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

Classification

SOCIAL SCIENCES – Psychological and Cognitive Sciences

Keywords

Working memory, attention, ecological rationality, models

Characters (with spaces): ~39,000

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

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 some kind of encoding resource, such as samples, spikes, slots, or bits. 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, for example 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 an 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. Our results suggest that set size effects have a rational basis and that ecological costs should be considered in models of human behavior.

Cognitive performance 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 (1–8). By contrast, set size effects seem to be absent in long-term memory, where fidelity has been found to be independent of set size (9). The existence of set size effects is thus not a general property of neural coding, but rather 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 postulate that stimuli are encoded using a fixed total amount of resources, formalized as "samples" (1, 3, 10), slots (11), information bit rate (12), Fisher information (8), or neural firing (13): 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. There are two problems with this explanation. Firstly, they predict that precision is inversely proportional to set size, which is often not the case (e.g., (14–16)). Secondly, it is unclear whether keeping the amount of encoding resources constant across set sizes serves any ecologically relevant function: why has the brain not evolved to counter set size effects by increasing the amount of allocated resource as more items are encoded, as seems to be the case in

long-term memory? To address the first problem, models have been proposed in which encoding precision is postulated to be a power-law function of set size (5, 6, 15–21). These models tend to provide excellent fits, but have been criticized for lacking a principled motivation for the postulated power-law (22, 23), thus failing to address the second problem.

Here, we take an ecological approach to explain set size effects, starting from the principle that neural firing is energetically costly (24–26). This cost may have pressured the brain to balance behavioral benefits of high precision against the neural loss that it induces (8, 25, 27, 28). What level of encoding precision establishes a good balance might depend on multiple factors, such as set size, task, and motivation. Indeed, performance on perceptual decision-making tasks can be improved by increasing monetary reward (29–31), which suggests that the total amount of resource spent on encoding has some flexibility that is driven by ecological factors. Based on these considerations, we hypothesize that set size effects on encoding precision reflect an ecologically rational strategy that balances behavioral performance against neural costs. Below, we derive formal models from this hypothesis for four visual working memory and attention tasks, fit them to data from eleven previously published experiments, and discuss implications of our findings.

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	(6)	Delayed estimation	Color	1, 2, 4, 8	15
DE2	(11)	Delayed estimation	Color	1, 2, 3, 6	8
DE3	(17)	Delayed estimation	Color	1, 2, 4, 6	12
DE4	(18)	Delayed estimation	Orientation	1-8	6
DE5	(18)	Delayed estimation	Color	1-8	13
DE6	(18)	Delayed estimation	Color	1-8	13
CD1	(14)	Change detection	Color	1, 2, 4, 8	7
CD2	(14)	Change detection	Orientation	2, 4, 6, 8	10
CL1	(18)	Change localization	Color	2, 4, 6, 8	7
CL2	(18)	Change localization	Orientation	2, 4, 6, 8	11
VS	(4)	Visual search	Orientation	1, 2, 4, 8	6

