PLOS | COMPUTATIONAL BIOLOGY

# No Evidence for an Item Limit in Change Detection

**Shaiyan Keshvari**[⌐¤], **Ronald van den Berg**[⌐], **Wei Ji Ma***

Department of Neuroscience, Baylor College of Medicine, Houston, Texas, United States of America

## Abstract

Change detection is a classic paradigm that has been used for decades to argue that working memory can hold no more than a fixed number of items ("item-limit models"). Recent findings force us to consider the alternative view that working memory is limited by the precision in stimulus encoding, with mean precision decreasing with increasing set size ("continuous-resource models"). Most previous studies that used the change detection paradigm have ignored effects of limited encoding precision by using highly discriminable stimuli and only large changes. We conducted two change detection experiments (orientation and color) in which change magnitudes were drawn from a wide range, including small changes. In a rigorous comparison of five models, we found no evidence of an item limit. Instead, human change detection performance was best explained by a continuous-resource model in which encoding precision is variable across items and trials even at a given set size. This model accounts for comparison errors in a principled, probabilistic manner. Our findings sharply challenge the theoretical basis for most neural studies of working memory capacity.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: wjma@bcm.edu

¤ Current address: Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America.

⌐ These authors contributed equally to this work.

## Introduction

Visual working memory, the ability to buffer visual information over time intervals of the order of seconds, is a fundamental aspect of cognition. It is essential for detecting changes [1–3], integrating information across eye fixations [4–5], and planning goal-directed reaching movements [6]. Numerous studies have found that visual working memory is limited, but the precise nature of its limitations is subject of intense debate [7–14]. The standard view is that visual working memory cannot hold more than about four items, with any excess items being discarded [7–9,15–18]. According to an alternative hypothesis, working memory limitations take the form of a gradual decrease in the encoding precision of stimuli with increasing set size [10–11,13,19–23]. In this view, encoding precision is a continuous quantity, and this hypothesis has therefore also been referred to as the continuous-resource hypothesis.

Historically, the leading paradigm for studying visual working memory has been change detection, a task in which observers report whether a change occurred between two scenes separated in time [2–3,24]. Not only humans, but also non-human primates can perform multiple-item change detection [25–28], and physiological studies have begun to investigate the neural mechanisms involved in this task [27]. Findings from change detection studies have been used widely to argue in favor of the item-limit hypothesis [2,8,15–18]. The majority of these studies, however, used stimuli that differed categorically from each other, such as line drawings of everyday objects or highly distinct and easily named colors. The logic is that for such stimuli, changes are

large relative to the noise, avoiding the problem of "comparison errors" [1,18,29–30] that would be associated with low encoding precision (high noise). When encoding precision is limited, an observer's stimulus measurements are noisy and will differ between displays for each item, even if the item did not change. The observer then has to decide whether a difference in measurements is due to noise only or to a change plus noise, which is especially problematic when changes are small. This signal detection problem results in comparison errors.

Attempts to avoid such errors by using categorical stimuli run into two objections: first, using such stimuli does not guarantee that comparison errors are absent and can be ignored in modeling; second, there is no good reason to avoid comparison errors, since the pattern of such errors can help to distinguish models. Ideally, change detection performance should be measured across a wide range of change magnitudes, including small values, as we do here. Comparison errors can, in fact, be modeled rather easily within the context of a Bayesian-observer model. Bayesian inference is the decision strategy that maximizes an observer's accuracy given noisy measurements [31–32], and was recently found to describe human decision-making in change detection well [33].

We conducted two change detection experiments, in the orientation and color domains, in which we varied both set size and the magnitude of change. We rigorously tested five models of working memory limitations, each consisting of an encoding stage and a decision stage. The encoding stage differed between the five models: the original item-limit model [2,15–16], two recent variants [9], and two continuous-resource models, one with equal

## Author Summary

Working memory is a fundamental aspect of human cognition. It allows us to remember bits of information over short periods of time and make split-second decisions about what to do next. Working memory is often tested using a change detection task: subjects report whether a change occurred between two subsequent visual images that both contain multiple objects (items). The more items are present in the images, the worse they do. The precise origin of this phenomenon is not agreed on. The classic theory asserts that working memory consists of a small number of slots, each of which can store one item; when there are more items than slots, the extra items are discarded. A modern model postulates that working memory is fundamentally limited in the quality rather than the quantity of memories. In a metaphor: instead of watering only a few plants in our garden, we water all of them, but the more plants we have, the less water each will receive on average. We show that this new model does much better in accounting for human change detection responses. This has consequences for the entire field of working memory research.

precision for all items [20,23], and one with item-to-item and trial-to-trial variability in precision [13,33]. The decision stage was Bayesian for every model. To anticipate our results, we find that variable precision coupled with Bayesian inference provides a highly accurate account of human working memory performance across change magnitudes, set sizes, and feature dimensions, and far outperforms models that postulate an item limit.

## Results

### Theory

We model a task in which the observer is presented with two displays, each containing $N$ oriented stimuli and separated in time by a delay period. On each trial, there is a 50% probability that one stimulus changes orientation between the first and the second display. The change can be of any magnitude. Observers report whether or not a change occurred. We tested five models of this task, which differ in the way they conceptualize what memory resource consists of and how it is distributed across items (Fig. 1a).

**Infinite-precision item-limit model.** In the infinite-precision (IP) item-limit model, the oldest item-limit model [2,8,15–16] and often called the "limited-capacity" or simply the "item-limit" model, memorized items are stored in one of $K$ available "slots". $K$ is called the capacity. Each slot can hold exactly one item. The memory of a stored item is perfect ("infinite precision"). If $N \leq K$, all items from the first display are stored. If $N > K$, the observer memorizes $K$ randomly chosen items from the first display. When a change occurs among the memorized items, the observer responds "change" with probability $1 - \varepsilon$. When no change occurs among the memorized items, the observer responds "change" with a guessing probability $g$.

**Precision and noise.** All models other than the IP model assume that the observer's measurement of each stimulus is corrupted by noise. We model the measurement $x$ of a stimulus $\theta$ as being drawn from a Von Mises (circular normal) distribution centered at $\theta$:

$$p(x|\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(x-\theta)}, \quad (1)$$

where $\kappa$ is called the concentration parameter and $I_0$ is the modified Bessel function of the first kind of order 0. (For convenience, we remap all orientations from $[-\pi/2, \pi/2]$ to $[-\pi, \pi]$.)

In all models with measurement noise, we identify memory resource with Fisher information, $J(\theta)$ [34]. The reasons for this choice are threefold [13]. First, regardless of the functional form of the distribution of the internal representation of a stimulus (in our formalism, of the scalar measurement), Fisher information determines the best possible performance of any estimator through the Cramér-Rao bound [34], of which a version on a circular space exists [35]. Second, when the measurement distribution is Gaussian, Fisher information is equal to the inverse variance, $J = \frac{1}{\sigma^2}$, which is, up to an irrelevant proportionality constant, the same relationship one would obtain by regarding resource as a collection of discrete observations or samples [20,23]. Third, when neural variability is Poisson-like, Fisher information is proportional to the gain of the neural population [36–38], and therefore the choice of Fisher information is consistent with regarding neural activity as resource [13]. We will routinely refer to Fisher information as precision. For the circular measurement distribution in Eq. (1), Fisher information is related to $\kappa$ through $J = \kappa \frac{I_1(\kappa)}{I_0(\kappa)}$ [13,33], where $I_1(\kappa)$ is the modified Bessel function of the first kind of order 1.

**Slots-plus-averaging model.** The SA model [9] is an item-limit model in which $K$ discrete, indivisible chunks of resource are allocated to items. When $N > K$, $K$ randomly chosen items receive a chunk and are encoded; the remaining $N - K$ items are not memorized. When $N \leq K$, chunks are distributed as evenly as possible over all items. For example, if $K = 4$ and $N = 3$, two items receive one chunk and one receives two. Resource per item, $J$, is proportional to the number of chunks allocated to it, denoted $S$: $J = S J_s$, where $J_s$ is the Fisher information corresponding to one chunk.

**Slots-plus-resources model.** The slots-plus-resources (SR) model [9] is identical to the SA model, except that resource does not come in discrete chunks but is a continuous quantity. When $N \leq K$, all items are encoded with precision $J = J_1/N$, where $J_1$ is the Fisher information for a single item. When $N > K$, $K$ randomly chosen items are encoded with precision $J = J_1/K$ and the remaining $N - K$ items are not memorized. Related but less quantitative ideas have been proposed by Alvarez and Cavanagh [14] and by Awh and colleagues [7,18].

**Equal-precision model.** According to the equal-precision (EP) model [10–11,20,23], precision is a continuous quantity that is equally divided over all items. Versions of this model have been tested before on change detection data [8,10,39]. If the total amount of memory precision were fixed across trials, we would expect an inverse proportionality between $J$ and set size. However, there is no strong justification for this assumption, we allow for a more flexible relationship by using a power-law function, $J = J_1 N^\alpha$.

