model the parameter estimates were close to the parameters that were used to generate the synthetic data, with the exception of $\bar{J}_{N=1}$ and τ in the VP and VPF models. Those parameters were sometimes both overestimated or both underestimated, indicating that the data were approximately equally well fitted by a lower $\bar{J}_{N=1}$ and a lower τ as by a higher $\bar{J}_{N=1}$ and a higher τ. We conclude that we can trust the model comparison results but that the estimates of $\bar{J}_{N=1}$ and τ should be taken with a grain of salt.

Other model comparison metrics

We also used two other model comparison metrics, which are based not on marginalizing over the parameters but solely on the maximum of the parameter log likelihood $\mathrm{LL}_{\max}$, with a correction for the number k of free parameters in the model. These metrics are the corrected Akaike information criterion $\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n_{\mathrm{trials}}-k-1}$ (Akaike, 1974; Hurvich & Tsai, 1989) and the Bayesian information criterion $\mathrm{BIC} = -2\mathrm{LL}_{\max} + k\log n_{\mathrm{trials}}$ (Schwarz, 1978). In order to make these metrics comparable in magnitude to the marginal log likelihood LML(model), we report each of them multiplied by −0.5, so that the leading term is $\mathrm{LL}_{\max}$: $\mathrm{AICc}^* = -0.5\mathrm{AICc}$ and $\mathrm{BIC}^* = -0.5\mathrm{BIC}$.

Bootstrapping

Since we had only three monkey subjects, we used bootstrapping (Efron, 1993) for each monkey separately to estimate the standard errors on all summary statistics. The original data set for each monkey consisted of 11,520 rows (each row represented a trial) and three columns (set size, change magnitude of the changed item, and whether the trial was correct or incorrect). We sampled the rows (trials) with replacement from the original data set to create 11,520-trial bootstrapped data sets. We repeated this process to create 100 bootstrapped data sets for each monkey. For each bootstrapped data set, we estimated the parameters, computed psychometric curves, calculated R^2, and computed AICc*, BIC*, and LML. The means each of these was 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

For both species, the proportion correct decreased monotonically as a function of set size, with humans being substantially more accurate than monkeys (Figure 3A). A more detailed representation of the data is provided by the proportion correct as a function of change magnitude for each of the four set sizes (Figure 3B, C). We found large effects of both set size and change magnitude on VWM performance in both species (humans: two-way repeated-measures ANOVA)—set size: $F(3, 27) = 64.05$, $p < 0.001$; change magnitude: $F(8, 72) = 80.36$, $p < 0.001$.

Model fitting

We used maximum-likelihood estimation to fit the parameters in each model. For humans, we fitted the data of individual subjects. For each monkey, we fitted the individual data sets that we sampled using bootstrapping from the monkey's raw data (this gives error bars on parameter estimates). Parameter estimates are given in Appendix B (Table A1). Model fits to the monkeys' actual data (without bootstrapping) are given in Appendix C.

Model comparison

In spite of the large performance differences between species, it is possible that the underlying VWM mechanisms are the same. To test this possibility, we compared the four leading models of VWM limitations as well as a new hybrid model (VPF) for each individual monkey and human. We first used Bayesian model comparison, a likelihood-based method that automatically corrects for the number of free parameters (see Methods: Model fitting and model comparison). We found that the mean log marginal likelihoods of the VP and VPF models exceed those of the EPF, EP, and IL models for both species (Figure 4; Table 1); the VP and VPF models are not distinguishable. Moreover, the results are highly consistent across individual monkey and human subjects. Model comparison results on the monkeys' actual data (without bootstrapping) are given in Appendix C.

Under Methods: Model fitting and model comparison, we commented on the choices of the parameter ranges R_j in Equation 9, which for monkey subjects were (weakly) informed by the data. Fortunately, our qualitative results are reasonably robust to these

