Efficient coding of numbers explains decision bias and noise

Arthur Prat-Carrabin[1,⋆] and Michael Woodford[1]

[1]Department of Economics, Columbia University, New York, USA
⋆*email: arthur.p@columbia.edu*

Abstract

Human subjects differentially weight different stimuli in averaging tasks. This has been interpreted as reflecting biased stimulus encoding, but an alternative hypothesis is that stimuli are encoded with noise, then optimally decoded. Moreover, with efficient coding, the amount of noise should vary across stimulus space, and depend on the statistics of stimuli. We investigate these predictions through a task in which participants are asked to compare the averages of two series of numbers, each sampled from a prior distribution that differs across blocks of trials. We show that subjects encode numbers with both a bias and a noise that depend on the number. Infrequently occurring numbers are encoded with more noise. A model combining efficient coding and Bayesian decoding best captures subjects' behaviour. Our results suggest that Wei and Stocker's "law of human perception", which relates the bias and variability of sensory estimates, also applies to number cognition.

In many decision problems, someone is presented with an array of variables that must be aggregated in order to identify the optimal decision. How humans combine several sources of information in their decision-making process is a long-standing debate in economics and cognitive science [1, 2, 3, 4, 5]. Recently, a series of experimental studies have focused on averaging tasks, in which subjects are presented with several stimuli (sometimes numbers, but sometimes visual stimuli characterized by their length, orientation, shape, or color) and asked to make a decision about the *average* magnitude of the presented stimuli [6, 7, 8, 9]. Although the contribution of each stimulus to the average should, in theory, be proportional to its true magnitude, the weights attributed to stimuli by human subjects, in their decisions, appear to be nonlinear functions of their magnitudes. Subjects asked to compare the averages of two series of digits, for instance, overweight larger digits when making a decision [6]. What is the origin of this seemingly suboptimal behaviour? Refs. [6, 7] show that if comparison of the average encoded values involves *noise*, then a *nonlinear transformation* of the presented stimuli can partially compensate for the performance loss induced by the noise. The nonlinear distortion of stimuli appears, under this proposal, as a consequence of an optimal encoding strategy, given unavoidable noise at decision time.

Here, we investigate the alternative hypothesis that a presented stimulus is encoded with noise, while the decoding rule used to produce a comparative judgment is deterministic, and indeed represents an optimal inference from the noisy encoded values (rather than a comparison of the simple sums of the encoded values, plus noise). Under the assumption that the estimate of each magnitude is optimally inferred from its encoded representation [10, 11], the noisy estimate that results will generally be biased, to a degree that will vary as a function of the stimulus [12, 13, 14, 15]. Thus the average estimate will be a nonlinear function of the true stimulus magnitude, though for a different reason than in the model of Ref. [6] mentioned above. This nonlinear function will depend on how the encoding noise varies over the stimulus space. Efficient coding theories suggest that the degree of noise with which a stimulus is encoded should be a decreasing function of its probability under the prior distribution of stimuli; in other words, less likely stimuli should be encoded with more noise [16, 17, 18, 19, 20]. This implies that the way that both estimation bias and estimation noise vary over the stimulus space should depend on the prior distribution over that space; we test for such dependence in our data.

In a related model that combines efficient coding with Bayesian decoding, Wei and Stocker derive a relation between estimation noise and estimation bias, a "law of human perception" supported by evidence from numerous sensory domains [21]. We find support for this law in our data on number comparisons as well, suggesting that it applies to the semantic representation of numerosity carried by Arabic numerals, and not only to the perception of physical stimuli. We also test other implications of both efficient coding and optimal decoding.

Each of these candidate theories highlights the potential role of the prior in human decision-making. Thus we design a task in which subjects are asked to compare the averages of two series of numbers, and manipulate the prior distribution of the presented numbers across different blocks of trials: it is sometimes uniform, but sometimes skewed toward smaller or larger numbers. In each of these three conditions, the weights of numbers in the decisions of subjects appear to vary nonlinearly over the range of presented numbers. We compare the behaviour of our subjects to the predictions of a family of noisy estimation models, and find that the patterns in their responses are best captured by a model in which a nonlinear transformation of a presented number is observed with a degree of noise that itself depends on the number.

Turning to encoding-decoding models, we show how a model of efficient coding combined with Bayesian decoding accounts both for the nonlinear transformation and for the noise, and best captures subjects' behaviour. We find that the encoding noise varies depending on the prior, and that this allows our subjects to increase their expected reward in the task.