**Variable-precision model.** In the variable-precision (VP) model [13], encoding precision is variable across items and trials, and average encoding precision depends on set size. We model variability in precision by drawing $J$ from a gamma distribution with mean $\bar{J}$ and scale parameter $\tau$ (Fig. 1b). The gamma distribution is a flexible, two-parameter family of distributions on the positive real line. The process by which a measurement $x$ is generated in the VP model is thus doubly stochastic: $x$ is drawn randomly from a Von Mises distribution with a given precision, while precision itself is stochastic. Analogous to $J$ in the EP model, we model the relationship between $\bar{J}$ and set size using a power law function, $\bar{J} = \bar{J}_1 N^\alpha$.
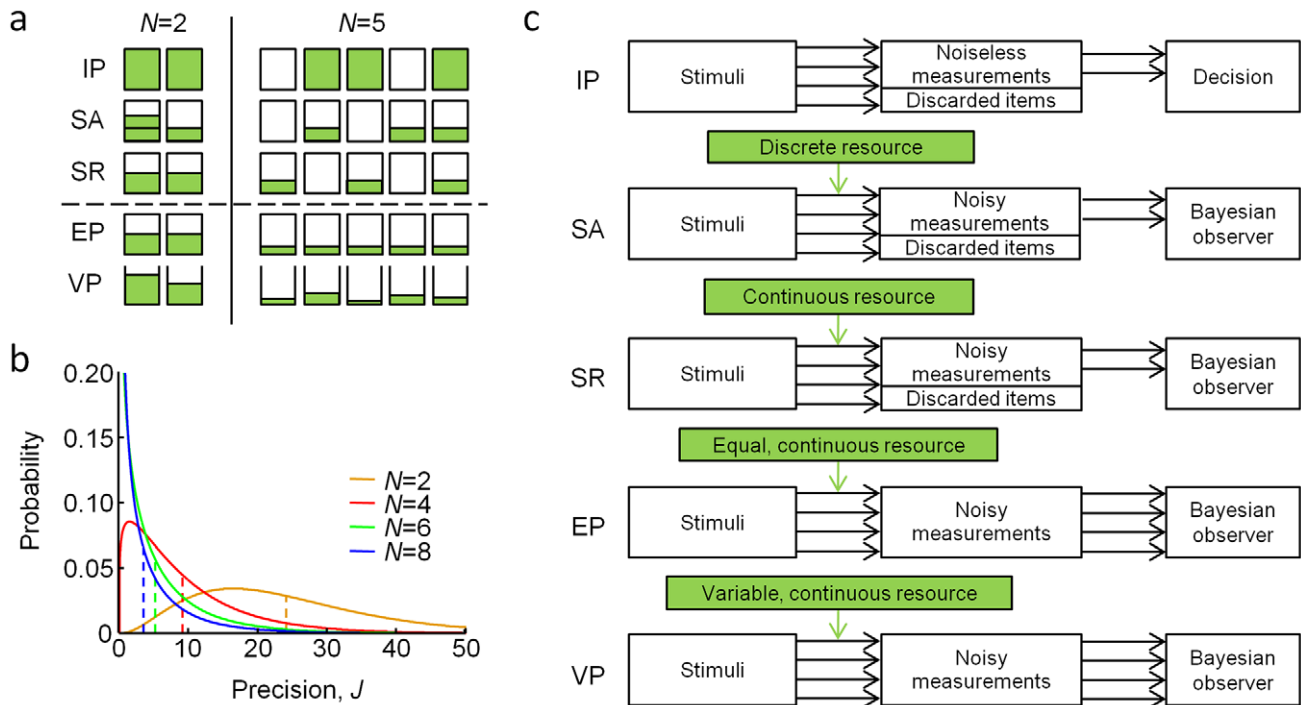
**Figure 1. Models of change detection.** Infinite-precision item limit (IP), slots plus averaging (SA), slots plus resources (SR), equal precision (EP), and variable precision (VP). The first three are item-limit models, the last two continuous-resource models. (**a**) Illustration of resource allocation in the models at set sizes 2 and 5, with a capacity of 3 slots/chunks for IP, SA, and SR. The VP model is distinct from the other models in that the amount of resource varies on a continuum without a hard upper bound. (**b**) Probability density functions over encoding precision in the VP model, for four set sizes. Parameters were taken from the best fit to the data of one human subject. Mean precision, indicated by a dashed line, is inversely proportional to set size. In the EP model, these distributions would be infinitely sharp (delta functions). (**c**) Decision process during change detection for each of the five models.
doi:10.1371/journal.pcbi.1002927.g001

**Bayesian inference.** In the models with noise (SA, SR, EP, VP), the observer decides whether or not a change occurred (denoted by $C = 1$ and $C = 0$) based on the noisy measurements in both displays (Fig. 1c). We use $x_i$ and $y_i$ to denote the noisy measurements at the $i$th location in the first and second displays, and $\kappa_{x,i}$ and $\kappa_{y,i}$ are their respective concentration parameters (see Eq. (1)). Due to the noise, the measurements of any one item will always differ between displays, even if the underlying stimulus value remains unchanged. Thus, also on no-change trials, the observer is confronted with two non-identical sets of measurements, making the inference problem difficult. While the noise precludes perfect performance, the observer still has a best possible strategy available, namely Bayesian MAP estimation. This strategy consists of computing, on each trial, the probability of a change based on the measurements, $p(C = 1 | \mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are the vectors of measurements $\{x_i\}$ and $\{y_i\}$, respectively. The observer then responds "change" if this probability exceeds 0.5, or in other words, when

$$d \equiv \frac{p(C = 1 | \mathbf{x}, \mathbf{y})}{p(C = 0 | \mathbf{x}, \mathbf{y})} > 1.$$

Making use of the statistical structure of the task (Fig. S1), the posterior ratio $d$ can be evaluated to

$$d = \frac{p_{\text{change}}}{1 - p_{\text{change}}} \frac{1}{N} \sum_{i=1}^{N} \frac{I_0(\kappa_{x,i}) I_0(\kappa_{y,i})}{I_0\left(\sqrt{\kappa_{x,i}^2 + \kappa_{y,i}^2 + 2\kappa_{x,i}\kappa_{y,i}\cos(y_i - x_i)}\right)}. \quad (2)$$

(see Text S1 and [40]). Here, $p_{\text{change}}$ is the prior probability that a change occurred. This decision rule automatically models errors arising in the comparison operation [1,18,29–30]: the difference $y_i - x_i$ is noisy, so that even when a change is absent, it might by chance be large, and even when a change is present, it might by chance be small.

In an earlier paper [40], we examined suboptimal alternative decision rules. A plausible one would be a "threshold" rule, according to which the observer compares the largest difference between measurements at the same location in the two displays to a fixed criterion. If the difference exceeds the criterion, the observer reports that a change occurred. We proposed this "maximum-absolute-difference" rule in our earlier continuous-resource treatment of change detection [10], but a comparison against the optimal rule showed it to be inadequate [40].

Another suboptimal strategy that deserves attention is probability matching or sampling [41–42]. Under this strategy, the observer computes the Bayesian posterior $p(C = 1 | \mathbf{x}, \mathbf{y})$, but instead of reporting a change ($\hat{C} = 1$) when this probability exceeds 0.5, reports a change with probability

$$p\left(\hat{C} = 1 | \mathbf{x}, \mathbf{y}\right) = \frac{p(C = 1 | \mathbf{x}, \mathbf{y})^k}{p(C = 0 | \mathbf{x}, \mathbf{y})^k + p(C = 1 | \mathbf{x}, \mathbf{y})^k}. \quad (3)$$

When $k = 0$, probability matching amounts to random guessing; when $k \to \infty$, it reduces to MAP estimation. Thus, probability matching consists of a family of stochastic decision rules interpolating between MAP estimation and guessing. Probability

matching turns out to be very similar to a modification of MAP estimation we considered in [40], namely adding zero-mean Gaussian noise to the logarithm of the decision variable in Eq. (2). To see this, we rewrite Eq. (3) as

$$p\left(\hat{C}=1|\mathbf{x},\mathbf{y}\right)=\frac{1}{1+\dfrac{p(C=0|\mathbf{x},\mathbf{y})^{k}}{p(C=1|\mathbf{x},\mathbf{y})^{k}}}=\frac{1}{1+e^{-k\log d}},$$

which is the logistic function with argument log $d$. On the other hand, adding zero-mean Gaussian noise $\eta$ with standard deviation $\sigma_{\eta}$ to log $d$ gives

$$p\left(\hat{C}=1|\mathbf{x},\mathbf{y}\right)=\Pr(\log d+\eta>0)=\Phi\left(\frac{\log d}{\sigma_{\eta}}\right),$$

where $\Phi$ is the cumulative of the standard normal distribution. It is easy to verify that the logistic function and the cumulative normal distribution are close approximations of each other (with a one-to-one relation between $k$ and $\sigma_{\eta}$), showing that both forms of suboptimality are very similar. Since an equal-precision model augmented with Gaussian decision noise far underperformed the variable-precision model [40], human data are unlikely to be explained by such decision noise (or equivalently by probability matching) in the absence of variable precision in the encoding stage. It is, however, possible that decision noise is present in addition to variability in encoding precision, but this would not invalidate our conclusions. Therefore, in the present paper, we will only examine the optimal Bayesian decision rule.

**Free parameters.** The IP, SA, SR, and EP models each have 3 free parameters, and the VP model has 4.

## Experiment: orientation change detection

We conducted an orientation change detection task in which we manipulated both set size and change magnitude (Fig. 2a). Consistent with earlier studies (e.g. [10,15,17]), we found that the ability of observers to detect a change decreased with set size, with hit rate $H$ monotonically decreasing and false-alarm rate $F$ monotonically increasing (Fig. 2b). Effects of set size were significant (repeated-measures ANOVA; hit rate: $F(3,27)=52.8$, $p<0.001$; false alarm rate: $F(3,27)=82.0$, $p<0.001$). The increase

in $F$ is inconsistent with the IP model, as this model would predict no dependence.