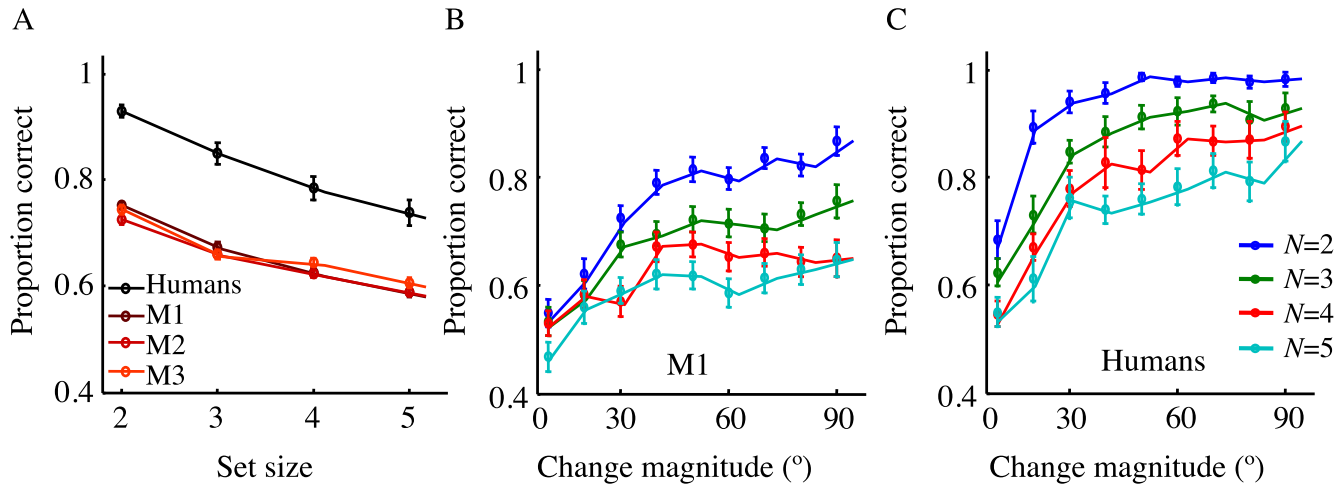

Figure 3. (A) Proportion correct as a function of set size for humans and three monkeys (M1, M2, M3). (B) Proportion correct as a function of set size N and change magnitude for M1 (mean ± standard error of the mean estimated from bootstrapped datasets). (C) The same as (B) but for humans (mean ± standard error of the mean across 10 subjects).

choices. If we assume that the parameter log likelihood is zero outside the narrower [0, 30] ranges of the $\overline{J}_{N=1}$ and τ parameters in the VP model chosen for monkey subjects, then the effect of changing these ranges to the wider [0, 100] ranges we used in humans is to reduce the VP log marginal likelihoods by $-2\log(30) - (-2\log(100)) = 2.4$, which would not change the finding that VP outperforms IL, EP, and EPF. Moreover, the effect of changing the [0, 30] ranges of the $\overline{J}_{N=1}$ and τ parameters in the VPF model for monkeys to the wider [0, 200] ranges we chose in humans is to reduce the VPF log marginal likelihoods by 3.8, which would not change the finding that VP and VPF are indistinguishable. The reason for this robustness against the choice of parameter ranges arises from the fact that our differences in log marginal likelihoods are largely driven by the $\mathrm{LL}_{\max}$ term.

Our results for both monkeys and humans also remain unchanged when we use AICc or BIC as an alternative model comparison metrics (Table 1). Unlike the log marginal likelihood, these model comparison metrics do not depend on parameter ranges. Again, this consistency follows from the model differences being dominated by differences in the $\mathrm{LL}_{\max}$ term.

Model checking

We substituted the fitted parameters into their respective models to create fits (predictions) for the summary statistics in Figure 3. The model fits to the

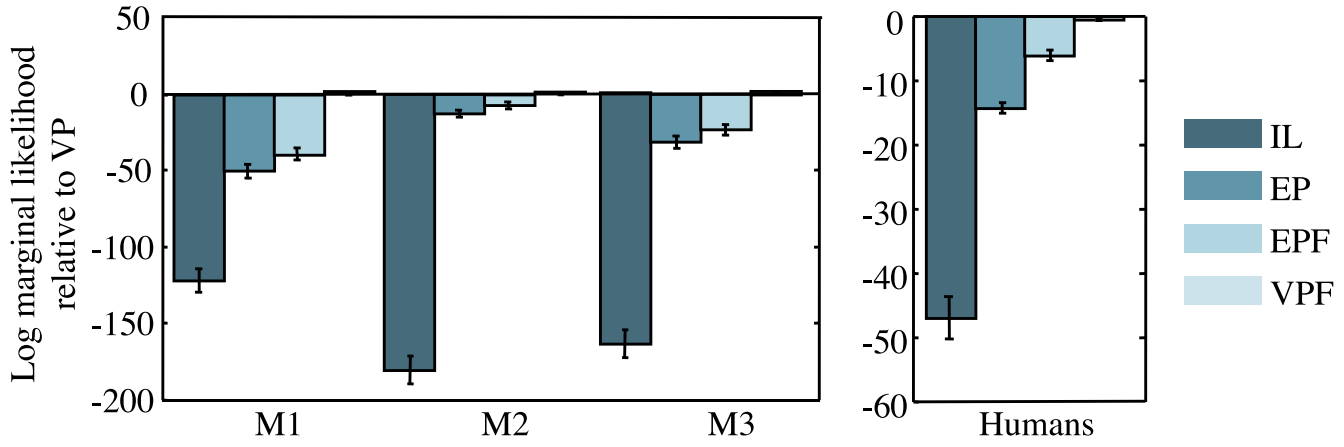

Figure 4. Bayesian model comparison for monkeys M1, M2, and M3 and humans, showing the log marginal likelihood of each model (IL, EP, EPF, VPF) minus that of the VP model (mean ± standard error of the mean). A value of $-x$ means that the data are e^x times more probable under the VP model. The VP and VPF models account equally well for the data, and much better than the other models.