Results

We first present our average-comparison experiment. In each trial of a computer-based task, ten numbers, each within the range $[10.00, 99.99]$, and alternating in color between red and green, are presented to the subject in rapid succession. The subject is then asked to choose whether the five red numbers or the five green numbers had the higher average (Fig. 1a). In different blocks of trials, numbers are sampled from different prior distributions: the **"Uniform"** distribution, under which all numbers in the $[10.00, 99.99]$ range are equally likely, an **"Upward"** distribution (under which higher numbers are more likely), and a **"Downward"** distribution (under which lower numbers are more likely) (Fig. 1b). Within a block of trials, both red and green numbers are sampled from the same distribution. Additional details on the task are provided in Methods.

Although the presented information would allow, in principle, an unambiguous identification of the color of the numbers that have the higher average, subjects make errors in their responses. These errors are not independent of the presented numbers: the proportion of trials in which they choose 'red' (the "choice probability") is an increasing, sigmoid-like function of the difference between the average of the red numbers and that of the green numbers, with about 50% red responses when this difference is close to zero. In other words, subjects make less errors when there is a marked difference between the two averages (Fig. 1c).

In order to isolate the influence of a single number on the decision, we can compute the probability of choosing 'red' conditional on a given (red) number x being presented, $P(\text{'red'}|x)$, and the absolute difference between this probability and 0.5: $\big|P(\text{'red'}|x) - 0.5\big|$. This quantity, which we call the "decision weight" (following Ref. [6]), is a measure of the average impact of a given number on the decision. For an ideal subject who makes no errors, the decision weight should be zero for x equal to the prior mean $\bar{x}$, and should increase approximately linearly with the absolute distance $|x - \bar{x}|$ from the prior mean (Fig. 1d,e, thin lines). In our subjects' data, instead, the value of x for which the decision weight is zero is somewhat larger than $\bar{x}$ for each prior, and the decision weight increases more steeply with increases in x above this value than it does with decreases in x below that value (Fig. 1d,e, thick lines).

We wish to explore models of noisy comparison that can account for these regularities. We begin by fitting models to our data that do not separate the processes of encoding and decoding, and instead simply posit that a given number x results in a noisy estimate $\hat{x}$ drawn from a conditional distribution $p(\hat{x}|x)$. We assume that the ten numbers presented are estimated with this procedure (with independent draws), and that the color 'red' is chosen if the average of the estimates of the red numbers is greater than that of the green numbers, i.e., if $\frac{1}{5}\sum_{i=1}^{5}\hat{x}_i^R > \frac{1}{5}\sum_{i=1}^{5}\hat{x}_i^G$, where $\hat{x}_i^R$ and $\hat{x}_i^G$ are the estimates for the red and green numbers. Note that this kind of characterization of the data is equally consistent with the theory of Ref. [6], in which $\hat{x}$ is equal to a deterministic transformation of x plus a random noise term added at the time of the comparison process, and a model in which x is encoded with noise and $\hat{x}$ is then an estimate of the value of x based on the noisy encoded value.

We consider several possible assumptions about the form of the conditional distribution of estimates, $p(\hat{x}|x)$. The simplest kind of model consistent with a sigmoid curve of the kind shown in Figure 1c would be one in which the estimate is normally distributed around the true value of the number, with a constant standard deviation, s, i.e.: $\hat{x}|x \sim N(x, s^2)$. In other words, numbers are perceived with constant Gaussian noise, but the estimates are unbiased. The probability of choosing the red color, conditioned on the ten presented numbers, would then be

$$P\left(\sum_{i=1}^{5}\hat{x}_i^R > \sum_{i=1}^{5}\hat{x}_i^G \,\middle|\, x_{1:5}^R, x_{1:5}^G\right) = \Phi\left(\frac{\sum_{i=1}^{5}x_i^R - \sum_{i=1}^{5}x_i^G}{\sqrt{10s^2}}\right), \tag{1}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. The choice probability in this model is thus a sigmoid function of the difference between the averages of red and green numbers, similarly to the choice probability of subjects (Fig. 2a).