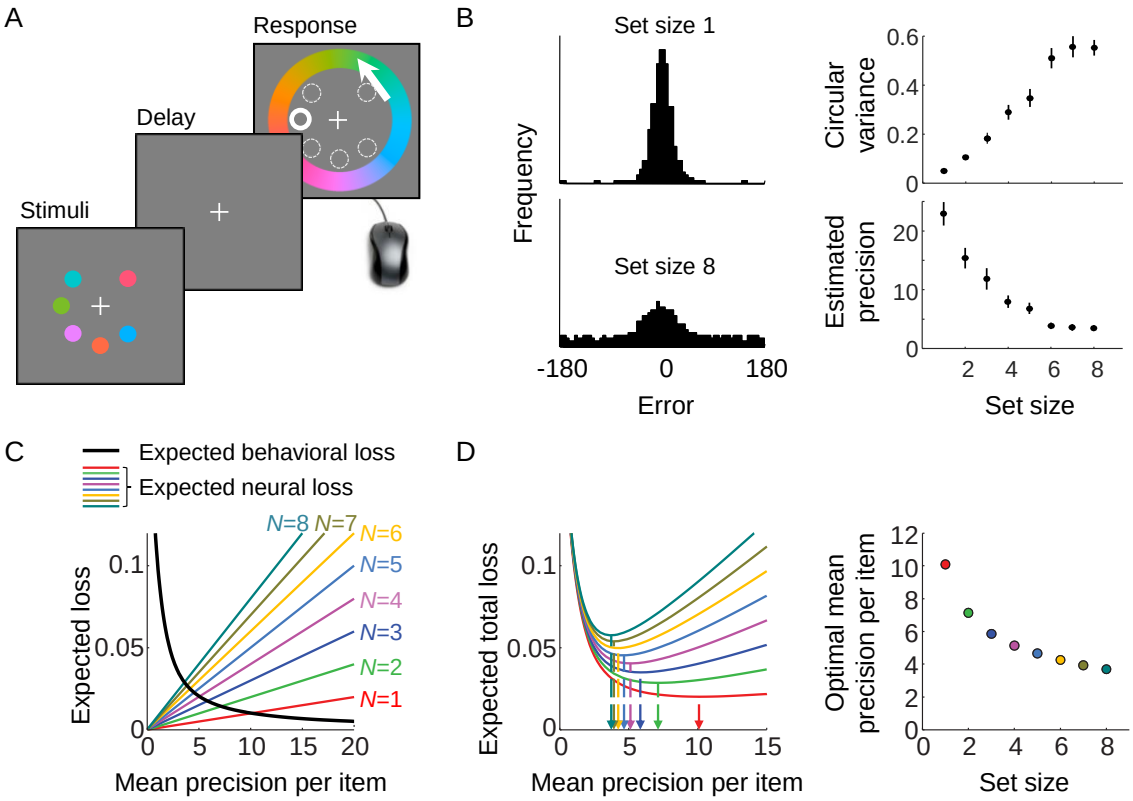

Figure 1. An ecologically rational model of set size effects in delayed estimation. (A) Example of a delayed-estimation experiment. The subject is briefly presented with a set of stimuli and, after a short delay, reports the value of a randomly chosen target item. (B) Estimation error distributions widen 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. 2C). (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 λ=0.01, β=2, and $\tau \downarrow 0$.

THEORY

We first present our theory in the context of the delayed-estimation paradigm (6) and will later show how it generalizes 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 are visible as a widening of the estimation error distribution (Fig. 1B). As in previous work (4, 5, 14, 15, 18, 32), 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 (33), 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 (5, 15, 18, 32, 34), which we model using a gamma distribution with a mean $\bar{J}$ and a scale parameter τ (see Supplementary Information).

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,

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

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).

The second kind of loss is a *neural loss* induced by the neural spiking activity that represents a stimulus. For many choices of spike variability, including the common one of Poisson-like variability (35), the precision (Fisher information) of a stimulus encoded in a neural population is proportional to the neural spiking rate (36, 37). Moreover, it has been estimated that the energetic loss induced by each spike increases with spiking rate (24, 25). When combining these two premises, the expected neural loss associated with the encoding of an item is a supralinear function of encoding precision, which can be modeled using for example a power-law function, $\bar{L}_{\text{neural}}\left(\bar{J}\right) = \alpha \bar{J}^{\omega}$, with α and ω as free parameters. However, to minimize the number of free parameters, we assume for the moment that the function is linear ($\beta = 1$) and will later present a mathematical proof that our theory generalizes to supralinear functions ($\beta > 0$; condition (iv) at the

end of this section). Further assuming that stimuli are encoded independently of each other, 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). However, it can be shown (see Supplementary Information) that the conditions under which this model predicts a set size effect generalize to any choice of loss functions, as long as the four, rather general conditions are satisfied: (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 (38): 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. Next, we evaluate the model by fitting it to data from a range of previously published experiments.