	AICc*(model) − AICc*(VP)	BIC*(model) − BIC*(VP)		LML(model) − LML(VP)	
Model	Mean	Mean	Standard error of the mean	Mean	Standard error of the mean
IL					
M1	−125	−122	15	−121	15
M2	−183	−180	18	−180	18
M3	−167	−164	18	−163	18
Humans	−47.2	−45.7	6.8	−47.1	6.6
EP					
M1	−47.5	−44.8	9.2	−48.9	9.1
M2	−12.8	−10.1	4.6	−12.7	4.8
M3	−30.3	−27.6	7.8	−31.3	8.1
Humans	−12.9	−11.4	1.5	−14.4	1.7
EPF					
M1	−40.2	−40.2	7.9	−39.0	7.8
M2	−9.3	−9.3	4.4	−6.7	4.6
M3	−24.0	−24.0	6.7	−22.6	6.9
Humans	−7.6	−7.6	1.5	−6.2	1.6
VPF					
M1	−1.3	−4.18	0.83	1.5	1.5
M2	−2.2	−4.00	0.91	1.20	0.81
M3	−0.56	−3.2	1.5	2.0	1.1
Humans	−1.46	−3.00	0.32	−0.57	0.31

Table 1. Model comparison. *Notes*: For model comparison metrics, we use scaled versions of the AICc and BIC defined by AICc* = −0.5AICc and so on, so that the leading term is the maximum log likelihood LL_{max} and these measures can be compared directly to the log marginal likelihood (LML). Values shown are the mean differences in the model comparison metrics between the IL, EP, EPF, and VPF models on the one hand and the VP model on the other hand. A negative value means that the VP model fits better. The standard error of the mean is the same between the AICc* and BIC* because these measures differ only in their penalty terms.

psychometric curve of performance as a function of set size were good for all models (Figure 5A). The added manipulation of change magnitude, however, clearly separates these model fits (Figure 5B and Figure 5C). The psychometric curves from both species are best described by the two variable-precision models, VP and VPF, followed by the EPF model, the EP model, and the IL model (Table 2).

Comparison between species

Our model comparison suggests that the fundamental nature of VWM limitations is the same in both species (Figure 5), with quantitative differences reflected only in the parameter values within the same model (Table A1). For example, mean precision $\overline{J}_{N=1}$ was much lower in monkeys than in humans, which might reflect attentional differences between the two species. The exponent α in the relationship between mean precision and set size was similar across monkeys and somewhat higher in humans. In both species, however, the values were more negative than −1, indicating steep decreases in mean precision as set size increases. In the VPF model, the number of remembered items K was fitted as 3.5 in monkeys and 4.1 in humans; while consistent with earlier reports of K, this parameter should be interpreted with caution: In light of the finding that the VPF model is indistinguishable from the VP model, we cannot rule out the possibility that there is no item limit at all.

	M1	M2	M3	Humans
IL	0.402 ± 0.041	0.222 ± 0.038	0.263 ± 0.049	0.228 ± 0.048
EP	0.718 ± 0.055	0.835 ± 0.031	0.817 ± 0.035	0.619 ± 0.037
EPF	0.755 ± 0.046	0.854 ± 0.030	0.817 ± 0.035	0.714 ± 0.043
VP	0.901 ± 0.023	0.885 ± 0.024	0.891 ± 0.020	0.799 ± 0.023
VPF	0.902 ± 0.022	0.887 ± 0.024	0.896 ± 0.020	0.803 ± 0.023

Table 2. R^2 values of the fits of the five models to the full psychometric curves (proportion correct as a function of set size and change magnitude) of both species. *Notes*: We note that R^2 is a much less principled measure of goodness of fit than AICc, BIC, or LML. If any conflicts were to exist, the latter three should be preferred. However, results are consistent across measures. See Figure 5B and C.

