A rational theory of set size effects in working memory and attention

Ronald van den Berg[1] and Wei Ji Ma[2]

[1]Department of Psychology, University of Uppsala, Uppsala, Sweden.

[2]Center for Neural Science and Department of Psychology, New York University, New York, USA

Running head: A rational theory of set size effects

Keywords

Working memory, attention, ecological rationality, normative models

Corresponding author

Ronald van den Berg

Department of Psychology

Von Kraemers Allé 1A, 75237

Uppsala, Sweden

Tel: +46184717277

E-mail: ronald.vandenberg@psyk.uu.se

ABSTRACT

The precision with which items are encoded in working memory and attention decreases with the number of encoded items. Current theories typically account for this "set size effect" by postulating a hard constraint on the allocated amount of encoding resource. While these theories have produced models that are descriptively successful, they offer no principled explanation for the very existence of set size effects: given their detrimental consequences for behavioral performance, why have these effects not been weeded out by evolutionary pressure, by allocating resources proportionally to the number of encoded items? Here, we propose a theory that is based on an ecological notion of rationality: set size effects are the result of a near-optimal trade-off between behavioral performance and the neural costs associated with stimulus encoding. We derive models for four visual working memory and attention tasks and show that they account well for data from eleven previously published experiments. Moreover, our results suggest that the total amount of resource that subjects allocate for stimulus encoding varies non-monotonically with set size, which is consistent with our rational theory of set size effects but not with previous descriptive theories. Altogether, our findings suggest that set size effects may have a rational basis and highlight the importance of considering ecological costs in theories of human cognition.

INTRODUCTION

Human cognition is strongly constrained by set size effects in working memory and attention: the precision with which these systems encode information rapidly declines with the number of items, as observed in for example delayed estimation, change detection, visual search, and multiple-object tracking tasks (Ma & Huang, 2009; Ma, Husain, & Bays, 2014; Mazyar, van den Berg, & Ma, 2012; Mazyar, Van den Berg, Seilheimer, & Ma, 2013; J Palmer, 1990; John Palmer, 1994; Shaw, 1980; Wilken & Ma, 2004). By contrast, set size effects seem to be absent in long-term memory, where fidelity has been found to be independent of set size (Brady, Konkle, Gill, Oliva, & Alvarez, 2013). The existence of set size effects is thus not a general property of stimulus encoding, but a phenomenon that requires explanation. Despite an abundance of models, such an explanation is still lacking.

A common way to model set size effects has been to assume that stimuli are encoded using a fixed total amount of resources, formalized as "samples" (Lindsay, Taylor, & Forbes, 1968; J Palmer, 1990; Sewell, Lilburn, & Smith, 2014; Shaw, 1980), slots (Zhang & Luck, 2008),

information bit rate (C. R. Sims, Jacobs, & Knill, 2012), Fisher information (Ma & Huang, 2009), or neural firing (Bays, 2014): the larger the number of encoded items, the lower the amount of resource available for each item and, therefore, the lower the precision per item. These models make a very specific prediction about set size effects: encoding precision is inversely proportional to set size. It has been found that this prediction is often inconsistent with empirical data, which has led more recent models to instead use a power law to describe set size effects (Bays, Catalao, & Husain, 2009; Bays & Husain, 2008; Devkar & Wright, 2015; Donkin, Kary, Tahir, & Taylor, 2016; Elmore et al., 2011; Keshvari, van den Berg, & Ma, 2013; Mazyar et al., 2012; van den Berg, Awh, & Ma, 2014; van den Berg, Shin, Chou, George, & Ma, 2012; Wilken & Ma, 2004). These more flexible power-law models tend to provide excellent fits to experimental data, but they have been criticized for lacking a principled motivation (Oberauer, Farrell, Jarrold, & Lewandowsky, 2016; Oberauer & Lin, 2017). Hence, previous research has evolved to power-law models that accurately describe *how* precision in working memory and attention depends on set size, but a principled theory that explains *why* these effects are best described by a power law – and why they exist at all – is still lacking. While there seems little room for further improvement in the descriptive power of these models, finding rational or normative answers to these more fundamental questions can deepen our understanding of the very origin of encoding limitations in working memory and attention.

Although several previous studies have used normative or rational theories to explain certain aspects of working memory and attention, none of them has accounted for set size effects in a principled way. One example is our own previous work on visual search (Mazyar et al., 2012, 2013), change detection (Keshvari, van den Berg, & Ma, 2012; Keshvari et al., 2013), and change localization (van den Berg et al., 2012), where we modelled the decision stage using optimal-observer theory, while assuming an ad hoc power law to model the relation between encoding precision and set size. Another example is the work by Sims and colleagues, who developed a normative framework in which working memory and perceptual systems are conceptualized as optimally performing information channels (C. R. Sims, 2016; C. R. Sims et al., 2012). Their framework offers parsimonious explanations for several aspects of stimulus encoding in visual working memory, such as the relation between stimulus variability and encoding precision (C. R. Sims et al., 2012) and the non-Gaussian shape of encoding noise (C. R. Sims, 2015). However, their framework does not offer a normative explanation of set size effects. In their early work (C. R. Sims et al., 2012) they accounted for these effects by assuming that total information capacity is fixed,

which is similar to other fixed-resource models and predicts an inverse proportionality between encoding precision and set size. In their later work (A Emin Orhan, Sims, Jacobs, Knill, & Orhan, 2014; C. R. Sims, 2016) they add to this the assumption that there is an inefficiency in distributing capacity across items and fit capacity as a free parameter at each set size. Neither of their assumptions is motivated by normative arguments.

Here, we propose that set size effects may be a near-optimal solution to an ecological trade-off. The starting point for our theory is the principle that stimulus encoding is costly (Attwell & Laughlin, 2001; Lennie, 2003; Sterling & Laughlin, 2015), which may have pressured the brain to balance behavioral benefits of high precision against neural costs (Christie & Schrater, 2015; Lennie, 2003; Ma & Huang, 2009; Pestilli & Carrasco, 2005). Indeed, consistent with this idea, it has been found that performance on perceptual decision-making tasks can be improved by increasing monetary reward (Baldassi & Simoncini, 2011; Della Libera & Chelazzi, 2006; Peck, Jangraw, Suzuki, Efem, & Gottlieb, 2009). However, what level of encoding precision establishes a good balance may depend not only on the level of reward, but possibly also on task-related factors such as set size. Based on these considerations, we hypothesize that set size effects are the result of an ecologically rational or normative strategy that balances behavioral performance against encoding costs. We next formalize this hypothesis, derive models from it for four visual working memory and attention tasks, and fit them to data from eleven previously published experiments.

THEORY

We first formalize and test our theory in the context of the delayed-estimation paradigm (Wilken & Ma, 2004) and will later examine its generalization to other tasks. In single-probe delayed-estimation tasks, subjects briefly hold a set of items in memory and report their estimate of a randomly chosen target item (Fig. 1A; Table 1). Estimation error ε is the (circular) difference between the subject's estimate and the true stimulus value s. Set size effects in this task manifest itself as a widening of the estimation error distribution (Fig. 1B). As in previous work (Keshvari et al., 2012, 2013, Mazyar et al., 2012, 2013, van den Berg et al., 2014, 2012), we assume that a memory x follows a Von Mises distribution with mean s and concentration parameter κ, and define encoding precision J as Fisher information (Cover & Thomas, 2005), which is one-to-one related to κ (see Supplementary Information). We assume that response noise is negligible, such that the estimation error is equal to the memory error, $\varepsilon = x - s$. Moreover, we assume variability in J across items and trials (Fougnie, Suchow, & Alvarez, 2012; Keshvari et al., 2012; Mazyar et al., 2012; van

den Berg et al., 2014, 2012), which we model using a gamma distribution with a mean $\bar{J}$ and a scale parameter τ (see Supplementary Information).