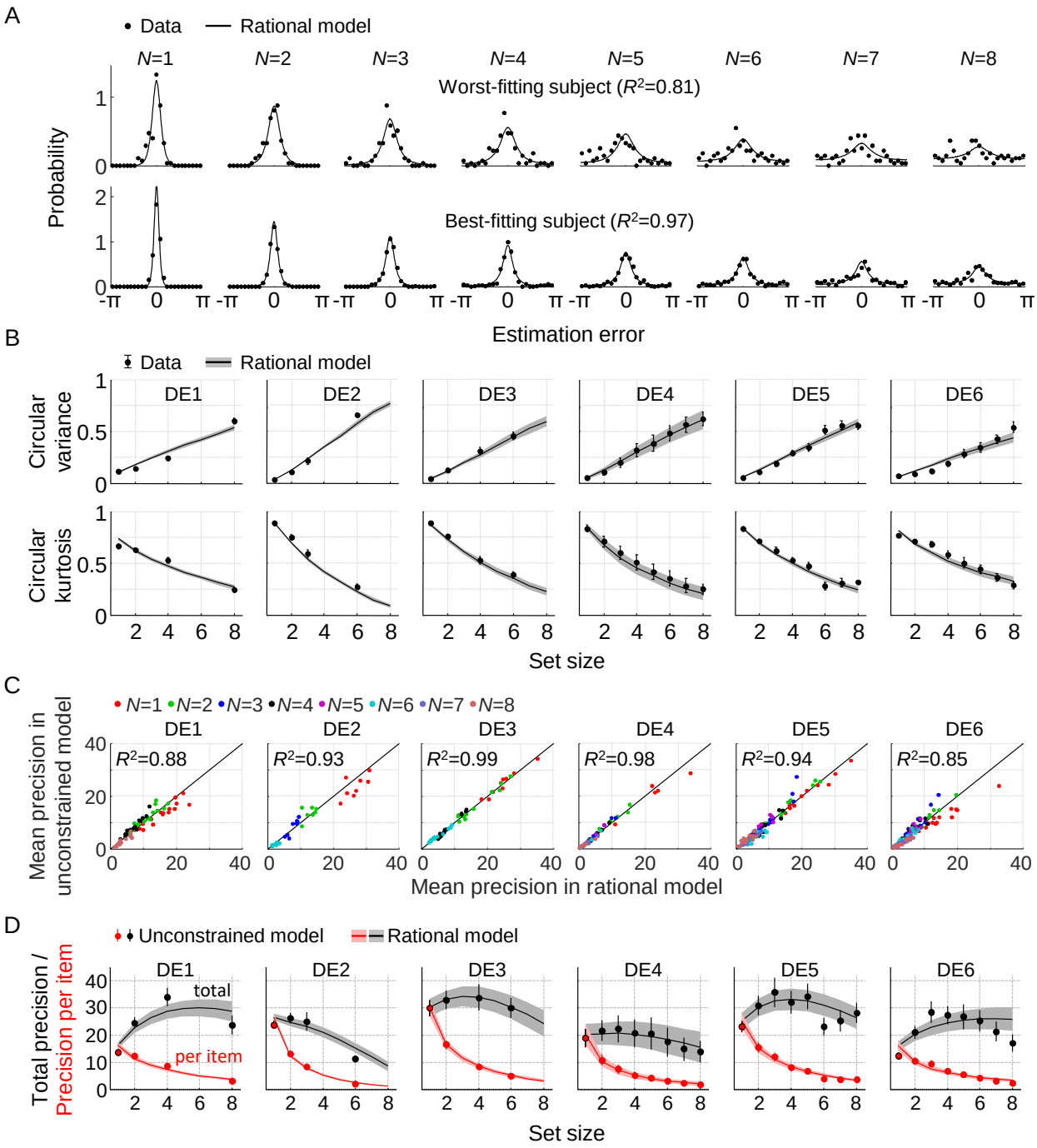

Figure 2. Model fits to data from six delayed-estimation experiments. (A) Maximum-likelihood fits to raw data of the worst-fitting and best-fitting subjects (based on R^2). (B) Subject-averaged fits to the two statistics that summarize the estimation error distributions (circular variance kurtosis) as 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. (C) Best-fitting precision values in the rational model scattered against the best-fitting precision values in the unconstrained model. Each dot represents the estimate for a single subject. (D) Estimated mean encoding precision per item (red) and total encoding precision (black) plotted against set size.

144

RESULTS

Model fits

We used maximum-likelihood estimation to fit the three model parameters ($\tilde{\lambda}$, τ, and β) to 67 individual-subject data sets from a delayed-estimation benchmark set* (Table 1). The model accounts well for the raw error distributions (Fig. 2A) and the two statistics that summarize these distributions (Fig. 2B). Model comparison based on the Akaike Information Criterion (AIC) (39) indicates that the goodness of fit is comparable to that of a descriptive model variant in which the relation between encoding precision and set size is assumed to follow a power law (ΔAIC=5.27±0.70 in favor of the rational model). Hence, the rational model provides a principled explanation of set size effects in delayed-estimation tasks without sacrificing quality of fit.