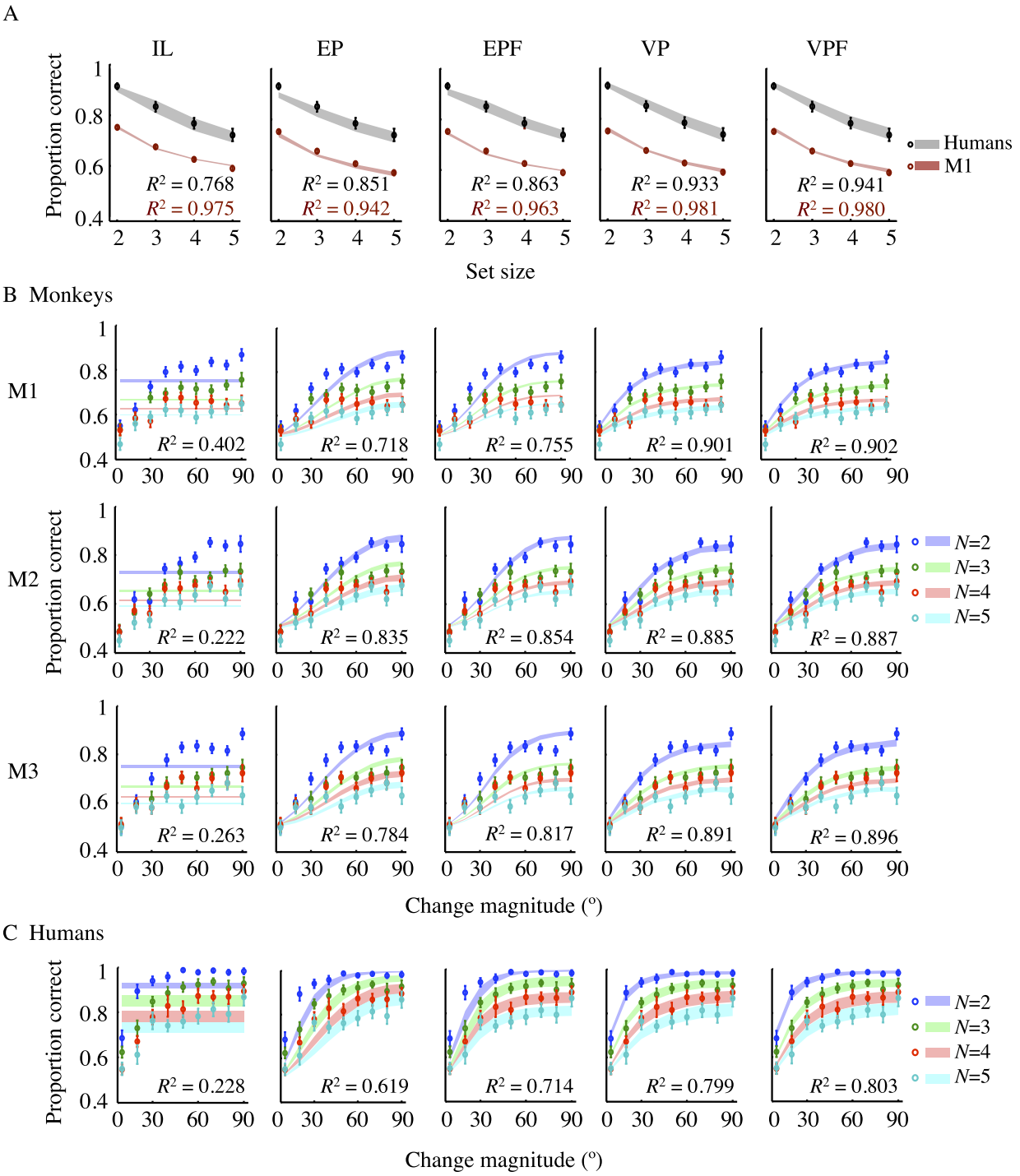

Figure 5. Model fits. (A) Proportion correct as a function of set size for M1 and humans. Circles and error bars are behavior; shaded areas are model fits. It is difficult to distinguish among the models based on these psychometric curves. (B) Proportion correct as a function of set size *N* and change magnitude, for monkeys (mean ± standard error of the mean estimated from bootstrapped data sets). Circles and error bars are behavior; shaded areas are model fits. (C) The same as (B) but for humans (mean ± standard error of the mean across subjects). This detailed representation of the data reveals that the VP and VPF models best account for the behavior of both species.