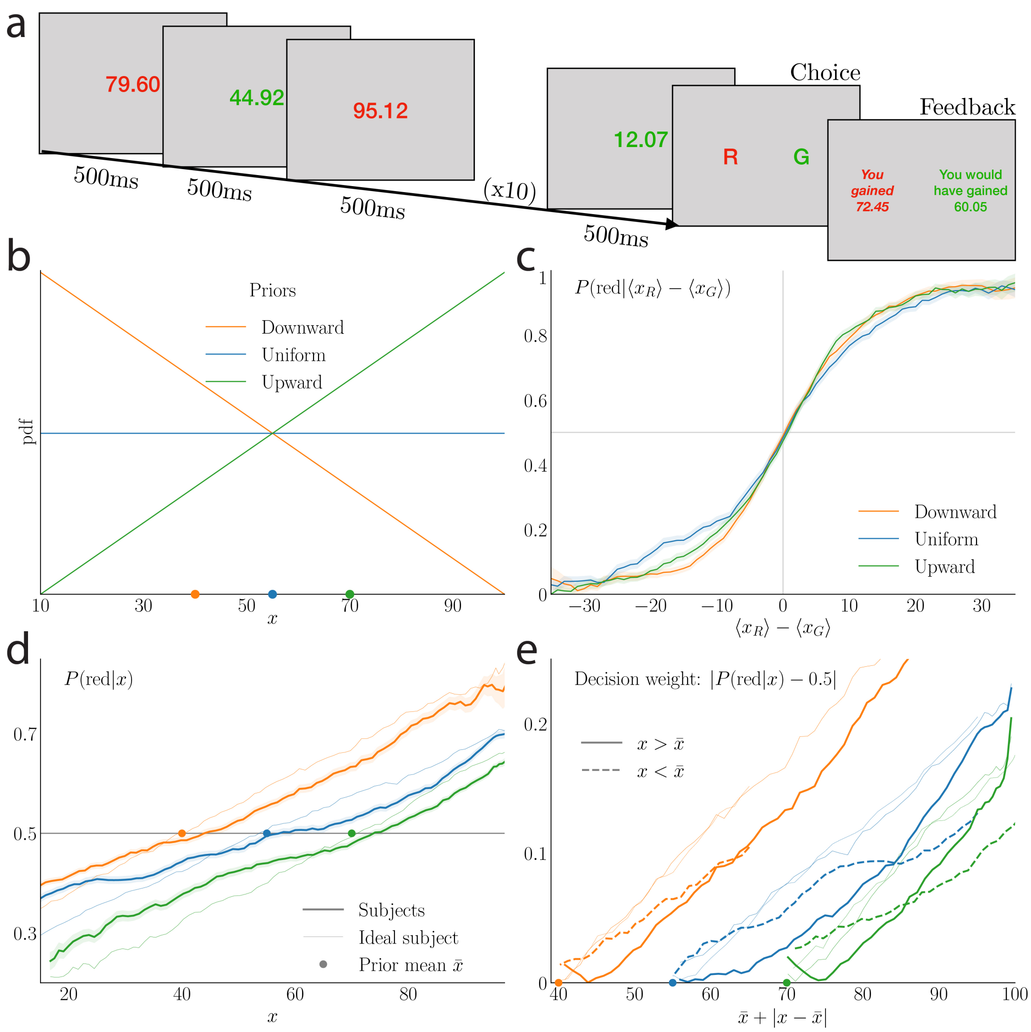

Figure 1: Illustration of the task and subjects choice behaviour. a. In each trial, ten numbers are sequentially presented, alternating in color between red and green. The subject can then choose a color, without time constraint, by pressing the corresponding key. Feedback is given, which reports the two averages and emphasizes the chosen one. **b.** In each trial, both red and green numbers are drawn from the same prior distribution over the range $[10.00, 99.99]$. In different blocks of consecutive trials, the prior is either Downward (triangular distribution with the peak at 10.00, orange line), Uniform (blue line), or Upward (triangular distribution with the peak at 99.99, green line). **c.** Choice probability: in each of the three prior conditions, the fraction of trials in which subjects choose the 'red' average, as a function of the true difference in the averages of the two series of numbers. While an ideal subject would be described by a step function, the choices of subjects result in sigmoid curves. **d.** Probability $P(\mathrm{red}|x)$ of choosing 'red' conditional on a red number x being presented, as a function of x, for the subjects (thick lines) and the ideal subject (thin lines). **e.** Decision weight, defined as $|P(\mathrm{red}|x) - 0.5|$, for the subjects (thick lines) and the ideal subject (thin lines), as a function of the sum of the prior mean, $\bar{x}$, and the absolute difference between the number and the prior mean, $|x - \bar{x}|$. Subjects' decision weights for numbers below the mean (dashed line) and above the mean (solid line) are appreciably different, whereas for the ideal subject there is only a small difference due to sampling error. In **c**, **d**, and **e**, each point of the curves is obtained by taking the average of the quantity of interest (ordinate) over a sliding window of length 10 of the quantity on the abscissa, incremented by steps of length 1. In **c** and **d**, the shaded areas represent the standard error of the mean.