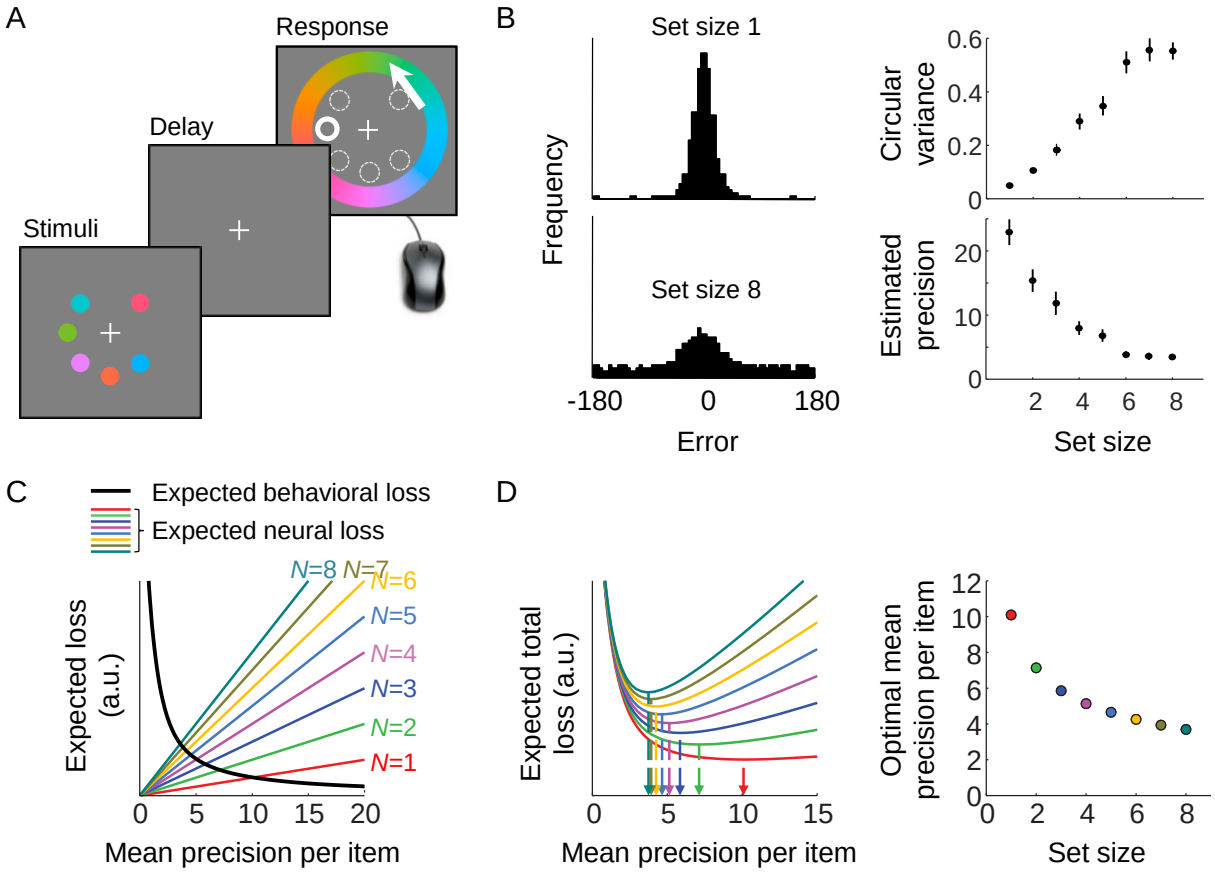


Figure 1. An ecologically rational model of set size effects in delayed estimation. (A) Example of a delayed-estimation experiment. The subject is briefly presented with a set of stimuli and, after a short delay, reports the value of a randomly chosen target item. (B) Estimation error distributions widen with set size, suggesting a decrease in encoding precision (data from Experiment DE5 in Table 1; estimated precision computed in the same way as in Fig. 3A). (C) Stimulus encoding is assumed to be associated with two kinds of loss: a behavioral loss that decreases with encoding precision and a neural loss that is proportional to both set size and precision. In the delayed-estimation task, the expected behavioral error loss is independent of set size. (D) Total expected loss has a unique minimum that depends on the number of remembered items. The mean precision per item that minimizes expected total loss is referred to as the optimal mean precision (arrows) and decreases with set size. The parameter values used to produce panels C and D were $\tilde{\lambda}$=0.01, β=2, and $\tau \downarrow 0$.

Table 1. Overview of experimental datasets used. Task responses were continuous in the delayed-estimation experiments and categorical in the other tasks. DE5 and DE6 differed in the way color was reported (DE5: color wheel; DE6: scroll).

Experiment	Reference	Task	Feature	Set sizes	#subj
DE1	(Wilken & Ma, 2004)	Delayed estimation	Color	1, 2, 4, 8	15
DE2	(Zhang & Luck, 2008)	Delayed estimation	Color	1, 2, 3, 6	8
DE3	(Bays et al., 2009)	Delayed estimation	Color	1, 2, 4, 6	12
DE4	(van den Berg et al., 2012)	Delayed estimation	Orientation	1-8	6
DE5	(van den Berg et al., 2012)	Delayed estimation	Color	1-8	13
DE6	(van den Berg et al., 2012)	Delayed estimation	Color	1-8	13
CD1	(Keshvari et al., 2013)	Change detection	Color	1, 2, 4, 8	7
CD2	(Keshvari et al., 2013)	Change detection	Orientation	2, 4, 6, 8	10
CL1	(van den Berg et al., 2012)	Change localization	Color	2, 4, 6, 8	7
CL2	(van den Berg et al., 2012)	Change localization	Orientation	2, 4, 6, 8	11
VS	(Mazyar et al., 2013)	Visual search	Orientation	1, 2, 4, 8	6

The key novelty of our theory is the idea that stimuli are encoded with a level of mean precision, $\bar{J}$, that minimizes a combination of *behavioral loss* and *neural loss*. Behavioral loss is induced by making an error ε, which we formalize using a mapping $L_{\text{behavioral}}(\varepsilon)$. This mapping may depend on both internal incentives (e.g., intrinsic motivation) and external ones (e.g., the reward scheme imposed by the experimenter). For the moment, we choose a power-law function, $L_{\text{behavioral}}(\varepsilon)=|\varepsilon|^{\beta}$ with $\beta>0$ as a free parameter, such that larger errors correspond with larger loss. The *expected* behavioral loss, denoted $\bar{L}_{\text{behavioral}}$, is obtained by averaging loss across all possible errors, weighted by the probability that each error occurs,

$$\bar{L}_{\text{behavioral}}\left(\bar{J},N\right)=\int L_{\text{behavioral}}\left(\varepsilon\right)p\left(\varepsilon\mid\bar{J},N\right)d\varepsilon, \qquad (1)$$

where $p\left(\varepsilon\mid\bar{J},N\right)$ is the estimation error distribution for given mean precision and set size. In single-probe delayed-estimation tasks, the expected behavioral loss is independent of set size and subject to the law of diminishing returns (Fig. 1C, black curve).

A second kind of loss is the energetic expenditure incurred by representing a stimulus. Since this loss is primarily rooted in neural spiking activity, we refer to it as "neural loss" and use neural theory to make an estimate of the relation between encoding precision and neural loss. For many choices of spike variability, including the common one of Poisson-like variability (Ma, Beck, Latham, & Pouget, 2006), the precision (Fisher information) of a stimulus encoded in a neural population is proportional to the trial-averaged neural spiking rate (Paradiso, 1988; Seung & Sompolinsky, 1993). Moreover, it has been estimated that the energetic loss induced by each spike increases with spiking rate (Attwell & Laughlin, 2001; Lennie, 2003). When combining these two premises, the expected neural loss associated with the encoding of an item is a supralinear function of encoding precision. However, to minimize free model parameters, we assume for the moment that the function is linear (at the end of this section we present a mathematical proof that the main qualitative prediction of our theory generalizes to any supralinear function). Further assuming that stimuli are encoded independently of each other, expected neural loss is also proportional to the number of encoded items, N. We thus obtain

$$\bar{L}_{\text{neural}}\left(\bar{J},N\right)=\alpha\bar{J}N, \tag{2}$$

where α is a free parameter that represents the amount of neural loss incurred by a unit increase in mean precision (Fig. 1C, colored lines).

We combine the two types of expected loss into a total expected loss function (Fig. 1D),

$$\begin{aligned}\bar{L}_{\text{total}}\left(\bar{J},N\right)&=\bar{L}_{\text{behavioral}}\left(\bar{J},N\right)+\lambda\bar{L}_{\text{neural}}\left(\bar{J},N\right)\\&=\bar{L}_{\text{behavioral}}\left(\bar{J},N\right)+\lambda\alpha\bar{J}N,\end{aligned} \tag{3}$$

where the weight $\lambda\geq 0$ represents the importance of keeping neural loss low relative to the importance of good performance. Since λ and α have interchangeable effects on the model predictions, they can be fitted as a single free parameter $\tilde{\lambda}\triangleq\lambda\alpha$. We refer to the level of mean precision that minimizes the total expected loss as *optimal mean precision*,

$$\bar{J}_{\text{optimal}}\left(\bar{J},N\right)=\underset{J}{\operatorname{argmin}}\,\bar{L}_{\text{total}}\left(\bar{J},N\right). \tag{4}$$

Under the loss functions proposed above, we find that $\bar{J}_{\text{optimal}}$ is a decreasing function of set size (Fig. 1D), which is qualitatively consistent with set size effects observed in experimental data (cf. Fig. 1B).

Generality

When formalizing the loss functions, we had to make specific assumptions about how behavioral errors map to behavioral loss and encoding precision to neural loss. Since these assumptions cannot yet be fully empirically substantiated, it is important to verify that our theory generalizes to other choices that we could have made. To this end, we asked under what conditions our general theory, Eq. (4), predicts a set size effect (i.e., a decline of encoding precision with set size). A mathematical proof (see Supplementary Materials) shows that the following four conditions are sufficient: (i) the expected behavioral loss is a strictly decreasing function of encoding precision, i.e., an increase in precision results in an increase in performance; (ii) the expected behavioral loss is subject to a law of diminishing returns (Mankiw, 2004): the higher the initial precision, the smaller the behavioral benefit obtained from an increase in precision; (iii) the expected neural loss is an increasing function of encoding precision; (iv) the expected neural loss associated with a fixed increase in precision increases with precision. Hence, the conditions under which our theory predicts set size effects are not limited to the specific loss functions that we formulated here, but represent a broad range of choices.

RESULTS

Model fits

To evaluate whether our theory can quantitatively account for experimental data, we fit the model formulated above to 67 individual-subject data sets from a delayed-estimation benchmark set[*] (Table 1). The maximum-likelihood fit accounts well for the raw error distributions (Fig. 2A) and the two statistics that summarize these distributions (Fig. 2B). Hence, these data are consistent with the theory that set size effects are the result of an ecologically rational trade-off between behavioral

[*] The original benchmark set (van den Berg et al., 2014) contains 10 data sets with a total of 164 individuals. Two of these data sets were published in papers that later got retracted and another one contained data for only two set sizes, which is not very informative for our present purposes. While our model accounts well for these data sets (Fig. S1 in Supplementary Information), we decided to exclude them from the main analyses.

performance and neural cost. Maximum-likelihood estimates of the three model parameters ($\tilde{\lambda}$, τ, and β) are provided in Supplementary Table S1.