Comparison with an unconstrained model

We next try to falsify our theory by testing whether a mapping between set size and encoding precision can be found that fits the data better than the relation imposed by the loss-minimization strategy of the rational model. To this end, we fit an unconstrained variant of the model in which memory precision is fitted as a free parameter at each set size. We find only a minimal difference in goodness of fit (ΔAIC=3.49±0.93 in favor of the unconstrained model), suggesting 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. 2C). Hence, it seems that no relation exists that fits these data substantially better than the constrained set of relations that are possible in the rational model.

Total amount of allocated resource as a function of set size

We estimate the total amount of allocated encoding resource as the mean precision (Fisher information) per item summed across all items, $\bar{J}_{\text{total}} \triangleq \bar{J}N$. In fixed-resource models $\bar{J}_{\text{total}}$ is by

* The original benchmark set (15) 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.

definition constant and in power-law models it is a monotonic function of set size. However, an interesting qualitative feature of the rational model is that in some of the experiments the best-fitting parameters produce a *non-monotonic* relation between $\bar{J}_{\text{total}}$ and set size (Fig. 2D, gray curves). This means that at small set sizes it apparently sometimes pays off (in terms of total-loss minimization) to increase the total amount of allocated resource when an item is added, while the opposite is true at large set sizes[†]. Using the best-fitting precision values in the unconstrained model as an estimate of how much encoding resource subjects allocated on average at each set size, we find that the data show clear signs of a similar non-monotonicity (Fig. 2D, black circles); to our knowledge, this has not previously been reported.

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 (40)). 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.

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. 3A). A change was present on half of the trials, at a random location

[†] Upon reflection, it is perhaps not surprising to occasionally find a non-monotonic relation: when multiplying a decreasing function of set size ($\bar{J}$ as a function of N) with an increasing one (N itself), it easy to obtain a function that is not monotonic but peaks at an intermediate value.

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 (14, 32), 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 degree of 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. 3B).

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. 3C).

The final task to which we apply our theory is a visual search experiment (4) (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. 3D). 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. 3E).