For a more detailed representation of the data, we binned magnitude of change on change trials into 10 bins (Fig. 2c). All no-change trials had magnitude 0 and sat in a separate bin. These psychometric curves clearly show that the probability of reporting a change increases with change magnitude at every set size ($p<0.001$). From Fig. 2c we could, in principle, compute a naïve estimate of memory capacity using the well-known formula from the IP model, $K=\mathcal{N}(H-F)/(1-F)$ [16]. However, since $H$ depends on the magnitude of change, the estimated $K$ would depend on the magnitude of change as well, contradicting the basic premise of a fixed capacity. For example, at set size 6, for change magnitudes between $0°$ and $9°$, Cowan's formula would estimate $K$ at exactly zero (no items retained at all), while for magnitudes between $81°$ and $90°$, it would estimate $K$ at 3.8, with a nearly linear increase in between. This serves as a first indication that the IP model in general and this formula in particular are wrong.

**Model fits.** We fitted all models using maximum-likelihood estimation, for each subject separately (see Text S1). Mean and standard error of all parameters of all models are shown in Table 1. The values of capacity $K$ in the IP, SA, and SR models were $3.10±0.28$, $4.30±0.47$, and $4.30±0.42$, respectively (mean and s.e.m.), in line with earlier studies [7–9,15–18]. Using the maximum-likelihood estimates of the parameters, we obtained hit rates, false-alarm rates, and psychometric curves for each model and each subject (Fig. 3).

Hit and false-alarm rates were best described by the VP model, per root-mean-square error (RMSE) of the subject means (0.040), followed by the SA and SR models (both 0.046), the equal-precision (EP) model (0.059), and the IP model (0.070). The same order was found for the psychometric curves (RMSE: 0.10 for VP, 0.11 for SA, 0.12 for SR, 0.13 for EP, and 0.21 for IP). The IP model predicts that performance is independent of magnitude of change and is therefore easy to rule out.

**Bayesian model comparison.** The RMS errors reported so far are rather arbitrary descriptive statistics. To compare the models in a more principled (though less visualizable) fashion, we performed Bayesian model comparison, also called Bayes factors [43–44] (see Text S1). This method returns the likelihood of each model given the data and has three desirable properties: it uses all data instead of only a subset (like cross-validation would) or
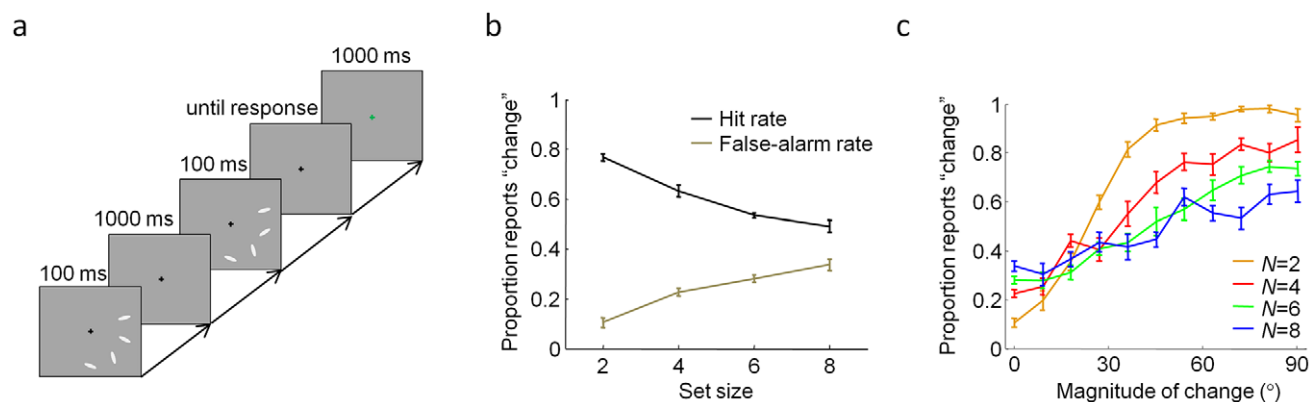


**Figure 2. Orientation change detection.** (**a**) Observers reported whether one of the orientations changed between the first and second displays. (**b**) Hit and false-alarm rates as a function of set size. (**c**) Psychometric curves, showing the proportion of "change" reports as a function of the magnitude of change, for each set size (mean ± s.e.m across subjects). Magnitude of change was binned into 9° bins. The first point on each curve (at 0°) contains all trials in which no change occurred, and thus represents the false-alarm rate. Using the standard formula for $K$ would return different estimates for different change magnitudes.
doi:10.1371/journal.pcbi.1002927.g002

**Table 1.** Fitted parameter ranges and estimates.

| Model | Parameter | Experiment 1 estimates | | Tested range | |
| | | Mean | s.e.m. | Min | Max |
| --- | --- | --- | --- | --- | --- |
| IP | $K$ | 3.10 | 0.28 | 1 | 8 |
| | $\varepsilon$ | 0.220 | 0.021 | 0 | 1 |
| | $G$ | 0.247 | 0.016 | 0 | 0.5 |
| SA | $J_s$ | 3.94 | 0.58 | 1 | 40 |
| | $K$ | 4.30 | 0.47 | 1 | 8 |
| | $p_{change}$ | 0.584 | 0.020 | 0.2 | 0.8 |
| SR | $J_1$ | 14.2 | 1.7 | 1 | 60 |
| | $K$ | 4.30 | 0.42 | 1 | 8 |
| | $p_{change}$ | 0.574 | 0.019 | 0.2 | 0.8 |
| EP | $J_1$ | 20.3 | 3.6 | 1 | 60 |
| | $\alpha$ | −1.28 | 0.11 | −2 | 0 |
| | $p_{change}$ | 0.492 | 0.007 | 0.2 | 0.8 |
| VP | $\bar{J}_1$ | 53.1 | 6.1 | 5 | 300 |
| | $\tau$ | 31.2 | 8.9 | 5 | 300 |
| | $\alpha$ | −0.88 | 0.08 | −2 | 0 |
| | $p_{change}$ | 0.532 | 0.005 | 0.2 | 0.8 |

Mean and standard error of the maximum-likelihood estimates and tested ranges of model parameters for Experiment 1 (orientation change detection).
doi:10.1371/journal.pcbi.1002927.t001

summary; it does not solely rely on point estimates of the parameters but integrates over parameter space, thereby accounting for the model's robustness against variations in the parameters; it automatically incorporates a correction for the number of free parameters. We found that the log likelihood of the VP model exceeds that of the IP, SA, SR, and EP models by $97\pm11$, $7.2\pm3.5$, $7.4\pm3.7$, and $19\pm3$, respectively (Fig. 4). This constitutes strong evidence in favor of the VP model, for example according to Jeffreys' scale [45]. Based on our data, we can convincingly rule out the three item-limit models (IP, SA, and SR) as well as the equal-precision (EP) model, as descriptions of human change detection behavior.

**Apparent guessing as an epiphenomenon.** In the delayed-estimation paradigm of working memory [10], data consist of subject's estimates of a memorized stimulus on a continuous space. Zhang and Luck [9] analyzed the histograms of estimation errors in this task by fitting a mixture of a uniform distribution (allegedly representing guesses) and a Von Mises distribution (allegedly representing true estimates of the target stimulus). They suggested that the mixture proportion of the uniform distribution represents the rate at which subjects guess randomly, and interpreted its increase with set size as evidence for a fixed limit on the number of remembered items. However, Van den Berg et al. [13] later showed that the variable-precision model reproduces the increase of the mixture proportion of the uniform distribution with set size well, even though the model does not contain any pure guessing. They suggested that the guesses reported in the mixture analysis were merely "apparent guesses".

We perform an analogous analysis for change detection here. We fitted, at each set size separately, a model in which subjects guess on a certain proportion of trials, and on other trials, respond like an EP observer. Free parameters, at each set size separately, are the guessing parameter, which we call apparent guessing rate (AGR), and the precision parameter of the EP observer. We found that AGR was significantly different from zero at every set size ($t(9)>4.5$, $p<0.001$) and increased with set size (Fig. 5; repeated-measures ANOVA, main effect of set size: $F(3,27)=21.1$, $p<0.001$), reaching as much as $0.60\pm0.06$ at set size 8.