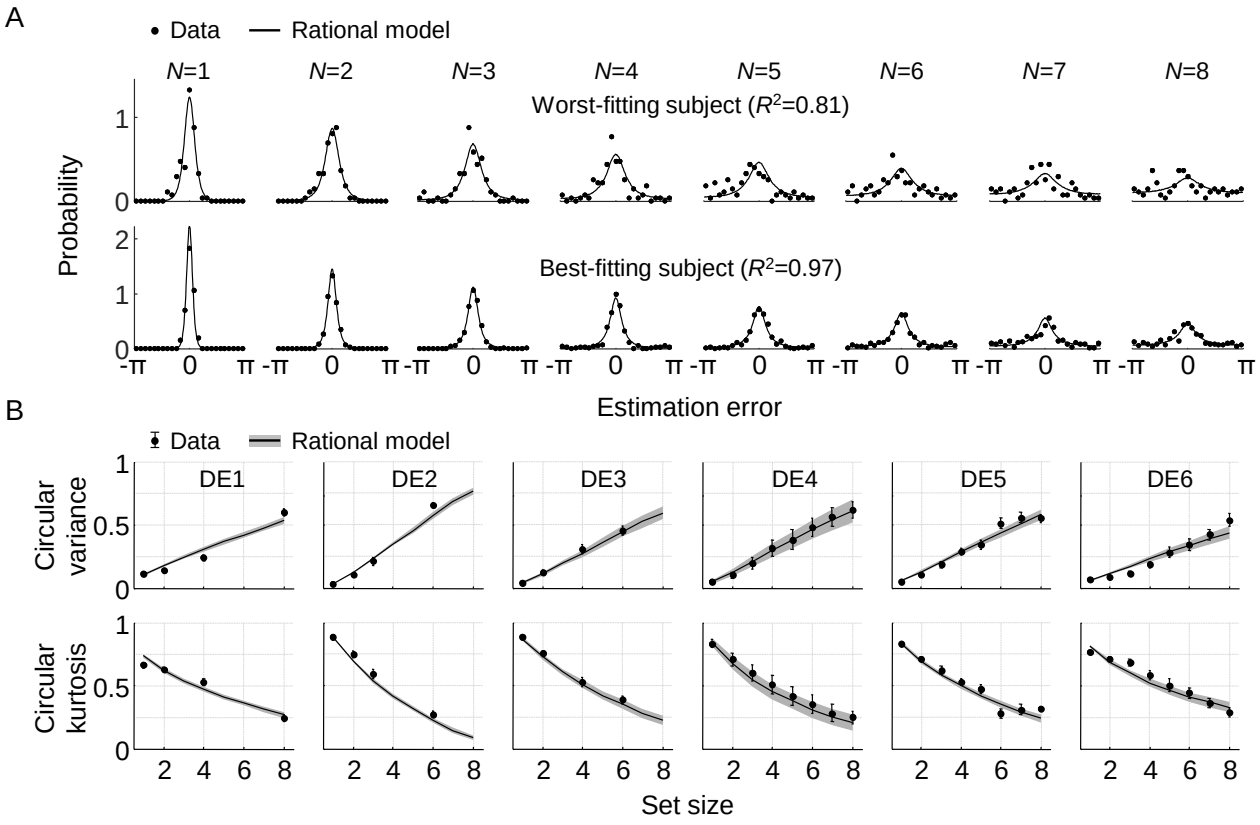


Figure 2. Model fits to data from six delayed-estimation experiments. (A) Maximum-likelihood fits to raw data of the worst-fitting and best-fitting subjects. Goodness of fit was measured as R^2, computed for each subject by concatenating histograms across set sizes. (B) Subject-averaged fits to the two statistics that summarize the estimation error distributions (circular variance and kurtosis) as a function of set size, split by experiment. Here and in subsequent figures, error bars and shaded areas represent 1 s.e.m. of the mean across subjects.

Comparison with a power-law model and an unconstrained model

To compare the goodness of fit of this model with that of previously proposed descriptive models, we next fit the same data using a model variant in which the relation between encoding precision and set size is assumed to be a power law. This variant is identical to the VP-A model in our earlier work (van den Berg et al., 2014). Model comparison based on the Akaike Information Criterion (AIC) (Akaike, 1974) indicates that the goodness of fit is comparable between the two models, with a small advantage for the rational model (ΔAIC=5.27±0.70; throughout the paper, X±Y indicates

mean±s.e.m. across subjects). Hence, the rational model provides a principled explanation of set size effects without sacrificing quality of fit compared to previous descriptive models.

To get an indication of the absolute goodness of fit of the rational model, we next examine how much room for improvement there is in the fits. We do this by fitting a model variant in which memory precision is a free parameter at each set size, while keeping all other aspects of the model the same (note that this model variant purely serves as a descriptive tool to obtain estimates of the empirical precision values, not as a process model of set size effects in visual working memory). We find a marginal AIC difference (ΔAIC=3.49±0.93, in favor of the unconstrained model), which indicates that the fits of the rational model are close to the best possible fits. This finding is corroborated by examination of the fitted parameter values: the estimated precision values in the unconstrained model closely match the precision values in the rational model (Fig. 3A).

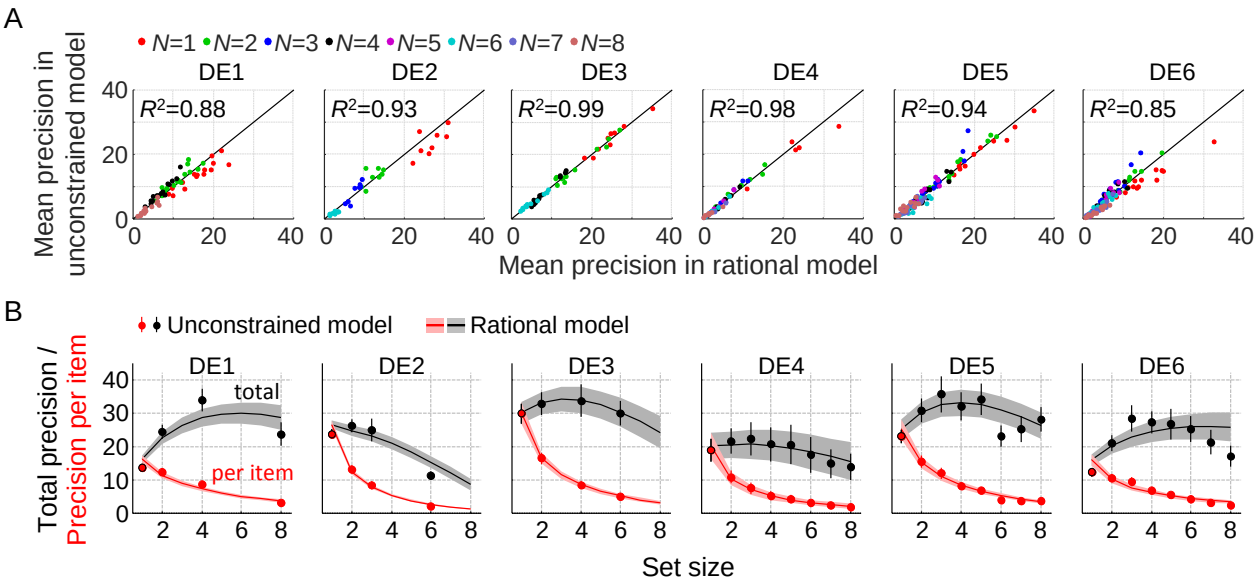


Figure 3. Estimated encoding precision in the delayed-estimation experiments. (A) Best-fitting precision values in the rational model scattered against the best-fitting precision values in the unconstrained model. Each dot represents the estimates for a single subject. (B) Estimated mean encoding precision per item (red) and total encoding precision (black) plotted against set size.

Total precision as a function of set size

One feature that sets our rational theory apart from previous theories is that it does not predict a trivial relationship between the total amount of allocated encoding resource and set size. To see this, we quantify the amount of allocated resources as the precision per item summed across all items,

$\bar{J}_{\text{total}} \triangleq \bar{J}N$. In fixed-resource models, this quantity is by definition constant and in power-law models it varies monotonically with set size. By contrast, we find that in the fits to several of the delayed-estimation experiments, total precision in the rational model varies *non-monotonically* with set size (Fig. 3B, gray curves). To examine whether there is evidence for such non-monotonic behavior in the subject data, we use the fitted precision values from the unconstrained model as our best empirical estimates of the precision with which subjects encoded items. We find that these empirical estimates show signs of similar non-monotonic relations in some of the experiments (Fig. 3B, black circles). To quantify this statistically, we performed Bayesian paired t-tests (JASP_Team, 2017) to compare the empirical $\bar{J}_{\text{total}}$ estimates at set size 3 with the estimates at set sizes 1 and 6 in the experiments that included these set sizes (DE2 and DE4-6; Table 1). These tests reveal strong evidence that total precision at set size 3 is higher than total precision at both set sizes 1 ($BF_{+0}=1.05\cdot10^{7}$) and 6 ($BF_{+0}=4.02\cdot10^{2}$). Moreover, across all six experiments, the subject-averaged set size at which $\bar{J}_{\text{total}}$ is highest in the unconstrained model is 3.52±0.18. These findings suggest that the total amount of resources that subjects allocate for stimulus encoding varies non-monotonically with set size, which is consistent with our rational model but not with previous descriptive models. To the best of our knowledge, this non-monotonic behavior has not been reported before and may be used to further constrain models of visual working memory and attention.

Alternative loss functions

To evaluate the necessity of a free parameter in the behavioral loss function, $L_{\text{behavioral}}(\varepsilon)$, we also test the following three parameter-free choices: $|\varepsilon|$, ε^2, and $-\cos(\varepsilon)$. Model comparison favors the original model with AIC differences of 14.0±2.8, 24.4±4.1, and 19.5±3.5, respectively. While there may be other parameter-free functions that give better fits, we expect that a free parameter is unavoidable here, as it is likely that the error-to-loss mapping differs across experiments (due to differences in external incentives) and possibly also across subjects within an experiment (due to differences in internal incentives). We also test a two-parameter function that was proposed recently (Eq. (5) in (C. R. Sims, 2015)). The main difference with our original choice is that this alternative function allows for saturation effects in the error-to-loss mapping. However, this extra flexibility does not increase the goodness of fit sufficiently to justify the additional parameter, as the original model outperforms this variant with an AIC difference of 5.3±1.8.