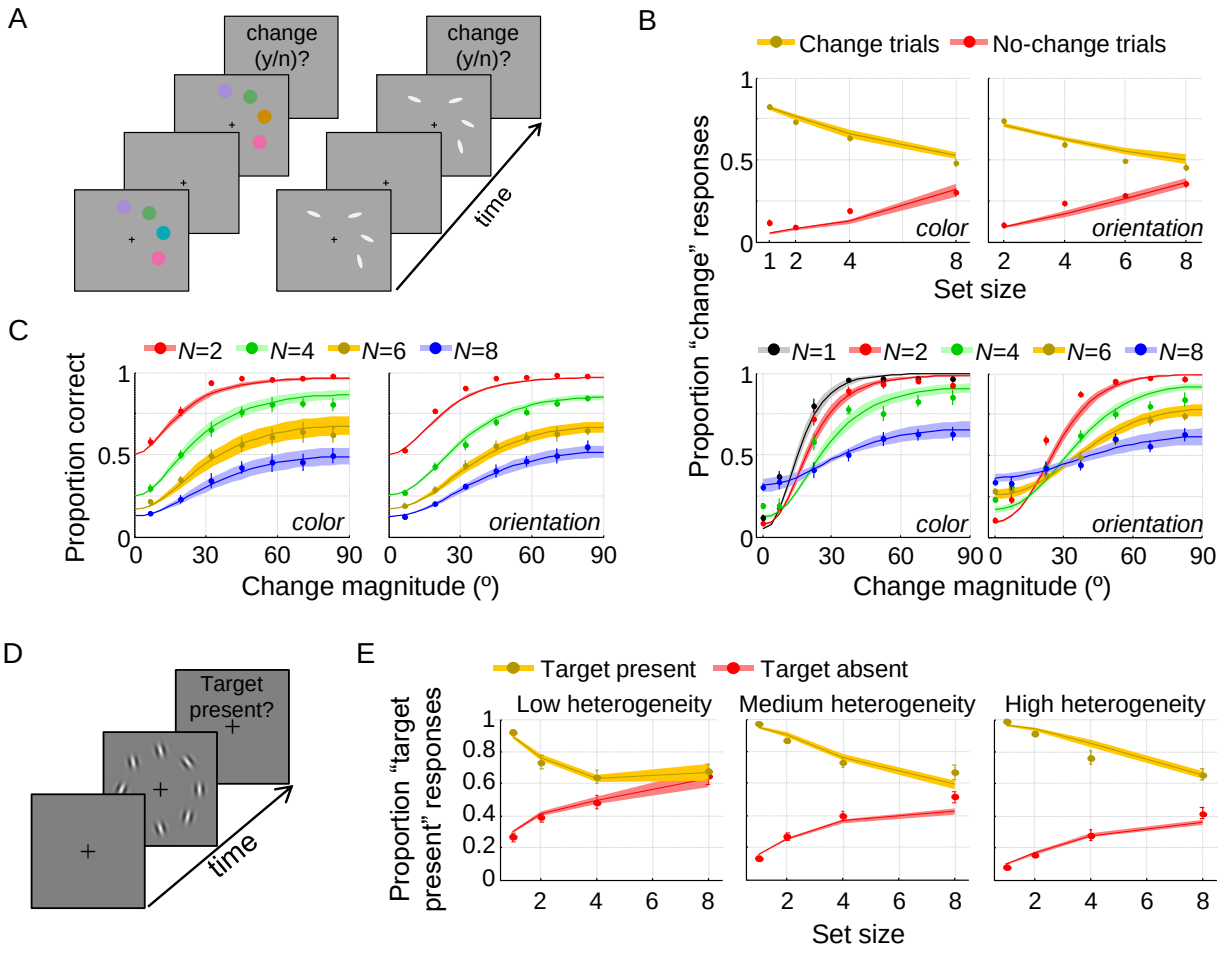

Figure 3. Model fits to three categorical decision-making tasks. (A) Experimental paradigm in the change-detection experiments. The paradigm for change localization was the same, except that a change was present on each trial and subjects reported the location of change. (B) Model fits to change-detection 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.

DISCUSSION

A key strength of our theory is that it uses a single principle of rationality and relatively few parameters to produce well-fitting models across a range of quite different tasks. Nevertheless, consideration of additional mechanisms could further improve the fits and lead to more complete models of human behavior. For example, previous studies have incorporated response noise (15, 18), non-target responses (17), and a (variable) limit on the number of remembered items (12, 15, 41) to improve fits. We did not consider such components here, as they come with additional parameters, some are task-specific (such as non-target responses), and they have so far not been

motivated in a principled manner. Regarding the latter point, it might be possible to treat some of these mechanisms using an ecologically rational approach as well. For example, the level of response noise might be set by optimizing a trade-off between performance and motor control effort (42).

Our findings suggest that set size effects in working memory and attention may reflect a near-optimal compromise reached by a system that strives to simultaneously maximize performance and minimize spiking activity. A possible explanation why long-term memory does not seem to suffer from set size effects (9) and has much larger capacity (43) is that loss incurred by maintaining synaptic connections is likely to be lower than the loss incurred by persistent activity.

The work presented in this paper 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 (44–47), rational approaches have been used to explain properties of receptive fields (48–50), tuning curves (51–53), neural wiring (54, 55), and neural network modularity (56). 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.

Although several previous models in the field of working memory and attention contain rational aspects, none of them accounts for set size effects in a principled way. Sims and colleagues have examined how errors in visual working memory can be minimized by optimally taking into account statistics of the stimulus distribution, but assume a fixed total amount of available encoding resource (12, 57). Moreover, in our own previous work on visual search (4, 5), change detection (14, 32), and change localization (18), we used optimal-observer models for the decision stage, but assumed an ad hoc power law for the encoding stage. 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 (23, 58–60): 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 (23). That being said, our theory does not rule out the possibility of interference, and it could be added onto any of the models we presented.

Our approach shares both similarities and differences with the concept of bounded rationality (61), 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 bounded rationality postulates these limitations as a given fact, while our approach explains them as rational outcomes of ecological optimization processes. The suggestion that cognitive limitations are themselves subject to optimization may also have implications for theories outside the field of psychology. One example concerns recent models of value-based decision-making that incorporate constraints imposed by working memory and attention limitations (e.g., (62)). Another example is the theory of "rational inattention" in behavioral economics, which examines optimal decision-making under the assumption that decision makers have a fixed limit on the total amount of attention that they can allocate to process economic data (63). It might be interesting to extend that theory by treating the amount of allocable attention as the outcome of an optimization process rather than a constant.