We then examined how well each of our five models can reproduce the increase of AGR. To do so, we computed AGR from synthetic data generated using each model, using maximum-likelihood estimates of the parameters as obtained from the subjects' data. We found that the VP model – which does not
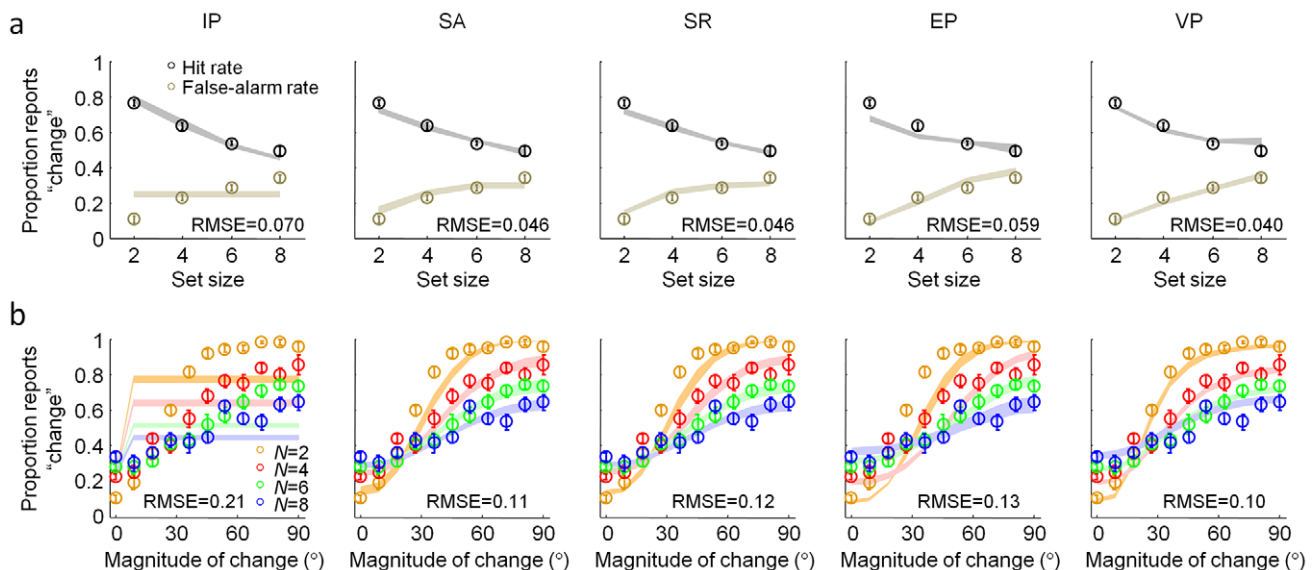


**Figure 3. Comparing models on summary statistics.** (**a**) Model fits to the hit and false-alarm rates. (**b**) Model fits to the psychometric curves. Shaded areas represent ±1 s.e.m. in the model. For the IL model, a change of magnitude 0 has a separate proportion reports "change", equal to the false-alarm rate shown in (a). In each plot, the root mean square error between the means of data and model is given.
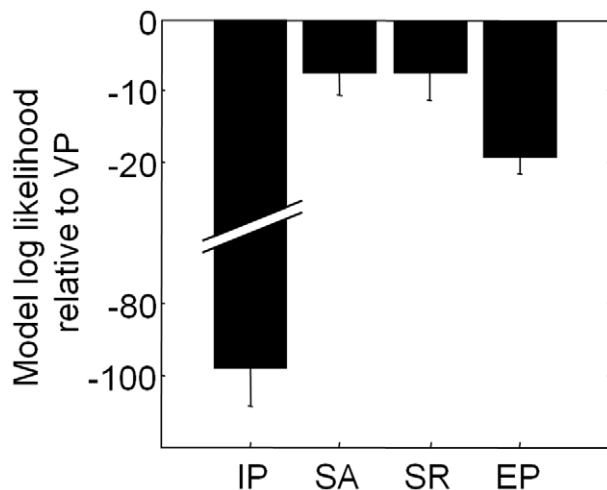doi:10.1371/journal.pcbi.1002927.g003

**Figure 4. Bayesian model comparison.** Model log likelihood of each model minus that of the VP model (mean ± s.e.m.). A value of $-x$ means that the data are $e^x$ times more probable under the VP model.
doi:10.1371/journal.pcbi.1002927.g004

contain any actual guessing – reproduces the apparent guessing rate better than the other models (Fig. 5; RMSE = 0.20 for VP). This means that the apparent presence of guessing does not imply that visual working memory is item-limited.

How the VP model can reproduce apparent guessing can be understood as follows. In the VP model, the distribution of precision is typically broad and includes a lot of small values, especially at larger set sizes (Fig. 1b). The EP model augmented with set size-dependent guessing would approximate this broad distribution by one consisting of two spikes of probability, one at a nonzero, fixed precision and one at zero precision. To mimic the VP precision distribution, the weight of the spike at zero must increase with set size, leading to an increase of AGR with set size. In sum, variability in precision produces apparent guessing as an epiphenomenon, a finding that is consistent with our results in the delayed-estimation task [13].

**Generalization.** To assess the generality of our results, we repeated the orientation change detection experiment with color stimuli and found consistent results (see Figs. S2, S3, S4, S5 and Text S1). Specifically, in Bayesian model comparison, the VP model outperforms all other models by log likelihood differences of at least 48.4±8.2, which constitutes further evidence against an item limit.

## Discussion

### Implications for working memory

Five models of visual working memory limitations have been proposed in the literature. Here, we tested all five using a change detection paradigm. Although change detection has been investigated extensively, several of the models had never been applied to this task and no previous study had compared all models. Compared to previous studies, our use of a continuous stimulus variable and changes drawn from a wide range of magnitudes enhanced our ability to tell apart the model predictions. Our results suggest that working memory resource is continuous and variable and do not support the notion of an item limit.

The variable-precision model of change detection connects a continuous-resource encoding model of working memory [13] with a Bayesian model for decision-making in change detection [33]. This improves on two related change detection studies that advocated for continuous resources. Wilken and Ma [10] introduced the concept of continuous resources, but only compared an EP model with a suboptimal decision rule to the IP model. Although the EP model won in this comparison, the more recent item-limit models (SA and SR) had not yet been proposed at that time. Our present results show that the SA and SR models are improvements over both the EP and IP models, but lose to the VP model. In a more recent study, we compared different variants of the Bayesian model of the decision process and found that the optimal decision rule outperformed suboptimal ones [33], but we did not vary set size or compare different models of working memory. Other tasks, such as change localization [13], visual search [21,23], and multiple-object tracking [19,46], can also be conceptualized using a resource-limited front end conjoined with a Bayesian-observer back end. Whether such a conceptualization will survive a deeper understanding of resource limitations remains to be seen.

It is instructive to consider each model in terms of the distribution over precision that it postulates for a given set size. In the IP model, this distribution has mass at infinity and, depending on set size, also at zero. In the SA and SR models, probability mass resides, depending on set size, at one or two nonzero values, or at zero and one nonzero value. The EP model has probability mass only at one nonzero value. The VP model is the only model considered that assigns probability to a broad, continuous range of precision values. Roughly speaking, the more values of precision a model allows, the better it seems to fit. Although we assumed in the VP model that precision follows a gamma distribution, it is possible that a different continuous distribution can describe variability in precision better. However, the amount of data needed to distinguish different
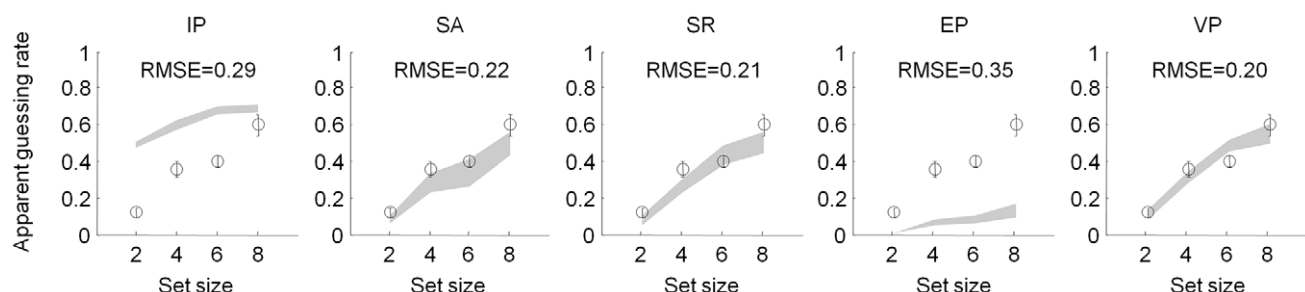


**Figure 5. Apparent guessing analysis.** Apparent guessing rate as a function of set size as obtained from subject data (circles and error bars) and synthetic data generated by each model (shaded areas). Even though the VP model does not contain any "true" guesses, it still accounts best for the apparent guessing rate.
doi:10.1371/journal.pcbi.1002927.g005

continuous precision distributions using psychophysics only might be prohibitive.

Work by Rouder et al. used a change detection task to compare a continuous-resource model based on signal detection theory to a variant of the IP model [8]. Manipulating bias, they measured receiver-operating characteristics (ROCs). The IP variant predicted straight-line ROCs, whereas the continuous-resource model predicted regular ROCs (i.e., passing through the origin). Unfortunately, each of the ROCs they measured contained only three points, and therefore the models were very difficult to distinguish. We ourselves, in an earlier study, had collected five-point ROCs using confidence ratings, allowing for an easier distinction between different ROC types; there, we found that the ROCs were regular [10], in support of a continuous-resource model. A difference between the Rouder study and our current study is that Rouder et al. used ten distinct colors instead of a one-dimensional continuum; this again has the disadvantage of missing the stimulus regime in which the signal-to-noise ratio is low. Moreover, the decision process in their continuous-resource model was not optimal; an optimal observer would utilize knowledge of the distribution of the stimuli and change magnitudes used in the experiment. It is likely that the optimal decision rule would have described human behavior in Rouder et al.'s experiment better than an ad-hoc suboptimal rule [33]. Finally, Rouder et al. did not consider variability in precision. In short, our current study does not contradict the results of Rouder et al., but offers a more plausible continuous-resource model and tests all models over a broader range of experimental conditions.

The notion of an item limit on the one hand and continuous or variable resources on the other hand are not mutually exclusive. In the SR model, for example, a continuous resource is split among a limited number of items. Although this model was not the best in the present study, many other "hybrid" models can be conceived – such as a VP model augmented with an item limit, or an IP or SA model with variable capacity [47–48] – and testing them is an important direction for future work. Our results, however, establish the VP model as the standard against which any new model of change detection should be compared.