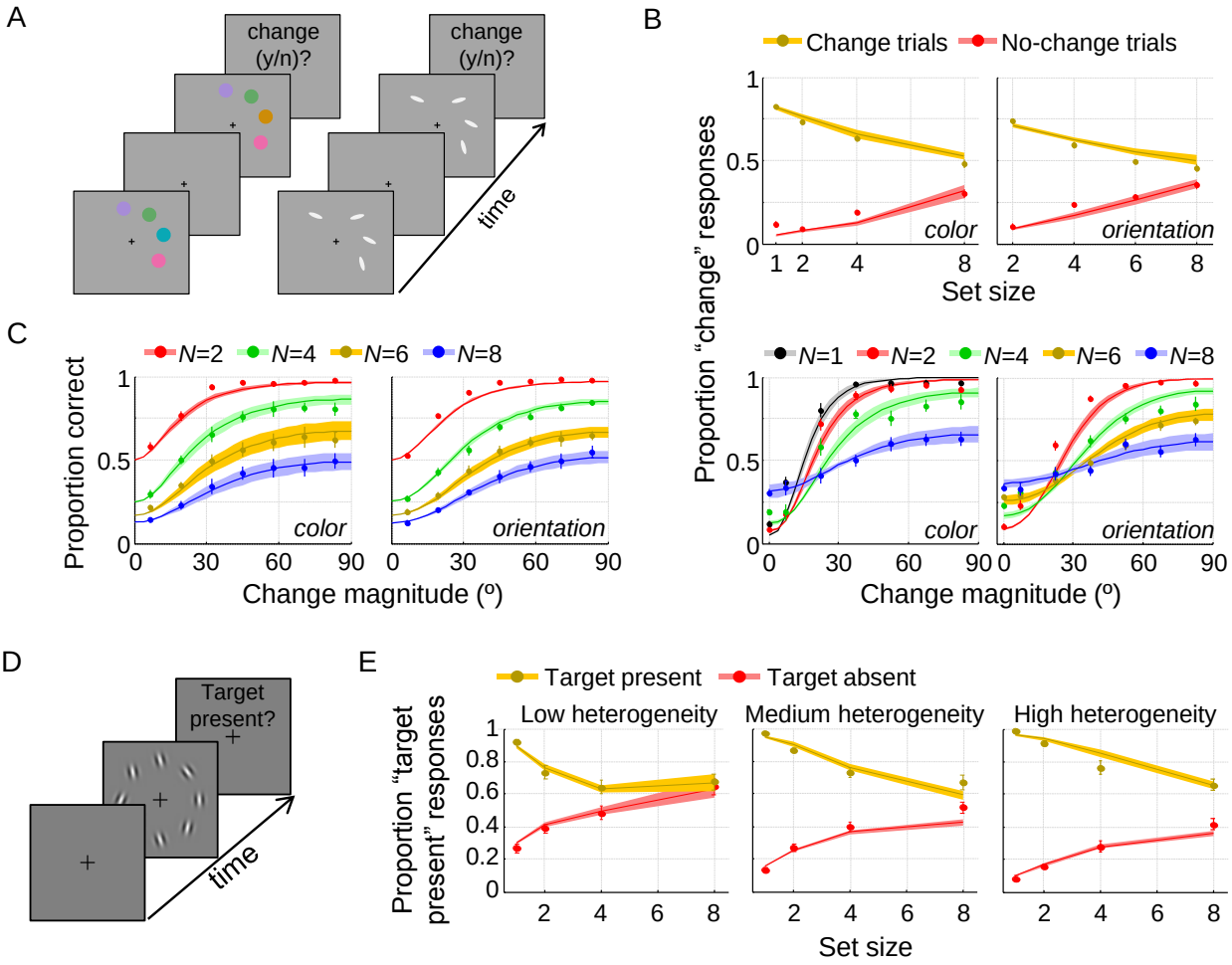


Figure 4. Model fits to three categorical decision-making tasks. (A) Experimental paradigm in the change-detection experiments. The paradigm for change localization was the same, except that a change was present on each trial and subjects reported the location of change. (B) Model fits to change-detection data. Top: hit and false alarm rates; bottom: psychometric curves. (C) Model fits to change-localization data. (D) Experimental paradigm in the visual-search experiment. (E) Model fits to visual-search data. Note that all models were fitted to raw response data, not to the summary statistics visualized here (see Methods).

Generalization to other tasks

We next examine the generality of our theory, by testing whether it can also explain set size effects in two change detection tasks (Table 1). In these experiments, the subject is on each trial sequentially presented with two sets of stimuli and reports whether there was a change at any of the stimulus locations (Fig. 4A). A change was present on half of the trials, at a random location and with a random change magnitude. The behavioral error, ε, takes only two values in this task: "correct" and "incorrect". Therefore, $p\left(\varepsilon \mid \bar{J}, N\right)$ specifies the probabilities of correct and incorrect responses for a given level of precision and set size, which depend on the observer's decision rule. Following previous work (Keshvari et al., 2012, 2013), we assume that subjects use the Bayes-

optimal rule (see Supplementary Information) and that there is random variability in encoding precision. This decision rule introduces one free parameter, p_{change}, specifying the subject's prior belief that a change will occur. Due to the binary nature of ε in this task, the free parameter of the behavioral loss function drops out of the model, as its effect is equivalent to changing parameter $\tilde{\lambda}$ (see Supplementary Information). The model thus has three free parameters ($\tilde{\lambda}$, τ, and p_{change}). We find that the maximum-likelihood fits account well for the data in both experiments (Fig. 4B).

So far, we have considered tasks with continuous and binary judgments. We next consider two change localization experiments (Table 1) in which judgments are non-binary but categorical. The task is identical to change detection, except that a change is present on every trial and the observer reports the location at which the change occurred (out of 2, 4, 6, or 8 locations). We again assume variable precision and an optimal decision rule (see Supplementary Information). Although the rational model has only two free parameters ($\tilde{\lambda}$ and τ), it accounts well for both datasets (Fig. 4C).

The final task to which we apply our theory is a visual search experiment (Mazyar et al., 2013) (Table 1). Unlike the previous three tasks, this is not a working memory task, as there was no delay period between stimulus offset and response. Set size effects in this experiment are thus likely to stem from limitations in attention rather than memory, but our theory applies without any additional assumptions. Subjects judged whether a vertical target was present among one of N briefly presented oriented ellipses (Fig. 4D). The distractors were drawn from a Von Mises distribution centered at vertical. The width of the distractor distribution determined the level of heterogeneity in the search display. Each subject was tested under three different levels of heterogeneity. We again assume variable precision and an optimal decision rule (see Supplementary Information). This decision rule has one free parameter, p_{present}, specifying the subject's prior degree of belief that a target will be present. We fit the three free parameters ($\tilde{\lambda}$, τ, and p_{present}) to the data from all three heterogeneity conditions at once and find that the model accounts well for the dependencies of the hit and false alarm rates on both set size and distractor heterogeneity (Fig. 4E).

DISCUSSION

Descriptive models of visual working memory and attention have evolved to a point where there is little room for improvement in how well they account for experimental data. However, the basic fact that encoding precision decreases with increasing set size still lacks a principled explanation.

Here, we examined a possible explanation based on normative and ecological considerations: set size effects may be the result of a rational trade-off between behavioral performance and costs induced by stimulus encoding. The models that we derived from this hypothesis account well for data across a range of quite different tasks, despite having relatively few parameters. Moreover, they account for a non-monotonicity that appears to exist between in the relation between set size and the total amount of resources that subjects allocate for stimulus encoding.

While the main purpose of our study was to make a conceptual advancement – by providing a principled theory for a phenomenon that has thus far been approached only descriptively – consideration of additional mechanisms could further improve the fits and lead to more complete models. For example, previous studies have incorporated response noise (van den Berg et al., 2014, 2012), non-target responses (Bays et al., 2009), and a (variable) limit on the number of remembered items (Dyrholm, Kyllingsbæk, Espeseth, & Bundesen, 2011; C. R. Sims et al., 2012; van den Berg et al., 2014) to improve fits. These mechanisms have not been motivated in a principled manner, but it might be possible to treat some of them using a rational approach similar to the one that we took here. For example, the level of response noise might be set by optimizing a trade-off between performance and motor control effort (Wolpert & Landy, 2012) and slot-like encoding could be a rational strategy if spreading encoding resources over multiple items incurs a metabolic loss, as has been suggested by previous work (Scalf & Beck, 2010).

More broadly, our work speaks to the relation between descriptive and rational theories in psychology and neuroscience. The main motivation for rational theories is to reach a deeper level of understanding by analyzing a system in the context of the ecological needs and constraints that it evolved under. Besides the large literature on ideal-observer decision rules (Geisler, 2011; Green & Swets, 1966; Körding, 2007; Shen & Ma, 2016), rational approaches have been used to explain properties of receptive fields (Liu, Stevens, & Sharpee, 2009; Olshausen & Field, 1996; Vincent, Baddeley, Troscianko, & Gilchrist, 2005), tuning curves (Attneave, 1954; Barlow, 1961; Ganguli & Simoncelli, 2010), neural wiring (Cherniak, 1994; Chklovskii, Schikorski, & Stevens, 2002), and neural network modularity (Clune, Mouret, & Lipson, 2013). A transition from descriptive to rational explanations might be an essential step in the maturation of theories of biological systems, and in psychology there certainly seems to be more room for this kind of explanation.

An alternative explanation of set size effects has been that the brain is unable to keep neural representations of multiple items segregated from one another (Endress & Szabó, 2017; Nairne, 1990; Oberauer & Lin, 2017; A E Orhan & Ma, 2015; Z. Wei, Wang, & Wang, 2012): as the

number of encoded items increases, so does the level of interference in their representations, resulting in lower task performance. However, these models offer no principled justification for the existence of interference and some require additional mechanisms to account for set size effects; for example, the model by Oberauer and colleagues requires three additional components – including a set-size dependent level of background noise – to fully account for set size effects (Oberauer & Lin, 2017). That being said, we do not deny there may be interference effects in working memory and adding them to models we presented here may improve their goodness of fit.

Our approach shares both similarities and differences with the concept of bounded rationality (Simon, 1957), which states that human behavior is guided by mechanisms that provide "good enough" solutions rather than optimal ones. The main similarity is that both approaches acknowledge that human behavior is constrained by various cognitive limitations. However, an important difference is that in the theory of bounded rationality, these limitations are postulates or axioms, while our approach explains them as rational outcomes of ecological optimization processes. This suggestion that cognitive limitations are subject to optimization instead of fixed may also have implications for theories outside the field of psychology. In the theory of "rational inattention" in behavioral economics, agents make optimal decisions under the assumption that there is a fixed limit on the total amount of attention that they can allocate to process economic data (C. A. Sims, 2003). This fixed-attention assumption is similar to the fixed-resource assumption in models of visual working memory and it could be interesting to explore the possibility that the amount of allocable attention is the outcome of a trade-off between expected economic performance and the expected cost induced by allocating attention to process economic data.