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 at 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 (64), as well as suboptimalities early in sensory processing (65). This prohibits subjects to reach the near-perfect performance levels that our model may predict when the behavioral loss associated to errors is huge. Second, precision might have a lower limit: task-irrelevant stimuli are sometimes automatically encoded (66), 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 (67, 68)). Whereas the systematic component could have a rational basis (e.g., higher precision for colors and orientations that occur more frequently in natural scenes (53, 69)), 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 form 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 only works well on average rather than being tailored to 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 fully optimal model that incorporates every such change.

Future work could further examine optimality of encoding precision in working memory and attention by experimentally varying factors that affect the loss functions. In delayed estimation, an obvious choice for this would be the delay period. Assuming that working memories are maintained in persistent activity (70, 71), a longer delay would induce a higher cost and decrease optimal encoding precision. Another experimental parameter that could be varied is the error-to-loss mapping. A previous study that performed this manipulation found a behavioral effect in one experiment, but did not vary set size (72). None of the experiments modeled here contained this manipulation (DE4-DE6 imposed an explicit loss function but did not vary it; the other experiments had no explicit scoring system). Future studies could measure effects of changes in explicitly imposed scoring systems and test how well a rational model accounts for such effects. Related to this, it would be relevant to examine whether subjects are able to internalize experimental loss functions in the timespan of a single experiment and, if not, to further characterize their "natural" loss functions (40). Another line of possible future work would be to examine whether our theory can be generalized to the level of objects (73, 74).

Developmental work has shown that working memory capacity estimates change with age (75, 76). 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 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 (13, 77) would be a likely candidate, as it has the effect of lowering the gain when set size is larger. Moreover, divisive normalization is already a key operation in neural models of attention (78) and visual working memory (13, 58).

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 $\boldsymbol{\theta} = \{\tilde{\lambda}, \beta, \tau\}$ that maximizes the log likelihood function, $\sum_{i=1}^{n} \log p(\varepsilon_i \mid N_i, \boldsymbol{\theta})$, 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 $\boldsymbol{\theta} = \{\tilde{\lambda}, \tau, p_{change}\}$ that maximizes $\sum_{i=1}^{n} \log p(R_i \mid \Delta_i, N_i, \boldsymbol{\theta})$, 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 $p(R=\text{“change”} \mid N, \Delta, \bar{J}, \tau, p_{\text{change}})$ 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 vector (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

1. Shaw ML (1980) Identifying attentional and decision-making components in information processing. *Attention and Performance VIII*, ed Nickerson RS (Erlbaum, Hillsdale, NJ, NJ), pp 277–296.
2. Ma WJ, Husain M, Bays PM (2014) Changing concepts of working memory. *Nat Neurosci* 17(3):347–56.
3. Palmer J (1990) Attentional limits on the perception and memory of visual information. *J Exp Psychol Hum Percept Perform* 16(2):332–350.
4. Mazyar H, Van den Berg R, Seilheimer RL, Ma WJ (2013) Independence is elusive : Set size effects on encoding precision in visual search. *J Vis* 13(5):1–14.
5. Mazyar H, van den Berg R, Ma WJ (2012) Does precision decrease with set size? *J Vis* 12(6):10.
6. Wilken P, Ma WJ (2004) A detection theory account of change detection. *J Vis* 4(12):1120–35.
7. Palmer J (1994) Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision Res* 34(13). doi:10.1016/0042-6989(94)90128-7.