## Neural implications

The neural basis of working memory limitations is unknown. In the variable-precision model, encoding precision is the central concept, raising the question which neural quantity corresponds to encoding precision. We hypothesize that precision relates to neural gain, according to the reasoning laid out in previous work [13,19,33]. To summarize, gain translates directly to precision in sensory population codes [49], increased gain correlates with increased attention [50], and high gain is energetically costly [51], potentially bringing encoding precision down as set size increases. The variable-precision model predicts that the gain associated with the encoding of each item exhibits large fluctuations across items and trials. There is initial neurophysiological support for this prediction [52–53]. Furthermore, if gain is variable, then spiking activity originates from a doubly stochastic process: spiking is stochastic for a given of value of gain, while gain is stochastic itself. Recent evidence points in this direction [54–55], although formal model comparison remains to be done. The variable-precision model also predicts that gain on average decreases with increasing set size. We proposed in earlier work that this could be realized mechanistically by divisive normalization [19]. Divisive normalization could act on the gains of the input populations by approximately dividing each gain by the sum of the gains across all locations raised to some power [56]. When set size is larger, the division would be by a larger number, resulting in a post-normalization gain that decreases with set size. A spiking neural network implementation of aspects of continuous-resource models was proposed recently [57]. Taken together, the variable-precision model has plausible neural underpinnings.

Our results have far-reaching implications for neural studies of working memory limitations. Throughout the field, taking a fixed item limit for granted has been the norm, and many studies have focused on finding its neural correlates [12,58]. Even if we restrict ourselves to change detection only, a fixed item limit has been assumed by studies that used fMRI [59–65], EEG [66–72], MEG [67,72–73], voxel-based morphometry [74], TMS [68,75], lesion patients [76], and computational models [77–78]. Our present results undermine the theoretical basis of all these studies. Neural studies that questioned the item-limit model or attempted to correlate neural measures with parameters in a continuous-resource model have been rare [27,57]. Perhaps, this is because no continuous-resource model has so far been perceived as compelling. The variable-precision model remedies this situation and might inspire a new generation of neural studies.

## Materials and Methods

### Stimuli

Stimuli were displayed on a 21″ LCD monitor at a viewing distance of approximately 60 cm. Stimuli were oriented ellipses with minor and major axes of 0.41 and 0.94 degrees of visual angle (deg), respectively. On each trial, ellipse centers were chosen by placing one at a random location on an imaginary circle of radius 7 deg around the screen center, placing the next one 45° counterclockwise from the first along the circle, etc., until all ellipses had been placed. Set size was 2, 4, 6, or 8. Each ellipse position was jittered by a random amount between −0.3 and 0.3 deg in both $x$- and $y$-directions to reduce the probability of orientation alignments between items. Stimulus and background luminances were 95.7 and 33.1 cd/m$^2$, respectively.

### Participants

Ten observers participated (4 female, 6 male; 3 authors). All were between 20 and 35 years old, had normal or corrected-to-normal vision, and gave informed consent.

### Procedure

On each trial, the first stimulus display was presented for 117 ms, followed by a delay period (1000 ms) and a second stimulus display (117 ms). In the first display, set size was chosen randomly and the orientation of each item was drawn independently from a uniform distribution over all possible orientations. The second display was identical to the first, except that there was a 50% chance that one of the ellipses had changed its orientation by an angle drawn from a uniform distribution over all possible orientations. The ellipse centers in the second screen were jittered independently from those in the first. Following the second display, the observer pressed a key to indicate whether there was a change between the first and second displays. A correct response caused the fixation cross to turn green and an incorrect response caused it to turn red. During the instruction phase, observers were informed in lay terms about the distributions from which the stimuli were drawn (e.g., "The change is equally likely to be of any magnitude."). Each observer completed three sessions of 600 trials each, with each session on a separate day, for a total of 1800 trials. There were timed breaks after every 100 trials. During each break, the screen displayed the observer's cumulative percentage correct.

## Model fitting and model comparison

Methods for model fitting and model comparison are described in the Text S1.

## Supporting Information

**Figure S1  Generative model.** The generative model shows the relevant variables in the change detection task and the statistical dependencies between them. $C$: change occurrence (0 or 1); $\Delta$: magnitude of change; $\boldsymbol{\Delta}$: vector of change magnitudes at all locations; $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$: vectors of stimuli in the first and second displays, respectively; $\mathbf{x}$ and $\mathbf{y}$: vectors of measurements in the first and second displays, respectively.
(TIF)

**Figure S2  Color change detection.** Observers reported whether one of the colors changed between the first and second displays.
(TIF)

**Figure S3  Color change detection: summary statistics and model fits.** (**a**) Model fits to the hit and false-alarm rates. (**b**) Model fits to the psychometric curves. Shaded areas represent $\pm 1$ s.e.m. in the model. For the IL model, a change of magnitude 0 has a separate proportion reports "change", equal to the false-alarm rate shown in (a). In each plot, the root mean square error between the means of data and model is given.
(TIF)

**Figure S4  Color change detection: Bayesian model comparison.** Model log likelihood of each model minus that of the VP model (mean $\pm$ s.e.m.). A value of $-x$ means that the data are $e^x$ times more probable under the VP model.
(TIF)

**Figure S5  Color change detection: apparent guessing analysis.** Apparent guessing rate as a function of set size as obtained from subject data (circles and error bars) and synthetic data generated by each model (shaded areas). Even though the VP model does not contain any "true" guesses, it still accounts best for the apparent guessing rate.
(TIF)

**Table S1  Mean and standard error of the maximum-likelihood estimates and tested ranges of model parameters for Experiment 2 (color change detection).**
(DOCX)

**Text S1  Supporting text.** Detailed derivation of Bayesian decision rule and explanation of model fitting and comparison methods. Explanation of color change detection experiment and results.
(DOCX)

## Author Contributions

Developed the theory: SK RvdB WJM. Conceived and designed the experiments: SK RvdB WJM. Performed the experiments: SK. Analyzed the data: SK RvdB WJM. Wrote the paper: SK RvdB WJM.

## References

1. Rensink RA (2002) Change detection. Annu Rev Psychol 53: 245–277.
2. Pashler H (1988) Familiarity and visual change detection. Percept Psychophys 44: 369–378.
3. Phillips WA (1974) On the distinction between sensory storage and short-term visual memory. Percept Psychophys 16: 283–290.
4. Irwin DE (1991) Information integration across saccadic eye movements. Cogn Psychology 23: 420–456.
5. Henderson JM (2008) Eye movements and visual memory. Visual Memory. Oxford: Oxford University Press. pp. 87–121.
6. Brouwer A-M, Knill DC (2007) The role of memory in visually guided reaching. Journal of Vision 7: 1–12.
7. Anderson DE, Vogel EK, Awh E (2011) Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. J Neurosci 31: 1128–1138.
8. Rouder J, Morey R, Cowan N, Morey C, Pratte M (2008) An assessment of fixed-capacity models of visual working memory. Proc Natl Acad Sci U S A 105: 5975–5979.
9. Zhang W, Luck SJ (2008) Discrete fixed-resolution representations in visual working memory. Nature 453: 233–235.
10. Wilken P, Ma WJ (2004) A detection theory account of change detection. J Vision 4: 1120–1135.
11. Bays PM, Husain M (2008) Dynamic shifts of limited working memory resources in human vision. Science 321: 851–854.
12. Fukuda K, Awh E, Vogel EK (2010) Discrete capacity limits in visual working memory. Curr Opin Neurobiol 20: 177–182.
13. Van den Berg R, Shin H, Chou W-C, George R, Ma WJ (2012) Variability in encoding precision accounts for visual short-term memory limitations. Proc Natl Acad Sci U S A 109: 8780–8785.
14. Alvarez GA, Cavanagh P (2004) The capacity of visual short-term memory is set both by visual information load and by number of objects. Psych Science 15: 106–111.
15. Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390: 279–281.
16. Cowan N (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci 24: 87–114.
17. Eng HY, Chen D, Jiang Y (2005) Visual working memory for simple and complex visual stimuli. Psychon B Rev 12: 1127–1133.
18. Awh E, Barton B, Vogel EK (2007) Visual working memory represents a fixed number of items regardless of complexity. Psych Science 18: 622–628.
19. Ma WJ, Huang W (2009) No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. J Vision 9: 3: 1–30.
20. Shaw ML (1980) Identifying attentional and decision-making components in information processing. In: Nickerson RS, editor. Attention and Performance. Hillsdale, NJ: Erlbaum. pp. 277–296.
21. Mazyar H, Van den Berg R, Ma WJ (2012) Does precision decrease with set size? J Vision 12: 10,11–16.
22. Bays PM, Catalao RFG, Husain M (2009) The precision of visual working memory is set by allocation of a shared resource. J Vision 9: 1–11.
23. Palmer J (1990) Attentional limits on the perception and memory of visual information. J Exp Psychol Hum Percept Perform 16: 332–350.
24. French RS (1953) The discrimination of dot patterns as a function of number and average separation of dots. J Exp Psychol 46: 1–9.
25. Elmore LC, Ma WJ, Magnotti JF, Leising KJ, Passaro AD, et al. (2011) Visual short-term memory compared in rhesus monkeys and humans. Curr Biol 21: 975–979.
26. Heyselaar E, Johnston K, Pare M (2011) A change detection approach to study visual working memory of the macaque monkey. J Vision 11: 11:11–10.
27. Buschman TJ, Siegel M, Roy RE, Miller EK (2011) Neural substrates of cognitive capacity limitations. Proc Natl Acad Sci 108: 11252–11255.
28. Lara AH, Wallis JD (2012) Capacity and precision in an animal model of short-term memory. J Vision 12: 1–12.
29. Scott-Brown KC, Baker MR, Orbach HS (2000) Comparison blindness. Visual Cognition 7: 253–267.
30. Simons DJ, Rensink RA (2005) Change blindness: past, present, and future. Trends Cogn Sci 9: 16–20.
31. Peterson WW, Birdsall TG, Fox WC (1954) The theory of signal detectability. Transactions IRE Profession Group on Information Theory, PGIT-4 : 171–212.
32. Knill DC, Richards W, eds. (1996) Perception as Bayesian Inference. New York: Cambridge University Press.
33. Keshvari S, Van den Berg R, Ma WJ (2012) Probabilistic computation in human perception under variability in encoding precision. PLoS ONE 7: e40216.
34. Cover TM, Thomas JA (1991) Elements of information theory. New York: John Wiley & Sons.
35. Mardia KV, Jupp PE (1999) Directional statistics: John Wiley and Sons. 350 p.
36. Seung H, Sompolinsky H (1993) Simple model for reading neuronal population codes. Proceedings of National Academy of Sciences USA 90: 10749–10753.
37. Paradiso M (1988) A theory of the use of visual orientation information which exploits the columnar structure of striate cortex. Biological Cybernetics 58: 35–49.
38. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9: 1432–1438.
39. Cowan N, Rouder JN (2009) Comment on "Dynamic shifts of limited working memory resources in human vision". Science 323: 877.
40. Keshvari S, Van den Berg R, Ma WJ (2012) Probabilistic computation in human perception under variability in encoding precision. PLoS One 7: e40216.
41. Vulkan N (2000) An economist's perspective on probability matching. J Economic Surveys 14: 101–118.