While our results show that set size effects can in principle be explained as the result of an optimization strategy, they do not necessarily imply that encoding precision is fully optimized on every trial in any given task. First, encoding precision in the brain most likely has an upper limit, due to irreducible sources of noise such as Johnson noise and Poisson shot noise (Faisal, Selen, & Wolpert, 2008; Smith, 2015), as well as suboptimalities early in sensory processing (Beck, Ma, Pitkow, Latham, & Pouget, 2012). This prohibits the brain from reaching the near-perfect performance levels that our model predicts when the behavioral loss associated to errors is huge. Second, precision might have a lower limit: task-irrelevant stimuli are sometimes automatically encoded (Shin & Ma, 2016; Yi, Woodman, Widders, Marois, & Chun, 2004), perhaps because in natural environments few stimuli are ever completely irrelevant. This would prevent subjects from sometimes encoding nothing at all, in contradiction to what our theory predicts to happen at very

large set sizes. Third, all models that we tested incorporated variability in encoding precision. Part of this variability is possibly due to stochastic factors such as neural noise, but part of it may also be systematic in nature (e.g., particular colors and orientations may be encoded with higher precision than others (Bae, Allred, Wilson, & Flombaum, 2014; Girshick, Landy, & Simoncelli, 2011)). Whereas the systematic component could have a rational basis (e.g., higher precision for colors and orientations that occur more frequently in natural scenes (Ganguli & Simoncelli, 2010; X.-X. Wei & Stocker, 2015)), this is unlikely to be true for the random component. Indeed, when we jointly optimize $\bar{J}$ and τ, we find estimates of τ that are consistently 0, meaning that any variability in encoding precision is suboptimal from the perspective of our model. Finally, even if set size effects are the result of a rational trade-off between behavioral and neural loss, it may be that the solution that the brain settled on works well on average, but is not tailored to provide an optimal solution in every possible situation. In that case, set size effects could be more rigid across environmental changes (e.g., in task or reward structure) than predicted by a model that incorporates every such change in a fully optimal manner.

One way to assess the plausibility and generality of a model is by examining whether variations in parameters map in a meaningful way to variations in experimental methods. Unfortunately, this approach was not possible here, because both the subject populations and experimental methods varied on a considerable number of dimensions across experiments, including stimulus time and contrast, delay time, instructions, scoring function, and the type and amount of reward. More controlled studies could be performed to further evaluate our theory, by varying a specific experimental factor that is expected to affect one of the loss functions, while keeping all other factors the same. For example, one way to manipulate the behavioral loss function would be to impose an explicit scoring function and vary this function across conditions while keeping all other factors constant. Interestingly, a previous study that performed such a manipulation in a delayed-estimation experiment found a behavioral effect in one experiment (Zhang & Luck, 2011), but unfortunately they did not vary set size. Another way to manipulate the behavioral loss function in working memory tasks is to use a cue to indicate which item is most likely going to be probed. Previous studies that used this manipulation (Bays, 2014; Klyszejko, Rahmati, & Curtis, 2014) found increased encoding precision in cued items compared to uncued items, consistent with an ideal observer strategy. It would be interesting to examine whether our model can quantitatively account for such data. Moreover, an intuitive argument suggests that our theory predicts set size effects on the cued item to become weaker as a function of cue validity. At minimum cue validity –

which is equivalent to using no cue, as in the experiments analyzed in this paper – our model predicts a decline of encoding precision with set size. At maximum validity, however, the loss-minimizing strategy is obviously to always encode the cued item with the level of precision that would be optimal for set size 1, thus entirely eliminating a set size effect. Our model makes precise quantitative predictions about this transition from strong set size effects at low cue validity to no set size effects at maximum cue validity. Moreover, the predicted set size effects are likely to differ between the cued and uncued items, which could be tested using the same experiment.

A seemingly obvious way to experimentally manipulate the neural loss function would be to vary the delay period. However, the neural mechanisms underlying working memory maintenance are still debated, which makes it difficult to derive model predictions for this manipulation. One possibility is that working memories are maintained in persistent activity (Funahashi, Bruce, & Goldman-Rakic, 1989; Fuster & Alexander, 1971), in which case it would be reasonable to assume that the neural cost related to maintenance increases linearly with delay time. If there is no initial cost associated to creating a memory, then a doubling of delay time should have the same effect as a doubling of set size. However, if there is an initial cost on top of the maintenance cost, then the effect of increasing delay period will be milder, especially if the initial cost is high. Moreover, it is has been argued that working memories may be maintained by increasing residual calcium levels at presynaptic terminals, which temporarily enhances synaptic strength and avoids the need for enhanced spiking (Mongillo, Barak, & Tsodyks, 2008). This way, an increase in delay time would induce little extra cost and our theory would predict only a mild effect of delay time on encoding precision, even in the absence of an initial cost. A recent study that varied delay period in a delayed-estimation task (Pertzov, Manohar, & Husain, 2017) indeed found only modest effects of delay time on estimation error. However, given the uncertainties about the relation between maintenance time and total neural cost, it would be premature to draw strong conclusions from this finding.

Developmental work has shown that working memory capacity estimates change with age (Simmering, 2012; Simmering & Perone, 2012). Viewed from the perspective of our proposed theory, this raises the question why the optimal trade-off between behavioral and neural loss would change with age. A speculative answer could be that a subject's encoding efficiency (formalized by parameter α in Eq. (2)) may improve during childhood. An increase in encoding efficiency (i.e., lower α) has the same effect in our model as a decrease in the set size (i.e., higher N), which we know is accompanied by an increase in optimal encoding precision. Hence, our model would

predict subjects to increase encoding precision over time, which is qualitatively consistent with the findings of the developmental studies.

Finally, our results raise the question what neural mechanisms could implement the kind of near-optimal resource allocation strategy that is the core of our theory. Some form of divisive normalization (Bays, 2014; Carandini & Heeger, 2012) would be a likely candidate, which is already a key operation in neural models of attention (Reynolds & Heeger, 2009) and visual working memory (Bays, 2014; Z. Wei et al., 2012). The essence of this mechanism is that it lowers the gain when set size is larger, without requiring knowledge of the set size prior to the presentation of the stimuli.

METHODS

Data and code sharing

All data analyzed in this paper and model fitting code are available at [url to be inserted].

Model fitting

Delayed estimation. We used Matlab's fminsearch function to find the parameter vector $\boldsymbol{\theta} = \left\{\tilde{\lambda}, \beta, \tau\right\}$ that maximizes the log likelihood function, $\sum_{i=1}^{n} \log p\left(\varepsilon_i \mid N_i, \boldsymbol{\theta}\right)$, where n is the number of trials in the subject's data set, ε_i the estimation error on the i^{th} trial, and N_i the set size on that trial. To reduce the risk of converging into a local maximum, initial parameter estimates were chosen based on a coarse grid search over a large range of parameter values. The predicted estimation error distribution for a given parameter vector $\boldsymbol{\theta}$ was computed as follows. First, $\bar{J}_{\text{optimal}}$ was computed by applying Matlab's fminsearch function to Eq. (5). In this process, the integrals over ε and J were approximated numerically by discretizing the distributions of these variables into 100 and 20 equal-probability bins, respectively. Next, the gamma distribution over precision with mean $\bar{J}_{\text{optimal}}$ and scale parameter τ was discretized into 20 equal-probability bins. Thereafter, the predicted estimation error distribution was computed under the central value of each bin. Finally, these 20 predicted distributions were averaged. We verified that our results are robust under changes in the number of bins used in the numerical approximations.

Change detection. Model fitting in the change detection task consisted of finding parameter vector $\boldsymbol{\theta} = \left\{ \tilde{\lambda}, \tau, \ p_{change} \right\}$ that maximizes $\sum_{i=1}^{n} \log p\left(R_i \mid \Delta_i, N_i, \boldsymbol{\theta} \right)$, where n is the number of trials in the subject's data set, R_i is the response ("change" or "no change"), Δ_i the magnitude of change, and N_i the set size on the i^{th} trial. For computational convenience, Δ was discretized into 30 equally spaced bins. To find the maximum-likelihood parameters, we first created a table with predicted probabilities of "change" responses for a large range of ($\bar{J}$, τ, p_{change}) triplets. One such table was created for each possible (Δ, N) pair. Each value $p(R$="change" $\mid N, \Delta, \bar{J}, \tau, p_{\text{change}})$ in these tables was approximated using the optimal decision rule (see Supplementary Information) applied to 10,000 Monte Carlo samples. Next, for a given set of parameter values, the log likelihood of each trial response was computed in two steps. First, the expected total loss was computed as a function of $\bar{J}$, using $\bar{L}_{\text{total}}\left(\bar{J}, N\right) = p_{\text{incorrect}}\left(\bar{J}, N\right) + \tilde{\lambda}\bar{J}N$, where $p_{\text{incorrect}}\left(\bar{J}, N\right)$ was estimated using the pre-computed tables. Second, we looked up $\log p(R_i \mid N_i, \Delta_i, \bar{J}_{\text{optimal}}, \tau, p_{\text{change}})$ from the pre-computed tables, where $\bar{J}_{\text{optimal}}$ is the value of $\bar{J}$ for which expected total loss was lowest. To estimate the best-fitting parameters, we performed a grid search over a large set of parameter combinations, separately for each subject.

Change localization and visual search. Model fitting methods for the change-localization and visual-search tasks were identical to the methods for the change-detection task, except for differences in the parameter vectors (no prior in the change localization task; p_{present} instead of p_{change} in visual search) and the optimal decision rules (see Supplementary Information).