Models with lapse rate

We have seen that the EP and EPF models do not describe the data as well as the VP model. However, it might be that subjects randomly guess on some fixed proportion of trials. This would be different from guessing due to an item limit, because the proportion of those guesses depends on set size. Therefore, we tested

	AICc*(EP+lapse) − AICc*(EP)	AICc*(EP+lapse) − AICc*(VP)	AICc*(EPF+lapse) − AICc*(EPF)	AICc*(EPF+lapse) − AICc*(VP)
M1	13.8 ± 3.4	−33.6 ± 7.3	26.1 ± 6.2	−14.2 ± 4.5
M2	2.6 ± 1.6	−10.2 ± 3.7	4.7 ± 2.9	−4.5 ± 2.8
M3	10.1 ± 2.8	−20.2 ± 5.9	17.6 ± 5.5	−6.5 ± 3.8
Humans	2.3 ± 1.3	−10.6 ± 1.9	0.64 ± 0.70	−6.9 ± 1.5

Table 3. The EP and EPF models with a lapse rate fit better than the corresponding models without a lapse rate; however, they still both fit worse than the VP model.

the EP and EPF models augmented with a lapse rate (Table 3). Both in monkeys and in humans, adding a lapse parameter improves the goodness of fit of the EP and EPF models. However, in both species, the VP model outperforms the EP and EPF models with lapse.

Discussion

We tested monkeys and humans in a nearly identical change localization paradigm and compared five models of VWM limitations. Like all previous change detection and change localization studies, both in humans (Keshvari et al., 2012, 2013; van den Berg et al., 2012; Wilken & Ma, 2004) and in monkeys (Buschman et al., 2011; Elmore et al., 2011; Heyselaar et al., 2011; Lara & Wallis, 2012), we found a decrease in performance with set size. Following Keshvari et al. (2012, 2013), Lara and Wallis (2012), and van den Berg et al. (2012), we systematically varied change magnitude to obtain a richer description of behavior, which we exploited to distinguish models that otherwise could not be distinguished.

Although change detection and change localization are classic paradigms in humans, formally comparing models on data from these paradigms is relatively new (van den Berg et al., 2012; Wilken & Ma, 2004), and no previous study has compared models in parallel across species. We tested the item-limit (IL) model, in which there is a fixed limit on the number of items that can be remembered and items are stored in an all-or-none fashion, as well as noise-based (or resource) models, in which items are encoded in VWM in a noisy way. The data from both species were well accounted for by a noise-based model in which memory precision is variable across items and trials (VP), but not by a noise-based model in which memory precision is equal across items and trials (EP), and not by the classic IL model. These findings are consistent with earlier ones in humans (Fougnie et al., 2012; Keshvari et al., 2012, 2013; van den Berg et al., 2012; van den Berg et al., 2014; van den Berg & Ma, 2014).

We also tested hybrid models that combine the concepts of noisy storage and an item limit. Adding an item limit to the EP model (as has been proposed by Zhang & Luck, 2008, and Anderson, Vogel, & Awh, 2011) helped, but not enough to make it fit as well as the VP model. Adding an item limit to the VP model did improve the fit, but not enough to convincingly exceed the penalty associated with adding an extra parameter to the model. Thus our model comparison neither yields any evidence for the existence of an item limit nor rules it out. This conclusion is consistent with a recent detailed model comparison on multiple data sets obtained using a delayed-estimation paradigm (van den Berg et al., 2014).

The success of the VP model brings to the fore the question of its mechanistic underpinnings. The essential components of the model are noisy storage, a decrease of average precision with increasing set size, and variability of precision across items and trials around this average. At the neural level, noisy storage could take the form of a Poisson-like neural population responding to the stimulus, in which case precision might correspond to either the gain or the total spike count in this population (Ma et al., 2014; van den Berg et al., 2012). A decrease of gain with set size has been observed in area LIP (Churchland, Kiani, & Shadlen, 2008) and superior colliculus (Basso & Wurtz, 1998), and might be implemented using divisive normalization (Bays, 2014; Ma & Huang, 2009). A Poisson-like population with gain fixed across items and trials (i.e., at a given set size) might already behave like a VP model (Bays, 2014). In addition, gain itself might be variable (Goris, Movshon, & Simoncelli, 2014), for example due to fluctuations in attention (Cohen & Maunsell, 2010) or to variability in memory decay rates (Fougnie et al., 2012). Other factors are also expected to contribute to fluctuations in precision, such as eye movements, and stimulus-related differences such as those due to cardinal orientations (Girshick, Landy, & Simoncelli, 2011) and configural grouping (Brady & Tenenbaum, 2013). Thus, although much more work is needed, the VP model is currently supported by a range of behavioral, physiological, and computational studies.