8. Ma WJ, Huang W (2009) No capacity limit in attentional tracking: evidence for probabilistic inference under a resource constraint. *J Vis* 9(11):3.1-30.
9. Brady TF, Konkle T, Gill J, Oliva A, Alvarez G a (2013) Visual long-term memory has the same limit on fidelity as visual working memory. *Psychol Sci* 24(6):981–90.
10. Lindsay PH, Taylor MM, Forbes SM (1968) Attention and multidimensional discrimination. *Percept Psychophys* 4(2):113–117.
11. Zhang W, Luck SJ (2008) Discrete fixed-resolution representations in visual working memory. *Nature* 453(7192):233–235.
12. Sims CR, Jacobs RA, Knill DC (2012) An ideal observer analysis of visual working memory. *Psychol Rev* 119(4):807–30.
13. Bays PM (2014) Noise in neural populations accounts for errors in working memory. *J Neurosci* 34(10):3632–45.
14. Keshvari S, van den Berg R, Ma WJ (2013) No Evidence for an Item Limit in Change Detection. *PLoS Comput Biol* 9(2).
15. van den Berg R, Awh E, Ma WJ (2014) Factorial comparison of working memory models. *Psychol Rev* 121(1):124–49.
16. Bays PM, Husain M (2008) Dynamic shifts of limited working memory resources in human vision. *Science (80-)* 321(5890):851–4.
17. Bays PM, Catalao RFG, Husain M (2009) The precision of visual working memory is set by allocation of a shared resource. *J Vis* 9(10):7.1-11.
18. 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* 109:8780–8785.
19. Devkar DT, Wright AA (2015) The same type of visual working memory limitations in humans and monkeys. *J Vis* 13(2015):1–18.
20. Elmore LC, et al. (2011) Visual short-term memory compared in rhesus monkeys and humans. *Curr Biol* 21(11):975–979.
21. Donkin C, Kary A, Tahir F, Taylor R (2016) Resources masquerading as slots: Flexible allocation of visual working memory. *Cogn Psychol* 85:30–42.
22. Oberauer K, Farrell S, Jarrold C, Lewandowsky S (2016) What Limits Working Memory Capacity? *Psychol Bull* 142(March):758–799.

23. Oberauer K, Lin H (2017) An Interference Model of Visual Working Memory. *Psychol Rev* 124(1):21–59.
24. Attwell D, Laughlin SB (2001) An energy budget for signaling in the grey matter of the brain. *J Cereb Blood Flow Metab* 21(10):1133–1145.
25. Lennie P (2003) The cost of cortical computation. *Curr Biol* 13(6):493–497.
26. Sterling P, Laughlin S (2015) *Principles of neural design.* (MIT Press).
27. Pestilli F, Carrasco M (2005) Attention enhances contrast sensitivity at cued and impairs it at uncued locations. *Vision Res* 45(14):1867–1875.
28. Christie ST, Schrater P (2015) Cognitive cost as dynamic allocation of energetic resources. *Front Neurosci* 9(JUL). doi:10.3389/fnins.2015.00289.
29. Della Libera C, Chelazzi L (2006) Visual selective attention and the effects of monetary rewards. *Psychol Sci a J Am Psychol Soc / APS* 17(3):222–227.
30. Peck CJ, Jangraw DC, Suzuki M, Efem R, Gottlieb J (2009) Reward modulates attention independently of action value in posterior parietal cortex. *J Neurosci* 29(36):11182–11191.
31. Baldassi S, Simoncini C (2011) Reward sharpens orientation coding independently of attention. *Front Neurosci* (FEB). doi:10.3389/fnins.2011.00013.
32. Keshvari S, van den Berg R, Ma WJ (2012) Probabilistic computation in human perception under variability in encoding precision. *PLoS One* 7(6).
33. Cover TM, Thomas JA (2005) *Elements of Information Theory* doi:10.1002/047174882X.
34. Fougnie D, Suchow JW, Alvarez GA (2012) Variability in the quality of visual working memory. *Nat Commun* 3:1229.
35. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. *Nat Neurosci* 9(11):1432–1438.
36. Paradiso M a (1988) A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biol Cybern* 58(1):35–49.
37. Seung HS, Sompolinsky H (1993) Simple models for reading neuronal population codes. *ProcNatlAcadSci* 90(22):10749–10753.
38. Mankiw NG (2004) *Principles of economics* doi:10.1017/CBO9780511511455.
39. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr* 19(6). doi:10.1109/TAC.1974.1100705.