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6). https://doi.org/10.1109/TAC.1974.1100705

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193. https://doi.org/10.1037/h0054663

Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism : Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, *21*(10), 1133–1145. https://doi.org/10.1097/00004647-200110000-00001

494 Bae, G., Allred, S. R., Wilson, C., & Flombaum, J. I. (2014). Stimulus-specific variability in color
495 working memory with delayed estimation. *Journal of Vision*, *14*(4), 1–23.
496 https://doi.org/10.1167/14.4.7.doi
497 Baldassi, S., & Simoncini, C. (2011). Reward sharpens orientation coding independently of
498 attention. *Frontiers in Neuroscience*, (FEB). https://doi.org/10.3389/fnins.2011.00013
499 Barlow, H. B. H. (1961). Possible principles underlying the transformation of sensory messages. In
500 *Sensory Communication* (pp. 217–234). https://doi.org/10.1080/15459620490885644
501 Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *The*
502 *Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(10), 3632–
503 45. https://doi.org/10.1523/JNEUROSCI.3204-13.2014
504 Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set
505 by allocation of a shared resource. *Journal of Vision*, *9*(10), 7.1-11.
506 https://doi.org/10.1167/9.10.7
507 Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human
508 vision. *Science*, *321*(5890), 851–4. https://doi.org/10.1126/science.1158023
509 Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not Noisy, Just Wrong:
510 The Role of Suboptimal Inference in Behavioral Variability. *Neuron*.
511 https://doi.org/10.1016/j.neuron.2012.03.016
512 Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. a. (2013). Visual long-term memory has
513 the same limit on fidelity as visual working memory. *Psychological Science*, *24*(6), 981–90.
514 https://doi.org/10.1177/0956797612465439
515 Carandini, M., & Heeger, D. (2012). Normalization as a canonical neural computation. *Nature*
516 *Reviews Neuroscience*, (November), 1–12. https://doi.org/10.1038/nrn3136
517 Cherniak, C. (1994). Component placement optimization in the brain. *The Journal of*
518 *Neuroscience : The Official Journal of the Society for Neuroscience*, *14*(April), 2418–2427.
519 https://doi.org/10.1016/S0166-2236(96)84416-X
520 Chklovskii, D. B., Schikorski, T., & Stevens, C. F. (2002). Wiring optimization in cortical circuits.
521 *Neuron*. https://doi.org/10.1016/S0896-6273(02)00679-7
522 Christie, S. T., & Schrater, P. (2015). Cognitive cost as dynamic allocation of energetic resources.
523 *Frontiers in Neuroscience*, *9*(JUL). https://doi.org/10.3389/fnins.2015.00289
524 Clune, J., Mouret, J.-B., & Lipson, H. (2013). The evolutionary origins of modularity. *Proceedings*
525 *of the Royal Society B: Biological Sciences*, *280*(1755), 20122863–20122863.

https://doi.org/10.1098/rspb.2012.2863

Cover, T. M., & Thomas, J. A. (2005). *Elements of Information Theory*. *Elements of Information Theory*. https://doi.org/10.1002/047174882X

Della Libera, C., & Chelazzi, L. (2006). Visual selective attention and the effects of monetary rewards. *Psychological Science : A Journal of the American Psychological Society / APS*, *17*(3), 222–227. https://doi.org/10.1111/j.1467-9280.2006.01689.x

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

Donkin, C., Kary, A., Tahir, F., & Taylor, R. (2016). Resources masquerading as slots: Flexible allocation of visual working memory. *Cognitive Psychology*, *85*, 30–42. https://doi.org/10.1016/j.cogpsych.2016.01.002

Dyrholm, M., Kyllingsbæk, S., Espeseth, T., & Bundesen, C. (2011). Generalizing parametric models by introducing trial-by-trial parameter variability: The case of TVA. *Journal of Mathematical Psychology*, *55*(6), 416–429. https://doi.org/10.1016/j.jmp.2011.08.005

Elmore, L. C., Ji Ma, W., 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. https://doi.org/10.1016/j.cub.2011.04.031

Endress, A., & Szabó, S. (2017). Interference and memory capacity limitations. *Psychological Review*, *In press*.

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews. Neuroscience*, *9*, 292–303. https://doi.org/10.1038/nrn2258

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229. https://doi.org/10.1038/ncomms2237

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. https://doi.org/10.1016/j.neuron.2012.12.039

Fuster, J. M., & Alexander, G. E. (1971). Neuron Activity Related to Short-Term Memory. *Science*. https://doi.org/10.1126/science.173.3997.652

Ganguli, D., & Simoncelli, E. P. (2010). Implicit encoding of prior probabilities in optimal neural populations. *Advances in Neural Information Processing Systems*, *2010*(December 2010), 658–666. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4209846&tool=pmcentrez&rendert

ype=abstract

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*. https://doi.org/10.1016/j.visres.2010.09.027

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. https://doi.org/10.1038/nn.2831

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *Society*, *1*, 521. https://doi.org/10.1901/jeab.1969.12-475

JASP_Team. (2017). JASP (Version 0.8.2) [Computer program].

Keshvari, S., van den Berg, R., & Ma, W. J. (2012). Probabilistic computation in human perception under variability in encoding precision. *PLoS ONE*, *7*(6).

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No Evidence for an Item Limit in Change Detection. *PLoS Computational Biology*, *9*(2).

Klyszejko, Z., Rahmati, M., & Curtis, C. E. (2014). Attentional priority determines working memory precision. *Vision Research*, *105*, 70–76. https://doi.org/10.1016/j.visres.2014.09.002

Körding, K. (2007). Decision theory: what "should" the nervous system do? *Science (New York, N.Y.)*, *318*, 606–610. https://doi.org/10.1126/science.1142998

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857. https://doi.org/10.1073/pnas.1101328108

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*(6), 493–497. https://doi.org/10.1016/S0960-9822(03)00135-0

Lindsay, P. H., Taylor, M. M., & Forbes, S. M. (1968). Attention and multidimensional discrimination. *Perception & Psychophysics1*, *4*(2), 113–117.

Liu, Y. S., Stevens, C. F., & Sharpee, T. (2009). Predictable irregularities in retinal receptive fields. *Proceedings of the National Academy of Sciences*, *106*(38), 16499–16504. https://doi.org/10.1073/pnas.0908926106

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

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. https://doi.org/10.1167/9.11.3

590 Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature*
591 *Neuroscience*, *17*(3), 347–56. https://doi.org/10.1038/nn.3655
592 Mankiw, N. G. (2004). *Principles of economics*. *Book* (Vol. 328).
593 https://doi.org/10.1017/CBO9780511511455
594 Mazyar, H., van den Berg, R., & Ma, W. J. (2012). Does precision decrease with set size? *Journal*
595 *of Vision*, *12*(6), 10. https://doi.org/10.1167/12.6.10
596 Mazyar, H., Van den Berg, R., Seilheimer, R. L., & Ma, W. J. (2013). Independence is elusive : Set
597 size effects on encoding precision in visual search. *Journal of Vision*, *13*(5), 1–14.
598 https://doi.org/10.1167/13.5.8.doi
599 Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*,
600 *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769
601 Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269.
602 https://doi.org/10.3758/BF03213879
603 Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What Limits Working Memory
604 Capacity? *Psychological Bulletin*, *142*(March), 758–799. https://doi.org/10.1037/bul0000046
605 Oberauer, K., & Lin, H. (2017). An Interference Model of Visual Working Memory. *Psychological*
606 *Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044
607 Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by
608 learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.
609 https://doi.org/10.1038/381607a0
610 Orhan, A. E., & Ma, W. J. (2015). Neural Population Coding of Multiple Stimuli. *Journal of*
611 *Neuroscience*, *35*(9), 3825–3841. https://doi.org/10.1523/JNEUROSCI.4097-14.2015
612 Orhan, A. E., Sims, C. R., Jacobs, R. A., Knill, D. C., & Orhan, E. (2014). The adaptive nature of
613 visual working memory. *Current Directions in Psychological Science*, *23*(3), 164–170.
614 https://doi.org/10.1177/0963721414529144
615 Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal*
616 *of Experimental Psychology. Human Perception and Performance*, *16*(2), 332–350.
617 https://doi.org/10.1037/0096-1523.16.2.332
618 Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the
619 stimulus for simple tasks. *Vision Research*, *34*(13). https://doi.org/10.1016/0042-
620 6989(94)90128-7
621 Paradiso, M. a. (1988). A theory for the use of visual orientation information which exploits the

columnar structure of striate cortex. *Biological Cybernetics*, *58*(1), 35–49. https://doi.org/10.1007/BF00363954

Peck, C. J., Jangraw, D. C., Suzuki, M., Efem, R., & Gottlieb, J. (2009). Reward modulates attention independently of action value in posterior parietal cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(36), 11182–11191. https://doi.org/10.1523/JNEUROSCI.1929-09.2009

Pertzov, Y., Manohar, S., & Husain, M. (2017). Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(4), 528–536. https://doi.org/10.1037/xlm0000328

Pestilli, F., & Carrasco, M. (2005). Attention enhances contrast sensitivity at cued and impairs it at uncued locations. *Vision Research*, *45*(14), 1867–1875. https://doi.org/10.1016/j.visres.2005.01.019

Reynolds, J. H., & Heeger, D. J. (2009). The Normalization Model of Attention. *Neuron*. https://doi.org/10.1016/j.neuron.2009.01.002

Scalf, P. E., & Beck, D. M. (2010). Competition in visual cortex impedes attention to multiple items. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(1), 161–169. https://doi.org/10.1523/JNEUROSCI.4207-09.2010

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proc.Natl.Acad.Sci.*, *90*(22), 10749–10753. https://doi.org/10.1073/pnas.90.22.10749

Sewell, D. K., Lilburn, S. D., & Smith, P. L. (2014). An information capacity limitation of visual short-term memory. *J Exp Psychol Hum Percept Perform*, *40*(6), 2214–2242. https://doi.org/10.1037/a0037744

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

Shen, S., & Ma, W. J. (2016). A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological Review*, *123*(4), 452–480. https://doi.org/10.1037/rev0000028

Shin, H., & Ma, W. J. (2016). Crowdsourced single-trial probes of visual working memory for irrelevant features Laboratory experiments. *Journal of Vision*, *16(5)*(10), 1–8. https://doi.org/10.1167/16.5.10.doi

Simmering, V. R. (2012). The development of visual working memory capacity during early childhood. *Journal of Experimental Child Psychology*, *111*(4), 695–707.

https://doi.org/10.1016/j.jecp.2011.10.007

Simmering, V. R., & Perone, S. (2012). Working memory capacity as a dynamic process. *Frontiers in Psychology*, *3*(January), 567. https://doi.org/10.3389/fpsyg.2012.00567

Simon, H. A. (1957). *Models of Man (Book). Operations Research* (Vol. 5). https://doi.org/10.2307/2281884

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690. https://doi.org/10.1016/S0304-3932(03)00029-1

Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision*, *15*(3), 2. https://doi.org/10.1167/15.3.2.doi

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807–30. https://doi.org/10.1037/a0029856

Smith, P. L. (2015). The Poisson shot noise model of visual short-term memory and choice response time: Normalized coding by neural population size. *Journal of Mathematical Psychology*, *66*, 41–52. https://doi.org/10.1016/j.jmp.2015.03.007

Sterling, P., & Laughlin, S. (2015). *Principles of neural design.* MIT Press.