However, the decision weight of a number, in this model, is an approximately linear function of the absolute difference between the number and the prior mean, and two numbers equally distant from the mean have the same weight (Fig. 2c, top left panel). Thus this simple model does not reproduce the subjects' unequal weighting of numbers in decisions documented in Figure 1e.

In the model just presented, estimates are unbiased ($\mathbb{E}\hat{x} = x$), and all numbers are estimated with the same amount of noise ($\text{Var}\hat{x} = s^2$). But we also consider models that are more flexible in either or both of these respects. For example, we can allow for bias by assuming that the estimate, $\hat{x}$, of a presented number, x, is normally distributed around a nonlinear transformation, $m(x)$, of the number: $\hat{x}|x \sim N(m(x), s^2)$. Such a model is equivalent to the one assumed in Refs. [6, 7], in which noise is added to the sum of nonlinearly transformed values of the stimuli. Alternatively, we might assume that the variability of estimates is not constant over the stimulus space, as is found to be true in many stimulus domains [14]. The simplest case of this kind would be one in which $\hat{x}|x \sim N(x, s^2(x))$, where now the estimation noise, $s(x)$, varies with the number. Our most general model combines the features of these last two, positing that a transformation of the number is observed with varying noise: $\hat{x}|x \sim N(m(x), s^2(x))$. In this model, the probability of choosing the red color, conditioned on the ten presented numbers, will be

$$P\left(\sum_{i=1}^{5} \hat{x}_i^R > \sum_{i=1}^{5} \hat{x}_i^G \middle| x_{1:5}^R, x_{1:5}^G \right) = \Phi\left(\frac{\sum_{i=1}^5 m(x_i^R) - \sum_{i=1}^5 m(x_i^G)}{\sqrt{\sum_{i=1}^5 s^2(x_i^R) + \sum_{i=1}^5 s^2(x_i^G)}} \right). \quad (2)$$

The first three models are special cases of the fourth one, with either an absence of bias (i.e., an identity transformation $m(x) = x$), a constant noise ($s(x) = s$), or both (Eq. (1)).

We fit these models to the behavioural data by maximizing their likelihoods. We flexibly specify the functions $m(x)$ and $s(x)$, by allowing them to be arbitrary low-order polynomials, defined by their values at a small number of points (between 2 and 8), with the order chosen to minimize the Bayesian Information Criterion (BIC). Running simulations of the fitted models, using the numbers presented to the subjects, we find that each of them predicts that choice frequency should be a similar sigmoid function of the difference between the two averages. The models differ, however, in the implied graphs for the decision weights. The unbiased model with varying noise ($N(x, s^2(x))$) still results in approximately linear decision weights, with equal weights for numbers at equal distances from the mean (Fig. 2c, top right panel). Instead the two models that include a transformation $m(x)$ of the presented numbers yield nonlinear decision weights resembling those of the subjects: numbers below the mean and close to it have higher weights that numbers above the mean and equally distant from it, while the opposite occurs for numbers further from the mean (Fig. 2c, bottom panels. More details on how the transformation $m(x)$ affects the decision weights can be found in Methods.) In summary, the two models which feature a transformation of the number, whether with a constant or variable noise, better capture the patterns observed in behavioural data.

In order to quantitatively compare the degree of explanatory success of the models, we compute the BIC for the best-fitting model in each class. We can either assume that the parameters of the models are identical under the three priors ("homogeneous parameters"), or, conversely, that the parameters depend on the prior ("prior-specific parameters"). The