In the field of comparative cognition, much research has been devoted to comparing absolute performance differences across various species—including pigeons,

rats, rhesus monkeys, baboons, and humans—on attention, visual search, spatial navigation, and categorization tasks (Wasserman & Zentall, 2009). However, in order to disentangle whether these performance differences are due to qualitative differences in the underlying mechanisms (differences in models) or simply quantitative in nature (differences in parameters), formal model comparison is needed. Here we have shown that despite interspecies performance differences, the same model fitted the data from both species best. This suggests qualitative similarity and evolutionary continuity of basic VWM mechanisms. This qualitative similarity supports the use of rhesus monkeys as a model system for studying the neural mechanisms of multiple-item VWM.

Keywords: visual working memory, visual short-term memory, change detection, change localization, capacity, precision, computational modeling, nonhuman primates

Acknowledgments

This research was supported by NIH grants MH091038 and MH072616 to AAW and NIH grant R01EY020958-01 and ARO grant W911NF-12-1-0262 to WJM. We thank John Magnotti for useful discussions and assistance with programming.

Commercial relationships: none.
Corresponding author: Deepna T. Devkar.
Email: deepna.devkar@nyu.edu.
Address: Department of Psychology, New York University, New York, NY, USA.

References

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

Anderson, D. E., Vogel, E. K., & Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. *The Journal of Neuroscience*, *31*(3), 1128–1138.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*(7), 622–628.

Basso, M. A., & Wurtz, R. H. (1998). Modulation of neuronal activity in superior colliculus by changes in target probability. *The Journal of Neuroscience*, *18*(18), 7519–7534.

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

Bays, P. M., & Husain, M. (2008, Aug 8). 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., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences, USA*, *108*(27), 11252–11255.

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*(6), 693–702.

Cohen, M. R., & Maunsell, J. H. (2010). A neuronal population measure of attention predicts behavioral performance on individual trials. *The 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. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Efron, B. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Elmore, L. C., Ma, W. J., Magnotti, J. F., Leising, K. J., Passaro, A. D., Katz, J. S., & Wright, A. A. (2011). Visual short-term memory compared in rhesus monkeys and humans. *Current Biology*, *21*(11), 975–979.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229–1242.

Fukuda, K., Awh, E., & Vogel, E. K. (2010). Discrete capacity limits in visual working memory. *Current Opinion in Neurobiology*, *20*(2), 177–182.

Funahashi, S., Bruce, C. J., & Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349.

Fuster, J. M., & Alexander, G. E. (1971, Aug 13).

Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654.

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

Goris, R. L. T., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, *17*(6), 858–865.

Heyselaar, E., Johnston, K., & Paré, 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]

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

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.

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

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

Lara, A. H., & Wallis, J. D. (2014). Executive control processes underlying multi-item working memory. *Nature Neuroscience*, *17*(6), 876–883.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400.

Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision*, *9*(11):3, 1–30, doi:10.1167/9.11.3. [PubMed] [Article]

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.

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

Mardia, K. V., & Jupp, P. E. (1999). *Directional statistics*. London: John Wiley and Sons.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *The Journal of Neuroscience*, *16*(16), 5154–5167.

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

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

Schwarz, G. (1978). Estimating the dimension of a model. *The 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*(22), 10749–10753.

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

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

van den Berg, R., & Ma, W. (2014). "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.

Warden, M. R., & Miller, E. K. (2007). The representation of multiple objects in prefrontal neuronal delay activity. *Cerebral Cortex, 17* (suppl 1), i41–i50.

Wasserman, E., & Zentall, T. (2009). *Comparative cognition: Experimental explorations of animal intelligence*. Oxford, UK: Oxford University Press.

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]

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

Appendix A: Derivation of the decision rule

Step 1: Generative model

Figure A1 shows the relevant variables: the location L of the change (1 or 2), the magnitude Δ of the change, the relevant sample orientations θ_1 and θ_2 (all other sample items 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 $p(L) = 0.5$.
- The change magnitude Δ and each of the sample orientations have discrete distributions, but we approximate them by uniform distributions $p(\Delta) = \frac{1}{2\pi}$ and $p(\theta_1, \theta_2) = \left(\frac{1}{2\pi}\right)^2$. We chose continuous uniform distributions rather than discrete distributions at the 18 presented orientations (or change magnitudes) because we think that it is unlikely that an observer learns those exact orientations (or change magnitudes); the choice of continuous uniform distributions also allows for a closed form for the decision rule.
- 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 (Equation 1).
- 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 $\varphi_2 = \theta_2 + \Delta$. We can formally denote this by $(\varphi_1, \varphi_2) = (\theta_1, \theta_2) + \Delta\mathbf{1}_L$, where $\mathbf{1}_L$ is equal to (1, 0) when $L = 1$ and (0, 1) when $L = 2$.