40. Sims CR (2015) The cost of misremembering: Inferring the loss function in visual working memory. *J Vis* 15(3):2.
41. Dyrholm M, Kyllingsbæk S, Espeseth T, Bundesen C (2011) Generalizing parametric models by introducing trial-by-trial parameter variability: The case of TVA. *J Math Psychol* 55(6):416–429.
42. Wolpert DM, Landy MS (2012) Motor control is decision-making. *Curr Opin Neurobiol* 22(6):996–1003.
43. Brady TF, Konkle T, Alvarez GA, Oliva A (2008) Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci* 105(38):14325–14329.
44. Green DM, Swets JA (1966) Signal detection theory and psychophysics. *Society* 1:521.
45. Körding K (2007) Decision theory: what "should" the nervous system do? *Science* 318:606–610.
46. Geisler WS (2011) Contributions of ideal observer theory to vision research. *Vision Res* 51(7):771–781.
47. Shen S, Ma WJ (2016) A detailed comparison of optimality and simplicity in perceptual decision making. *Psychol Rev* 123(4):452–480.
48. Vincent BT, Baddeley RJ, Troscianko T, Gilchrist ID (2005) Is the early visual system optimised to be energy efficient? *Network* 16:175–190.
49. Liu YS, Stevens CF, Sharpee T (2009) Predictable irregularities in retinal receptive fields. *Proc Natl Acad Sci* 106(38):16499–16504.
50. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609.
51. Attneave F (1954) Some informational aspects of visual perception. *Psychol Rev* 61(3):183–193.
52. Barlow HBH (1961) Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pp 217–234.
53. Ganguli D, Simoncelli EP (2010) Implicit encoding of prior probabilities in optimal neural populations. *Adv Neural Inf Process Syst* 2010(December 2010):658–666.
54. Cherniak C (1994) Component placement optimization in the brain. *J Neurosci* 14(April):2418–2427.
55. Chklovskii DB, Schikorski T, Stevens CF (2002) Wiring optimization in cortical circuits.

Neuron 34(3):341–347.

56. Clune J, Mouret J-B, Lipson H (2013) The evolutionary origins of modularity. *Proc R Soc B Biol Sci* 280(1755):20122863–20122863.
57. Sims CR (2016) Rate–distortion theory and human perception. *Cognition* 152:181–198.
58. 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. *J Neurosci* 32(33):11228–11240.
59. Orhan AE, Ma WJ (2015) Neural Population Coding of Multiple Stimuli. *J Neurosci* 35(9):3825–3841.
60. Endress A, Szabó S (2017) Interference and memory capacity limitations. *Psychol Rev* In press.
61. Simon HA (1957) *Models of Man (Book)* doi:10.2307/2281884.
62. Krajbich I, Rangel A (2011) Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proc Natl Acad Sci* 108(33):13852–13857.
63. Sims CA (2003) Implications of rational inattention. *J Monet Econ* 50(3):665–690.
64. Faisal AA, Selen LPJ, Wolpert DM (2008) Noise in the nervous system. *Nat Rev Neurosci* 9:292–303.
65. Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A (2012) Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron* 74(1):30–39.
66. Yi D-J, Woodman GF, Widders D, Marois R, Chun MM (2004) Neural fate of ignored stimuli: dissociable effects of perceptual and working memory load. *Nat Neurosci* 7(9):992–996.
67. Bae G, Allred SR, Wilson C, Flombaum JI (2014) Stimulus-specific variability in color working memory with delayed estimation. *J Vis* 14(4):1–23.
68. Girshick AR, Landy MS, Simoncelli EP (2011) Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat Neurosci* 14(7):926–932.
69. Wei X-X, Stocker AA (2015) A Bayesian observer model constrained by efficient coding can explain “anti-Bayesian” percepts. *Nat Neurosci* 18(10):1509–1517.
70. Fuster JM, Alexander GE (1971) Neuron Activity Related to Short-Term Memory. *Science (80-)* 173(3997):652–654.

71. Funahashi S, Bruce CJ, Goldman-Rakic PS (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J Neurophysiol* 61(2):331–349.
72. Zhang W, Luck SJ (2011) The Number and Quality of Representations in Working Memory. *Psychol Sci* 22(11):1434–1441.
73. Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. *Nature* 390(6657):279–281.
74. Wheeler ME, Treisman AM (2002) Binding in short-term visual memory. *J Exp Psychol Gen* 131(1):48–64.
75. Simmering VR, Perone S (2012) Working memory capacity as a dynamic process. *Front Psychol* 3(January):567.
76. Simmering VR (2012) The development of visual working memory capacity during early childhood. *J Exp Child Psychol* 111(4):695–707.
77. Carandini M, Heeger D (2012) Normalization as a canonical neural computation. *Nat Rev Neurosci* (November):1–12.
78. Reynolds JH, Heeger DJ (2009) The Normalization Model of Attention. *Neuron* 61(2):168–185.