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124–49. https://doi.org/10.1037/a0035234

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*. https://doi.org/10.1073/pnas.1117465109

Vincent, B. T., Baddeley, R. J., Troscianko, T., & Gilchrist, I. D. (2005). Is the early visual system optimised to be energy efficient? *Network*, *16*, 175–190. https://doi.org/10.1080/09548980500290047

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain "anti-Bayesian" percepts. *Nature Neuroscience*, *18*(10), 1509–1517. https://doi.org/10.1038/nn.4105

Wei, Z., Wang, X.-J., & Wang, D.-H. (2012). From Distributed Resources to Limited Slots in Multiple-Item Working Memory: A Spiking Network Model with Normalization. *The Journal of Neuroscience*, *32*(33), 11228–11240. https://doi.org/10.1523/JNEUROSCI.0735-12.2012

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–35. https://doi.org/10:1167/4.12.11

686 Wolpert, D. M., & Landy, M. S. (2012). Motor control is decision-making. *Current Opinion in*
687 *Neurobiology*. https://doi.org/10.1016/j.conb.2012.05.003
688 Yi, D.-J., Woodman, G. F., Widders, D., Marois, R., & Chun, M. M. (2004). Neural fate of ignored
689 stimuli: dissociable effects of perceptual and working memory load. *Nature Neuroscience*,
690 *7*(9), 992–996. https://doi.org/10.1038/nn1294
691 Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working
692 memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860
693 Zhang, W., & Luck, S. J. (2011). The Number and Quality of Representations in Working Memory.
694 *Psychological Science*, *22*(11), 1434–1441. https://doi.org/10.1177/0956797611417006
695
696

SUPPLEMENTARY INFORMATION

A rational theory of set size effects in working memory and attention

Ronald van den Berg & Wei Ji Ma

Contents

MODEL DETAILS

Relation between *J* and κ

We measure encoding precision as Fisher Information, denoted *J*. As derived in earlier work (Keshvari, van den Berg, & Ma, 2012), the mapping between *J* and the concentration parameter κ of a Von Mises encoding noise distribution is $J(\kappa) = \kappa \frac{I_1(\kappa)}{I_0(\kappa)}$, where I_1 is the modified Bessel function of the first kind of order 1. Larger values of *J* map to larger values of κ, corresponding to narrower noise distributions.

Variable precision

In all our models, we incorporated variability in precision (Fougnie, Suchow, & Alvarez, 2012; van den Berg, Shin, Chou, George, & Ma, 2012) by drawing the precision for each encoded item independently from a Gamma distribution with mean $\bar{J}$ and scale parameter τ. We denote the

725 distribution of a single precision value by $p\left(J \mid \bar{J}, \tau\right)$ and the joint distribution of the precision

726 values of all *N* items in a display by $p\left(\mathbf{J} \mid \bar{J}, \tau\right) = \prod_{i=1}^{N} p\left(J_i \mid \bar{J}, \tau\right)$.

727

728 **Expected behavioral loss function by task**

729 As a consequence of variability in precision, computation of expected behavioral loss requires

730 integration over both the behavioral error, ε, and the vector with precision values, **J**,

731
$$\bar{L}_{\text{behavioral}}\left(\bar{J}, N\right) = \begin{cases} \sum_{\varepsilon} \int_{0}^{\infty} L_{\text{behavioral}}\left(\varepsilon\right) p\left(\varepsilon \mid \mathbf{J}, N\right) p\left(\mathbf{J} \mid \bar{J}, \tau\right) d\mathbf{J} & \text{if } \varepsilon \text{ is discrete} \\ \int_{0}^{\infty}\int_{0}^{\infty} L_{\text{behavioral}}\left(\varepsilon\right) p\left(\varepsilon \mid \mathbf{J}, N\right) p\left(\mathbf{J} \mid \bar{J}, \tau\right) d\mathbf{J} d\varepsilon & \text{if } \varepsilon \text{ is continuous} \end{cases}$$

732 The distribution of precision, $p\left(\mathbf{J} \mid \bar{J}, \tau\right)$, is the same in all models, but $L_{\text{behavioral}}(\varepsilon)$ and $p(\varepsilon|\mathbf{J},N)$ are

733 task-specific. We next specify these two components separately for each task.

734 *Delayed estimation.* In delayed estimation, the behavioral error only depends on the memory

735 representation of the target item. We assume that this representation is corrupted by Von Mises

736 noise,

737
$$p\left(\varepsilon \mid \mathbf{J}, N\right) = p\left(\varepsilon \mid J_{\text{T}}\right) = \frac{1}{2\pi I_0\left(F\left(J_{\text{T}}\right)\right)} e^{F\left(J_{\text{T}}\right)\cos\left(\varepsilon\right)},$$

738 where J_{T} is the precision of the target item and $F(.)$ maps Fisher Information to a concentration

739 parameter κ; we implement this mapping by numerically inverting the mapping specified in the

740 previous section. Furthermore, the behavioral loss function is assumed to be a power-law function

741 of the absolute estimation error, $L_{\text{behavioral}}=|\varepsilon|^{\beta}$, where $\beta>0$ is a free parameter.

742 *Change detection.* We assume that subjects report "change present" whenever the posterior

743 ratio for a change exceeds 1,

744
$$\frac{p\left(\text{change present} \mid \mathbf{x}, \mathbf{y}\right)}{p\left(\text{change absent} \mid \mathbf{x}, \mathbf{y}\right)} > 1,$$

745 where **x** and **y** denote the vectors of noisy measurements of the stimuli in the first and second

746 displays, respectively. Under the Von Mises assumption, this rule evaluates to (Keshvari, van den

747 Berg, & Ma, 2013)

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

where p_{change} is a free parameter representing the subject's prior belief that a change will occur, and $\kappa_{\text{x},i}$ and $\kappa_{\text{y},i}$ denote the concentration parameters of the Von Mises distributions associated with the observations of the stimuli at the i^{th} location in the first and second displays, respectively.

The behavioral error, ε, takes only two values in this task: correct and incorrect. We assume that observers map each of these values to a loss value,

$$L_{\text{behavioral}}\left(\varepsilon\right)=\begin{cases}L_{\text{incorrect}} & \text{if } \varepsilon \text{ is "incorrect"}\\ L_{\text{correct}} & \text{if } \varepsilon \text{ is "correct".}\end{cases}$$

For example, an observer might assign a loss of 0 to any correct decision and a loss of 1 to any incorrect decision. The expected behavioral loss is a weighted sum of $L_{\text{incorrect}}$ and L_{correct},

$$\bar{L}_{\text{behavioral}}\left(\bar{J},N\right)=p_{\text{correct}}\left(\bar{J},N\right)L_{\text{correct}}+\left(1-p_{\text{correct}}\left(\bar{J},N\right)\right)L_{\text{incorrect}},$$

where $p_{\text{correct}}\left(\bar{J},N\right)$ is the probability of a correct decision. This probability is not analytic, but can be easily be approximated using Monte Carlo simulations.

Change localization. Expected behavioral loss is computed in the same way as in the change-detection task, except that a different decision rule must be used to compute $p_{\text{correct}}\left(\bar{J},N\right)$. As shown in earlier work (van den Berg et al., 2012), the Bayes-optimal rule for the change-localization task is to report the location that maximizes $\frac{I_0\left(\kappa_{\text{x},i}\right)I_0\left(\kappa_{\text{y},i}\right)}{I_0\left(\sqrt{\kappa_{\text{x},i}^2+\kappa_{\text{x},i}^2+2\kappa_{\text{x},i}\kappa_{\text{y},i}\cos\left(y_i-x_i\right)}\right)}$, where all terms are defined in the same way as in the model for the change-detection task.

Visual search. The expected behavioral loss in the model for visual search is also computed in the same way as in the model for change detection, again with the only difference being the decision rule used to compute $p_{\text{correct}}\left(\bar{J},N\right)$. The Bayes-optimal rule for this task is to report "target

present" when $\frac{p_{\text{present}}}{1-p_{\text{present}}}\frac{I_0\left(\kappa_{\text{D}}\right)e^{\kappa_i\cos\left(x_i-s_{\text{T}}\right)}}{I_0\left(\sqrt{\kappa_i^2+\kappa_{\text{D}}^2+2\kappa_i\kappa_{\text{D}}\cos\left(x_i-s_{\text{T}}\right)}\right)}$, where p_{present} is the subject's prior belief that the target will be present, κ_{D} the concentration parameter of the distribution from which the distractors are drawn, κ_{i} the concentration parameter of the noise distribution associated to the stimulus at location i, x_{i} the noisy observation of the stimulus at location i, and s_{T} the value of the target (see (Mazyar, Van den Berg, Seilheimer, & Ma, 2013) for a derivation).

The behavioral loss function drops out when the behavioral error is binary

When the behavioral error ε takes only two values, the behavioral loss can also take only two values. The integral in the expected behavioral loss (Eq (2) in the main text) then simplifies to a sum of two terms,

$$\begin{aligned}\bar{L}_{\text{behavioral}}\left(\bar{J},N\right) &= p_{\text{correct}}\left(\bar{J},N\right)L_{\text{correct}}+\left(1-p_{\text{correct}}\left(\bar{J},N\right)\right)L_{\text{incorrect}}\\ &= p_{\text{correct}}\left(\bar{J},N\right)\left(L_{\text{correct}}-L_{\text{incorrect}}\right)+L_{\text{incorrect}}.\end{aligned}$$

The optimal (loss-minimizing) value of $\bar{J}$ is then

$$\begin{aligned}\bar{J}_{\text{optimal}}\left(N\right) &= \underset{\bar{J}}{\text{argmin}}\left[p_{\text{correct}}\left(\bar{J},N\right)\left(L_{\text{correct}}-L_{\text{incorrect}}\right)+L_{\text{incorrect}}+\tilde{\lambda}\bar{L}_{\text{neural}}\left(\bar{J},N\right)\right]\\ &= \underset{\bar{J}}{\text{argmin}}\left[p_{\text{correct}}\left(\bar{J},N\right)\Delta_{\text{L}}+\tilde{\lambda}\bar{L}_{\text{neural}}\left(\bar{J},N\right)\right],\end{aligned}$$

where $\Delta_{\text{L}} \equiv L_{\text{correct}} - L_{\text{incorrect}}$. Since Δ_{L} and $\tilde{\lambda}$ have interchangeable effects on $\bar{J}_{\text{optimal}}$, we fix Δ_{L} to 1 and fit only $\tilde{\lambda}$ as a free parameter.

Conditions under which optimal precision declines with set size

In this section, we show that when the expected behavioral loss is independent of set size (as in single-probe delayed estimation and change detection), the rational model predicts optimal precision to decline with set size whenever the following four conditions are satisfied:

1) The expected behavioral loss is a strictly decreasing function of encoding precision, i.e., an increase in precision results in an increase in behavioral performance.