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; we also assume that the observer knows the values of κ_1 and κ_2. An ideal observer infers L by computing the posterior distribution over L, $p(L|x_1, x_2, \varphi_1, \varphi_2)$. Since L is binary, all

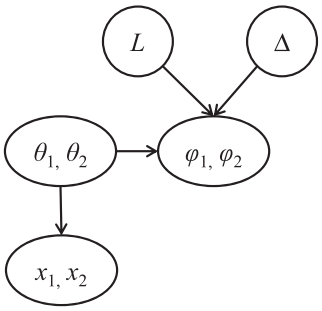

Figure A1. Graphical depiction of the generative model on which the decision rule is based.

information about the posterior is contained in the log posterior ratio, which can be rewritten using Bayes's rule:

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

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):

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

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

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

Combining, we find the log posterior ratio

$$\log\frac{p(L=1|x_1,x_2,\varphi_1,\varphi_2)}{p(L=2|x_1,x_2,\varphi_1,\varphi_2)}$$

		Monkeys									Humans				
		Tested range			M1		M2		M3		Tested range				
Model	Parameter	Min	Step	Max	Mean	*SEM*	Mean	*SEM*	Mean	*SEM*	Min	Step	Max	Mean	*SEM*
IL	K	1	1	5	1	0	1	0	1	0	1	1	5	1.50	0.17
	ε	0	0.003	3	0.248	0.0055	0.269	0.0062	0.249	0.0065	0	0.003	3	0.079	0.011
EP	$J_{N=1}$	0	0.13	25	3.51	0.78	2.41	0.68	2.71	0.69	0	0.13	25	17.7	2.3
	α	0	0.015	3	2.35	0.21	1.98	0.27	1.98	0.26	0	0.015	3	2.07	0.12
EPF	K	1	1	5	1	0	1.16	0.79	1.12	0.62	1	1	5	2.20	0.36
	$J_{N=1}$	0	0.13	25	1.23	0.085	1.12	0.16	1.35	0.31	0	0.13	25	15.3	2.5
	α	0	0.015	3	1.35	0.60	1.57	0.74	1.89	0.83	0	0.015	3	1.43	0.25
VP	$\bar{J}_{N=1}$	0	0.30	30	11.0	1.8	3.82	0.87	7.0	1.8	0	1.01	100	65.8	8.7
	τ	0.1	0.40	30	24.9	4.4	6.2	2.6	15.7	5.9	0.1	1.11	100	29.3	8.5
	α	0	0.03	3	1.47	0.14	1.32	0.14	1.31	0.13	0	0.03	3	1.82	0.13
VPF	$\bar{J}_{N=1}$	0	0.30	30	10.2	2.7	3.7	1.4	7.7	2.7	0	2.02	200	83.0	17.9
	τ	0.1	0.40	30	23.8	5.1	5.6	2.8	13.9	5.2	0.1	2.13	200	25.3	5.4
	α	0	0.03	3	1.55	0.49	1.47	0.48	1.5	0.41	0	0.061	3	1.97	0.18
	K	1	1	5	3.6	1.8	3.4	1.5	3.2	1.1	1	1	5	4.10	0.28

Table A1. Parameter ranges and parameter estimates. *Notes*: For monkeys, means and standard errors of the mean (*SEM*s) were estimated from 100 bootstrapped data sets. For humans, means and standard errors were computed across subjects.