42. Mamassian P, Landy MS (1998) Observer biases in the 3D interpretation of line drawings. Vision Research 38: 2817–2832.
43. Wasserman L (2000) Bayesian model selection and model averaging. J Math Psych 44: 92–107.
44. Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90: 773–795.
45. Jeffreys H (1961) The theory of probability. Oxford University Press. 470 p.
46. Vul E, Frank M, Alvarez GA, Tenenbaum JB (2009) Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. Neural Information Processing Systems 22: 1955.
47. Dyrholm M, Kyllingsbaek S, Espeseth T, Bundesen C (2011) Generalizing parametric models by introducing trial-by-trial parameter variability: The case of TVA. J Math Psych 55: 416–429.
48. Sims CR, Jacobs RA, Knill DC (2012) An ideal-observer analysis of visual working memory. Psychol Rev 119: 807–30.
49. Pouget A, Dayan P, Zemel RS (2003) Inference and computation with population codes. Annual Review of Neuroscience 26: 381–410.
50. Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18: 193–222.
51. Lennie P (2003) The cost of cortical computation. Curr Biol 13: 493–497.
52. Cohen MR, Maunsell JHR (2010) A neuronal population measure of attention predicts behavioral performance on individual trials. J Neurosci 30: 15241–15253.
53. Nienborg H, Cumming BG (2009) Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature 459: 89–92.
54. Churchland AK, Kiani R, Chaudhuri R, Wang X-J, Pouget A, et al. (2011) Variance as a signature of neural computations during decision-making. Neuron 69: 818–831.
55. Churchland MM, Yu BM, Cunningham JP, Sugrue LP, Cohen MR, et al. (2010) Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nat Neurosci 13: 369–378.
56. Reynolds JH, Heeger DJ (2009) The normalization model of attention. Neuron 61: 168–185.
57. Wei Z, Wang XJ, Wang DH (2012) From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. J Neurosci 32: 11228–11240.
58. Marois R, Ivanoff J (2005) Capacity limits of information processing in the brain. Trends Cogn Sci 9: 296–305.
59. Xu Y, Chun MM (2006) Dissociable neural mechanisms supporting visual short-term memory for objects. Nature 440: 91–95.
60. Todd JJ, Marois R (2004) Capacity limit of visual short-term memory in human posterior parietal cortex. Nature 428: 751–754.
61. Mitchell (2008) Flexible, capacity-limited activity of posterior parietal cortex in perceptual as well as visual short-term memory tasks. Cereb Cortex 18: 1788–1798.
62. Konstantinou N, Bahrami B, Rees G, Lavie N (2012) Visual short-term memory load reduces retinotopic cortex response to contrast. J Cogn Neurosci 24: 2199–210.
63. Xu Y (2010) The neural fate of task-irrelevant features in object-based processing. J Neurosci 30: 14020–14028.
64. Schneider-Garces N, Gordon B, Brumback-Peltz C, Shin E, Lee Y, et al. (2010) Span, CRUNCH, and beyond: working memory capacity and the aging brain. J Cogn Neurosci 2010: 4.
65. Harrison A, Jolicoeur P, Marois R (2010) "What" and "where" in the intraparietal sulcus: an fMRI study of object identity and location in visual short-term memory. Cereb Cortex 20: 2478–2485.
66. Vogel EK, Machizawa MG (2004) Neural activity predicts individual differences in visual working memory capacity. Nature 428: 748–751.
67. Palva J, Monto S, Kulashekhar SP, S (2010) Neuronal synchrony reveals working memory networks and predicts individual memory capacity. Proc Natl Acad Sci U S A 107: 7580–7585.
68. Sauseng P, Klimesch W, Heise KF, Glennon M, Gerloff C, et al. (2009) Brain oscillatory substrates of visual short-term memory capacity. Curr Biol 19: 1846–1852.
69. Luria R, Sessa P, Gotler A, Jolicoeur P, Dell'acqua R (2009) Visual short-term memory capacity for simple and complex objects. J Cogn Neurosci 22: 496–512.
70. Grimault S, Robitaille N, Grova C, Lina J-M, Dubarry A-S, et al. (2009) Oscillatory activity in parietal and dorsolateral prefrontal cortex during retention in visual short-term memory: additive effects of spatial attention and memory load. Human Brain Mapping 30: 3378–3392.
71. Emrich S, Al-Aidroos N, Pratt J, Ferber S (2009) Visual search elicits the electrophysiological marker of visual working memory. PLoS ONE 4: e8042.
72. Palva S, Kulashekhar S, Hämäläinen M, Palva J (2011) Localization of cortical phase and amplitude dynamics during visual working memory encoding and retention. J Neurosci 31: 5013–5025.
73. Roux F, Wibral M, Mohr H, Singer W, Uhlhaas P (2012) Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. J Neurosci 32: 12411–12420.
74. Sligte I, Scholte H, Lamme V (2009) Grey matter volume explains individual differences in visual short-term memory capacity. J Vision 9: 598.
75. Tseng P, Hsu T-Y, Chang C-F, Tzeng O, Hung D, et al. (2012) Unleashing potential: transcranial direct current stimulation over the right posterior parietal cortex improves change detection in low-performing individuals. J Neurosci 32: 10554–10561.
76. Jeneson A, Wixted J, Hopkins R, Squire L (2012) Visual working memory capacity and the medial temporal lobe. J Cogn Neurosci 32: 3584–3589.
77. Lisman JE, Idiart MA (1995) Storage of 7+/−2 short-term memories in oscillatory subcycles. Science 267: 1512–1515.
78. Raffone A, Wolters G (2001) A cortical mechanism for binding in visual working memory. J Cogn Neurosci 13: 766–785.

**No evidence for an item limit in change detection**
Shaiyan Keshvari, Ronald van den Berg, Wei Ji Ma

**SUPPORTING TEXT**

**1. The Bayesian decision process in change detection**
The first step in Bayesian modeling of a perceptual task is to specify the statistical structure of the task, also called the generative model (Fig. S1). We denote change occurrence by $C$, a binary variable taking values 0 and 1 (0: no change; 1: change). The $N$ stimuli in the first display, denoted by a vector $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_N)$, take values on the circle. They are drawn independently from a uniform distribution over the stimulus space, so that $p(\boldsymbol{\theta})=1/(2\pi)^N$. We denote by boldface $\boldsymbol{\Delta}$ the vector of change magnitudes at all locations. It is a vector of zeros except at the location where the change, if any, is located. All locations are equally likely to contain the change. The magnitude of change, denoted by non-boldface $\Delta$, is drawn from a uniform distribution, $p(\Delta)=1/(2\pi)$. The $N$ stimuli in the second display, denoted by $\boldsymbol{\varphi}=(\varphi_1,\ldots,\varphi_N)$, are equal to the sum of the initial stimulus vector $\boldsymbol{\theta}$ and the change vector $\boldsymbol{\Delta}$: $\boldsymbol{\varphi} = \boldsymbol{\theta} + \boldsymbol{\Delta}$. Finally, the measurements in the first and second displays, $\mathbf{x}=(x_1,\ldots,x_N)$ and $\mathbf{y}=(y_1,\ldots,y_N)$, respectively, are drawn from independent Von Mises distributions with corresponding concentration parameters $\boldsymbol{\kappa}_x=(\kappa_{x,1},\ldots,\kappa_{x,N})$ and $\boldsymbol{\kappa}_y=(\kappa_{y,1},\ldots,\kappa_{y,N})$:

$$p\left(\mathbf{x}\mid\boldsymbol{\theta};\boldsymbol{\kappa}_x\right)=\prod_{i=1}^{N}\frac{1}{2\pi I_0\left(\kappa_{x,i}\right)}e^{\kappa_{x,i}\cos\left(x_i-\theta_i\right)},$$

$$p\left(\mathbf{y}\mid\boldsymbol{\varphi};\boldsymbol{\kappa}_y\right)=\prod_{i=1}^{N}\frac{1}{2\pi I_0\left(\kappa_{y,i}\right)}e^{\kappa_{y,i}\cos\left(y_i-\varphi_i\right)}.$$

This allows for the possibility that the concentration parameter of the noise at a given location is different between the first and the second displays. In the variable-precision model, this will in general be the case, since these values are modeled as being drawn independently.