798 2) The expected behavioral loss is subject to a law of diminishing returns (Mankiw, 2004): the
799 behavioral benefit obtained from a unit increase in precision decreases with precision. This
800 law will hold when condition 1 holds and the loss function is bounded from below, which is
801 generally the case as errors cannot be negative.
802 3) The expected neural loss is an increasing function of encoding precision.
803 4) The expected neural loss per unit of precision is a non-decreasing function of precision. On
804 the premise that precision is proportional to spike rate (Paradiso, 1988; Seung &
805 Sompolinsky, 1993), this condition is satisfied if loss per spike increases with spike rate,
806 which has been found to be the case (Sterling & Laughlin, 2015).

807 These conditions translate to the following constraints on the first and second derivatives of the
808 expected loss functions,

809
$$1.\ \bar{L}'_{\text{behavioral}}\left(\bar{J}\right) < 0$$
$$2.\ \bar{L}''_{\text{behavioral}}\left(\bar{J}\right) > 0$$
$$3.\ \bar{L}'_{\text{neural}}\left(\bar{J}\right) > 0$$
$$4.\ \bar{L}''_{\text{neural}}\left(\bar{J}\right) \geq 0.$$

810 The loss-minimizing value of precision is found by setting the derivative of the expected total loss
811 function to 0,
812
813
$$0 = \bar{L}'_{\text{total}}\left(\bar{J}\right) = \bar{L}'_{\text{behavioral}}\left(\bar{J}\right) + \tilde{\lambda} N \bar{L}'_{\text{neural}}\left(\bar{J}\right),$$
814
815 which is equivalent to
816
$$-\frac{\bar{L}'_{\text{behavioral}}\left(\bar{J}\right)}{\bar{L}'_{\text{neural}}\left(\bar{J}\right)} = \tilde{\lambda} N. \qquad \text{(S5)}$$
817
818 The left-hand side is strictly positive for any $\bar{J}$, because of constraints 1 and 3 above. In addition, it
819 is a strictly decreasing function of $\bar{J}$, because
820
821
$$-\frac{d}{d\bar{J}}\frac{\bar{L}'_{\text{behavioral}}\left(\bar{J}\right)}{\bar{L}'_{\text{neural}}\left(\bar{J}\right)} = \frac{\bar{L}''_{\text{behavioral}}\left(\bar{J}\right)\bar{L}'_{\text{neural}}\left(\bar{J}\right) - \bar{L}'_{\text{behavioral}}\left(\bar{J}\right)\bar{L}''_{\text{neural}}\left(\bar{J}\right)}{\left(\bar{L}'_{\text{neural}}\left(\bar{J}\right)\right)^2}$$

is necessarily greater than 0 due to the four constraints specified above. As illustrated in Supplementary Figure S2, Eq. (S5) can be interpreted as the intersection point between the function specified by the left-hand side (solid curve) and a flat line at a value $\tilde{\lambda}N$ (dashed lines). The value of $\bar{J}$ at which this intersection occurs (i.e., $\bar{J}_{\text{optimal}}$) necessarily decreases with *N*.

Hence, in tasks where the expected behavioral loss is independent of set size, our model predicts a decline of precision with set size whenever the above four, rather general conditions hold. When expected behavioral loss does depend on set size (such as in whole-array change detection or change localization), the proof above does not apply and we were not able to extend the proof to this domain. (Anderson & Awh, 2012) (Anderson, Vogel, & Awh, 2011) (Rademaker, Tredway, & Tong, 2012)

PARAMETER ESTIMATES

Table S1. Subject-averaged parameter estimates of the rational model fitted to data from 11 previously published experiments. See Table 1 in main text for details about the experiments and references to the papers in which the experiments were originally published.

Experiment	$\tilde{\lambda}$	τ	β	p_{change}	p_{present}
DE1	$(4.48 \pm 0.66)\cdot 10^{-3}$	16.7±2.1	1.97±0.43		
DE2	$(2.66 \pm 0.28)\cdot 10^{-3}$	12.6±1.2	$(7.1\pm1.2)\cdot 10^{-2}$		
DE3	$(3.47 \pm 0.34)\cdot 10^{-3}$	17.9±2.1	0.228±0.040		
DE4	$(4.5 \pm 1.0)\cdot 10^{-3}$	7.3±1.7	0.176±0.064		
DE5	$(4.15 \pm 0.37)\cdot 10^{-3}$	18.5±3.0	0.426±0.089		
DE6	$(4.89 \pm 0.57)\cdot 10^{-3}$	7.6±1.4	0.66±0.14		
CD1	$(5.01 \pm 0.98)\cdot 10^{-3}$	6.2±1.7		0.517±0.007	
CD2	$(2.21 \pm 0.57)\cdot 10^{-3}$	27.6±5.6		0.526±0.009	
CL1	$(5.93 \pm 0.45)\cdot 10^{-3}$	20.9±1.8			
CL2	$(3.64 \pm 0.69)\cdot 10^{-3}$	55±12			
VS	$(3.01 \pm 0.04)\cdot 10^{-1}$	155±42			0.509±0.004

SUPPLEMENTARY REFERENCES

Anderson, D. E., & Awh, E. (2012). The plateau in mnemonic resolution across large set sizes indicates discrete resource limits in visual working memory. *Attention, Perception & Psychophysics*, *74*(5), 891–910. https://doi.org/10.3758/s13414-012-0292-1

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), 1128–38. https://doi.org/10.1523/JNEUROSCI.4125-10.2011

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working

memory. *Nature Communications*, *3*, 1229. https://doi.org/10.1038/ncomms2237

Keshvari, S., van den Berg, R., & Ma, W. J. (2012). Probabilistic computation in human perception under variability in encoding precision. *PLoS ONE*, *7*(6).

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No Evidence for an Item Limit in Change Detection. *PLoS Computational Biology*, *9*(2).

Mankiw, N. G. (2004). *Principles of economics*. *Book* (Vol. 328). https://doi.org/10.1017/CBO9780511511455

Mazyar, H., Van den Berg, R., Seilheimer, R. L., & Ma, W. J. (2013). Independence is elusive : Set size effects on encoding precision in visual search. *Journal of Vision*, *13*(5), 1–14. https://doi.org/10.1167/13.5.8.doi

Paradiso, M. a. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics*, *58*(1), 35–49. https://doi.org/10.1007/BF00363954

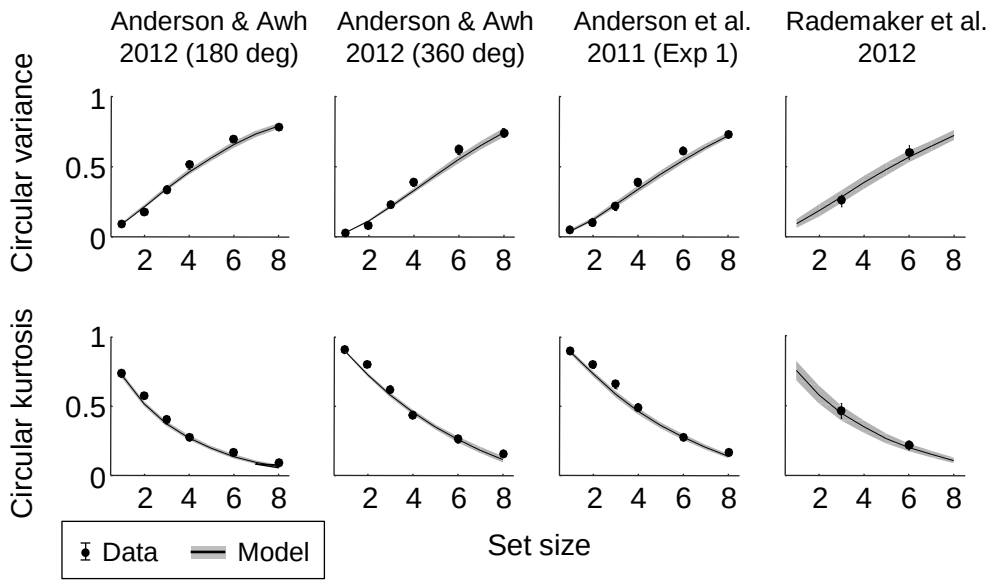
Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13), 21. https://doi.org/10.1167/12.13.21

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proc.Natl.Acad.Sci.*, *90*(22), 10749–10753. https://doi.org/10.1073/pnas.90.22.10749

Sterling, P., & Laughlin, S. (2015). *Principles of neural design.* MIT Press.

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*. https://doi.org/10.1073/pnas.1117465109

SUPPLEMENTARY FIGURES



Supplementary figure S1. Fits to the three delayed-estimation benchmark data sets that were excluded from the main analyses. Circular variance (top) and circular kurtosis (bottom) of the estimation error distributions as a function of set size, split by experiment. Error bars and shaded areas represent 1 s.e.m. of the mean across subjects. The first three datasets were excluded from the main analyses on the ground that they were published in papers that were later retracted (Anderson & Awh, 2012; Anderson et al. 2011). The Rademaker et al. (2012) dataset was excluded from the main analyses because it contains only two set sizes, which makes it less suitable for a fine-grained study of the relationship between encoding precision and set size.

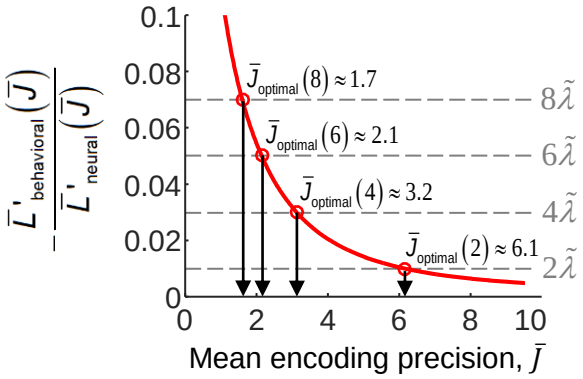


Figure S2. Graphical illustration of Eq. (S1). The value of $\bar{J}$ at which the equality described by Eq. (S1) holds is the intersection point between the function specified by the left-hand side (red curve) and a flat line at a value $N\lambda$. Since the left-hand side is strictly positive and also a strictly decreasing function of $\bar{J}$, the value at which this intersection occurs (i.e., $\bar{J}_{\text{optimal}}$) necessarily decreases with N.