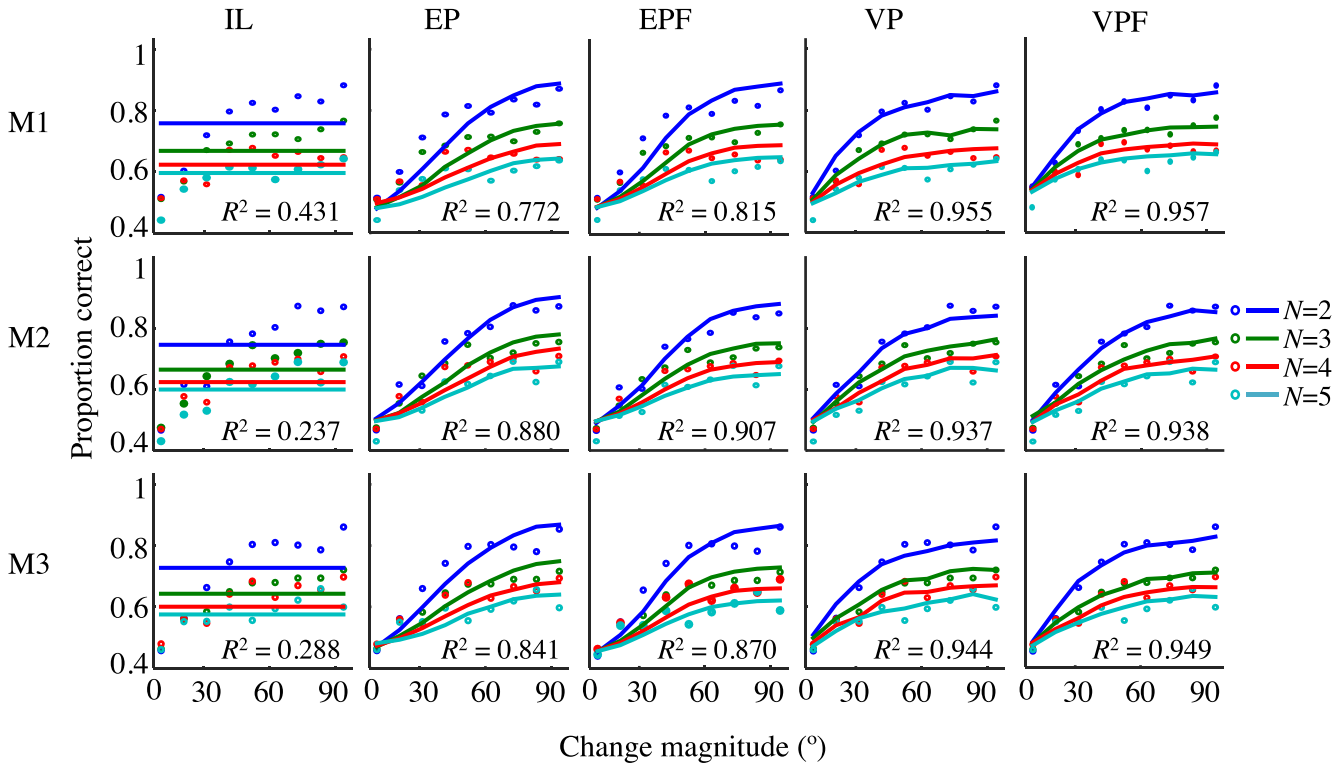

Figure A2. Model fits. Proportion correct as a function of set size N and change magnitude, for monkeys M1, M2, and M3 across nonbootstrapped data sets. Circles are behavior; solid lines are model fits.

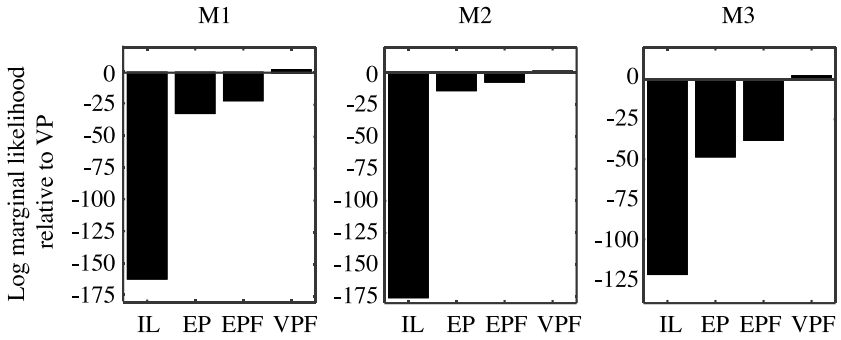

Figure A3. Bayesian model comparison for monkeys M1, M2, and M3, showing the log marginal likelihood of each model minus that of the VP model. The VP and VPF models account about equally well for the data, and better than the other models.

$$= \log \frac{\frac{1}{2\pi}\frac{1}{2\pi I_0(\kappa_2)} e^{\kappa_2 \cos((x_2-\varphi_2)}}{\frac{1}{2\pi}\frac{1}{2\pi I_0(\kappa_1)} e^{\kappa_1 \cos(x_1-\varphi_1)}}$$

$$= \log \frac{I_0(\kappa_1)}{I_0(\kappa_2)} + \kappa_2 \cos(x_2 - \varphi_2) - \kappa_1 \cos(x_1 - \varphi_1).$$

The ideal observer responds that the change occurred at location 1 when the log posterior ratio is positive:

$$\log \frac{I_0(\kappa_1)}{I_0(\kappa_2)} + \kappa_2 \cos(x_2 - \varphi_2) - \kappa_1 \cos(x_1 - \varphi_1) > 0.$$

This is Equation 3 in the main text.

Appendix B: Parameter estimates

Table A1 shows our approximations to the maximum-likelihood estimates of all parameters in all models in all subjects (but averaged over human subjects).

Appendix C: Model fits to nonbootstrapped monkey data sets

Figures A2 and A3 show model fits and Bayesian model comparison on the nonbootstrapped data sets for each individual monkey. Our results are consistent with those on bootstrapped datasets.