Now we are ready to work out the observer's inference process. Bayesian inference amounts to computing the probability of $C$ given the noisy measurements in the first and second displays on a given trial, $\mathbf{x}$ and $\mathbf{y}$. The Bayesian decision variable $d$ is the ratio between the probability that a change occurred and the probability that no change occurred, given $\mathbf{x}$ and $\mathbf{y}$:

$$d = \frac{p\left(C=1\mid\mathbf{x},\mathbf{y}\right)}{p\left(C=0\mid\mathbf{x},\mathbf{y}\right)} = \frac{p\left(\mathbf{x},\mathbf{y}\mid C=1\right)p\left(C=1\right)}{p\left(\mathbf{x},\mathbf{y}\mid C=0\right)p\left(C=0\right)}, \tag{1}$$

where $p(C)$ is the prior probability of change occurrence. $d$ is called the *posterior ratio* of change occurrence and reflects the observer's belief about the occurrence of a change on one trial. To compute $d$, we need $p(\mathbf{x},\mathbf{y}|C)$, the distribution of the measurements in either the no-change or the

1

change conditions. This distribution is obtained by marginalizing the joint distribution of all variables over $\Delta$ (scalar), $\boldsymbol{\Delta}$ (vector), $\boldsymbol{\theta}$, and $\boldsymbol{\varphi}$. The result is

$$p(\mathbf{x},\mathbf{y}\,|\,C) = \left(\frac{1}{2\pi}\right)^{N+1} \frac{1}{N} \sum_{i=1}^{N} \left( \prod_{j\neq i} \int p(x_j\,|\,\theta_j)\,p(y_j\,|\,\theta_j)\,d\theta_j \right) \iint p(x_i\,|\,\theta_i)\,p(y_i\,|\,\theta_i + C\Delta)\,d\theta_i\,d\Delta. \quad (2)$$

The sum over $i$ runs over the $N$ possible locations of the change. In Eq. (2), $p(x_i|\theta_i)$ and $p(y_i|\varphi_i)$ are interpreted as likelihood functions over the stimulus – single-trial quantities that reflect the observer's belief about the stimulus ($\theta_i$ or $\varphi_i$) based on the measurement on that trial ($x_i$ or $y_i$). The width of the stimulus likelihood function reflects the observer's uncertainty about the stimulus.

Twice substituting Eq. (2) in Eq. (1) and simplifying, we find

$$d = \frac{p_{\text{change}}}{1 - p_{\text{change}}} \frac{1}{2\pi N} \sum_{i=1}^{N} \frac{\iint p(x_i\,|\,\theta_i)\,p(y_i\,|\,\theta_i + \Delta)\,d\theta_i\,d\Delta}{\int p(x_i\,|\,\theta_i)\,p(y_i\,|\,\theta_i)\,d\theta_i},$$

where $p_{\text{change}}$ is the observer's prior belief that a change occurred. Finally, substituting the Von Mises noise distributions and evaluating the integrals yields [1]

$$d = \frac{p_{\text{change}}}{1 - p_{\text{change}}} \frac{1}{N} \sum_{i=1}^{N} \frac{I_0\left(\kappa_{\text{x},i}\right) I_0\left(\kappa_{\text{y},i}\right)}{I_0\left(\sqrt{\kappa_{\text{x},i}^2 + \kappa_{\text{y},i}^2 + 2\kappa_{\text{x},i}\kappa_{\text{y},i}\cos\left(y_i - x_i\right)}\right)}. \quad (3)$$

The observer responds "change" ($\hat{C}=1$) when $d>1$ and "no change" ($\hat{C}=0$) when $d<1$. This is called maximum-a-posteriori estimation and it is the strategy that maximizes accuracy if $p_{\text{change}}$ equals the true frequency of change trials, i.e. 0.5 in our experiments.

This model of change detection automatically encompasses both the integration of information across items and the comparison between displays. The former is reflected by the sum in Eq. (3), the latter by the difference between the measurements in the first and second displays, $y_i - x_i$. Since this difference is subject to noise, errors will be made. Thus, the notion of "comparison errors" in change detection [2] is incorporated automatically and in a principled manner in the the Bayesian model [1]. In previous work, we proposed the decision rule $\max_i |y_i - x_i| > k$, where $|\cdot|$ denotes circular distance and $k$ is an unknown criterion [3], but because this rule is not equivalent to Eq. (3), it is not optimal. Another important feature of Eq. (3) is that the local concentration parameters in both displays, $\kappa_{\text{x},i}$ and $\kappa_{\text{y},i}$, play a role. This is analogous to trial-to-trial weighting by reliability in cue integration [4,5]: evidence from less noisy measurements is weighted more heavily.

## 2. Model fitting and model comparison

*Model predictions.* In the IL model, simple expressions exist for the probability of reporting "change": $p(\hat{C}=1|C=1)=K/N\cdot(1-\varepsilon)+(1-K/N)\cdot g$ when $N>K$, $p(\hat{C}=1|C=1)=1-\varepsilon$ when $N\leq K$, and $p(\hat{C}=1|C=0)=g$. In the other models, the probability of reporting "change" is equal to the probability that $d>1$. This probability depends on set size $N$, change magnitude $\Delta$, and the model parameters. Because no analytical expression exists for this probability in any of the models, we estimated it for each model using Monte Carlo simulation: for each experimental trial and parameter combination, we drew 10,000 samples of $(\mathbf{x},\mathbf{y})$ pairs from the generative model. For each sample, the model's decision rule was applied, and the proportion of "change" responses among all samples was determined. This returned the model's probability of reporting "change" on each trial for the given parameter values.

*Parameter fitting.* For a given model $m$, we denote the vector of model parameters by $\mathbf{t}$. The likelihood of a specific combination of parameter values is the probability of the observer's empirical responses given that parameter combination: $L_m(\mathbf{t}) = p(\text{data}\,|\,\mathbf{t}) = \prod_{k=1}^{N_{\text{trials}}} p\left(\hat{C}_k\,|\,N_k,\Delta_k,\mathbf{t}\right)$, where $N_{\text{trials}}$ is the total number of trials, and $N_k$, $\Delta_k$, and $\hat{C}_k$, denote set size, change magnitude, and the subject's response on the $k^{\text{th}}$ trial, respectively. The maximum-likelihood estimate of the parameters is the value of $\mathbf{t}$ that maximizes $L_m(\mathbf{t})$.

*Bayesian model comparison.* Bayesian model comparison is a method to compare models that automatically penalizes models with more free parameters [6,7]. Each model $m$ produces a prediction for probability of a response $\hat{C}_k$ on each trial, $p(\hat{C}_k|N_k,\Delta_k,\mathbf{t})$. Bayesian model comparison consists of calculating, for each model, the joint probability of a subject's empirical responses, averaged over free parameters. This gives the model likelihood, denoted $L(m)$:

$$L(m) = p(\text{data}\,|\,m) = \int p(\text{data}\,|\,m,\mathbf{t})\,p(\mathbf{t}\,|\,m)\,d\mathbf{t} = \int \left( \prod_{k=1}^{N_{\text{trials}}} p\left(\hat{C}_k\,|\,N_k,\Delta_k,\mathbf{t}\right) \right) p(\mathbf{t}\,|\,m)\,d\mathbf{t}.$$

We chose the prior over the $j^{\text{th}}$ parameter to be uniformly distributed on an interval of size $R_j$. The intervals are given in Tables 1 and S1. After rewriting for numerical convenience, we arrive at $\log L(m) = L_{\max}(m) - \sum_{j=1}^{\dim \mathbf{t}} \log R_j + \log \int e^{\log L_m(\mathbf{t}) - \log L_{\max}(m)}\,d\mathbf{t}$, where $\dim \mathbf{t}$ is the number of parameters and $L_{\max}(m) = \max_t L_m(\mathbf{t})$. We computed the integrals over parameters numerically by using the trapezoidal rule.

## 3. Supplementary results: color change detection

To examine the generality of our conclusions, we repeated the orientation change detection experiment described in the main text with color stimuli (Fig. S2). This experiment was identical

except for the following differences. Stimuli were colored discs with a radius of 0.36 deg. The colors were drawn independently from 180 color values uniformly distributed along a circle of radius 60 in CIE 1976 (*L\**, *a\**, *b\**) color space. This circle had constant luminance (*L\**=58) and was centered at the point (*a\**=12, *b\**=13). Stimulus positions were not jittered. On each trial, set size was 1, 2, 4, or 8, with equal probabilities. Seven subjects participated (4 female, 3 male; 1 author).

The results are qualitatively consistent with those of the main experiment. We found a significant effect of set size on hit rate ($F(3,18)$=40.4, $p$<0.001) and false-alarm rate ($F(3,18)$=26.2, $p$<0.001) (Fig. S3a). The probability of reporting a change increased with change magnitude at every set size (Fig. S3b). The VP model best fitted the hit and false-alarm rates (RMSE of 0.046 versus 0.059 and higher) as well as the psychometric curves (0.11 versus 0.14 and higher). The VP model also won in Bayesian model comparison (Fig. S4), surpassing the SA model by 48.4 ± 8.2 points. As for orientation, we found apparent guessing rates that increased with set size ($F(3,18)$=31.8, $p$<0.001) and were non-zero for all set sizes ($t(6)$>5.45, $p$<0.001) (Fig. S5). The VP model was better able to account for this than the next best model (the SA model), with RMSE values of 0.09 and 0.17, respectively. Parameter estimates are given in Table S1.

1. Keshvari S, Van den Berg R, Ma WJ (2012) Probabilistic computation in human perception under variability in encoding precision. PLoS ONE 7: e40216.
2. Awh E, Barton B, Vogel EK (2007) Visual working memory represents a fixed number of items regardless of complexity. Psych Science 18: 622-628.
3. Wilken P, Ma WJ (2004) A detection theory account of change detection. J Vision 4: 1120-1135.
4. Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27: 712-719.
5. Yuille AL, Bulthoff HH (1996) Bayesian decision theory and psychophysics. In: Knill DC, Richards W, editors. Perception as Bayesian Inference. New York: Cambridge: University Press.
6. MacKay DJ (2003) Information theory, inference, and learning algorithms. Cambridge, UK: Cambridge University Press.
7. Wasserman L (2000) Bayesian model selection and model averaging. J Math Psych 44: 92-107.

**SUPPORTING TABLE**

| | | Experiment 2 estimates | | Tested range | |
|---|---|---|---|---|---|
| Model | Parameter | Mean | S.e.m. | Min | Max |
| IL | $K$ | 4.57 | 0.43 | 1 | 8 |
| | $\varepsilon$ | 0.229 | 0.009 | 0 | 1 |
| | $G$ | 0.172 | 0.005 | 0 | 0.5 |
| SA | $J_1$ | 3.97 | 0.72 | 1 | 40 |
| | $K$ | 4.14 | 0.59 | 1 | 8 |
| | $p_{change}$ | 0.591 | 0.016 | 0.2 | 0.8 |
| SR | $J_1$ | 13.8 | 2.0 | 1 | 60 |
| | $K$ | 4.14 | 0.46 | 1 | 8 |
| | $p_{change}$ | 0.574 | 0.019 | 0.2 | 0.8 |
| EP | $J_1$ | 13.6 | 4.1 | 1 | 60 |
| | $\alpha$ | -0.914 | 0.195 | -2 | 0 |
| | $p_{change}$ | 0.506 | 0.008 | 0.2 | 0.8 |
| VP | $\bar{J}_1$ | 111 | 14.0 | 5 | 300 |
| | $\tau$ | 163 | 35 | 5 | 300 |
| | $\alpha$ | -0.471 | 0.042 | -2 | 0 |
| | $p_{change}$ | 0.557 | 0.005 | 0.2 | 0.8 |

**Table S1.** Mean and standard error of the maximum-likelihood estimates and tested ranges of model parameters for Experiment 2 (color change detection).
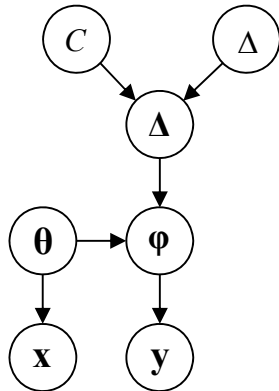

**SUPPORTING FIGURES**



**Fig. S1.** The generative model shows the relevant variables in the change detection task and the statistical dependencies between them. $C$: change occurrence (0 or 1); $\Delta$: magnitude of change; $\boldsymbol{\Delta}$: vector of change magnitudes at all locations; $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$: vectors of stimuli in the first and second displays, respectively; $\mathbf{x}$ and $\mathbf{y}$: vectors of measurements in the first and second displays, respectively.
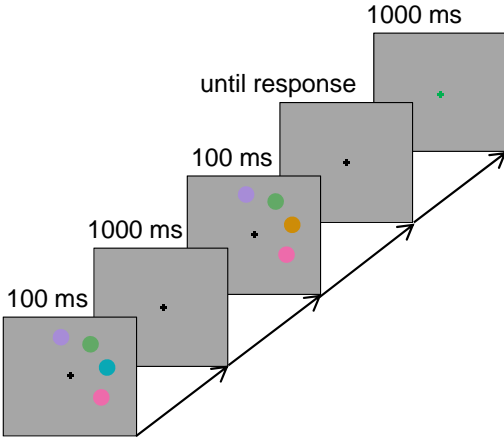
**Fig. S2. Color change detection.** Observers reported whether one of the colors changed between the first and second displays.
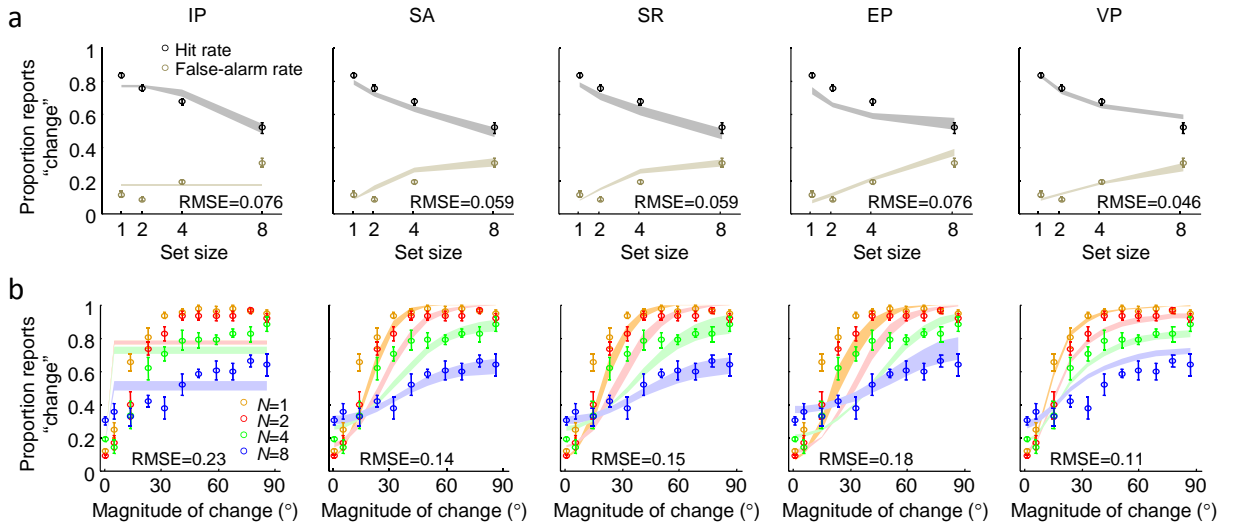


**Fig. S3. Color change detection: summary statistics and model fits. (a)** Model fits to the hit and false-alarm rates. **(b)** Model fits to the psychometric curves. Shaded areas represent ± 1 s.e.m. in the model. For the IL model, a change of magnitude 0 has a separate proportion reports "change", equal to the false-alarm rate shown in (a). In each plot, the root mean square error between the means of data and model is given.
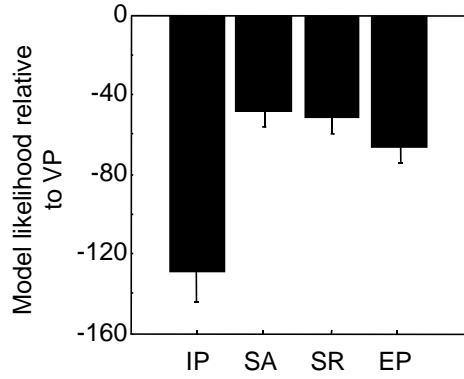
**Fig. S4. Color change detection: Bayesian model comparison.** Model log likelihood of each model minus that of the VP model (mean ± s.e.m.). A value of $-x$ means that the data are $e^x$ times more probable under the VP model.
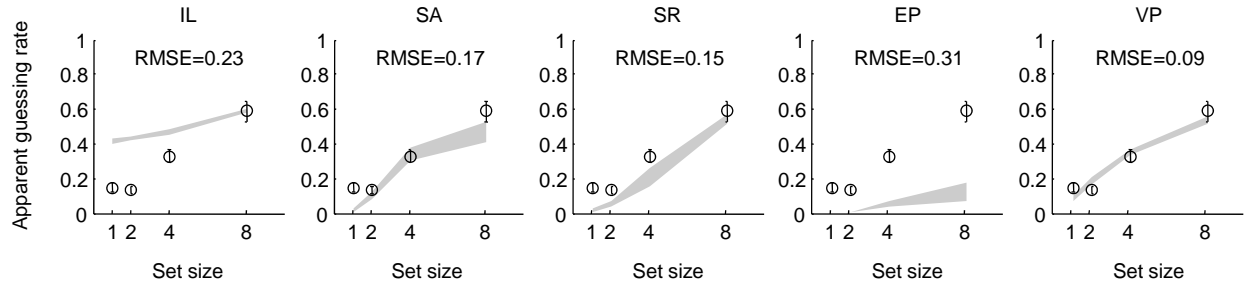


**Fig. S5. Color change detection: apparent guessing analysis.** Apparent guessing rate as a function of set size as obtained from subject data (circles and error bars) and synthetic data generated by each model (shaded areas). Even though the VP model does not contain any "true" guesses, it still accounts best for the apparent guessing rate.