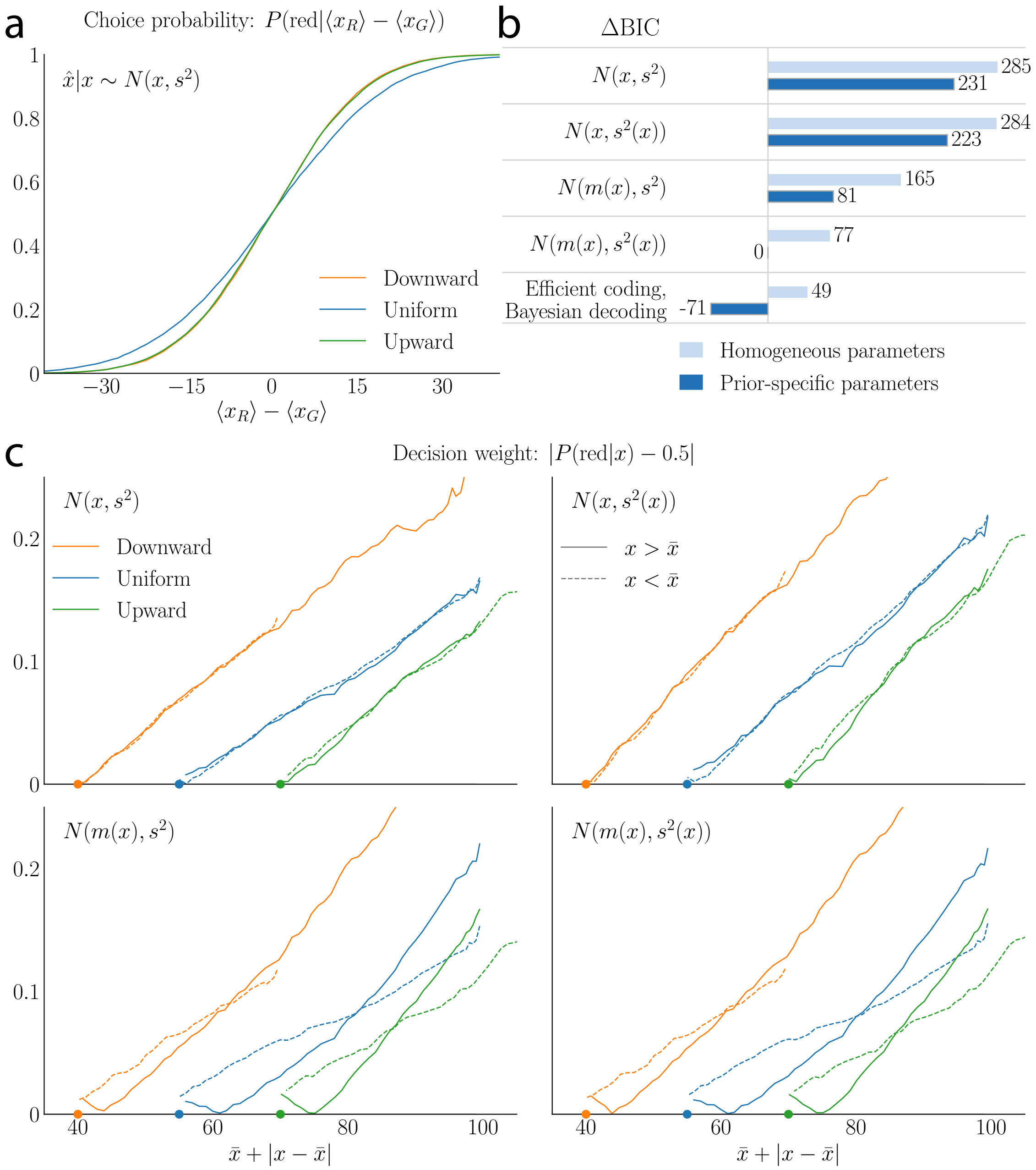

Figure 2: Models with both nonlinear transformation and varying noise best capture subjects' behaviour. a. Choice probability under the unbiased, constant-noise model ($N(x, s^2)$) as a function of the difference in the averages of the presented numbers, for the three prior conditions. **b.** Model comparison statistics for the four one-stage models of noisy estimation and for the efficient-coding, Bayesian-decoding model, both with homogeneous parameters and with prior-specific parameters. For each model class, the difference in BIC from the least restrictive model class (general transformation, variable noise, prior-specific) is reported. The lower BIC for the efficient-coding, Bayesian-decoding model (a special case of the general one-stage model) indicates that it captures subjects' behaviour more parsimoniously. **c.** Decision weights $|P(\text{red}|x) - 0.5|$ for the unbiased, constant-noise model (top left), the unbiased, varying-noise model (top right), the model with transformation of the number and constant noise (bottom left), and the model with transformation of the number and varying noise (bottom right). Only the two models with transformation of the number reproduce the behaviour of subjects shown in Fig. 1e.

latter results in a larger number of parameters, which is penalized by the BIC. In spite of this penalty, in all four models the BIC is lower with prior-specific parameters than with homogeneous parameters, suggesting that subjects' decision-making process is adapted to the prior distribution (Fig. 2b). The two best-fitting one-stage models include a transformation $m(x)$ of the presented number, in accordance with our observations about the decision weights, and the best-fitting model also features a variable noise $s(x)$. We note that the BIC in this case is significantly lower than with a transformation but maintaining constant noise ($\Delta\text{BIC} = 81$), in spite of the additional parameters of the noise function $s(x)$. In Methods we discuss the finer features of subjects' behaviour that are better captured by allowing for variable noise, showing that a nonlinear transformation alone (as in Ref. [6, 7]) does not suffice to account for our data. In addition, to investigate the robustness of our results, we conduct a random-effect analysis that takes into account the subjects' heterogeneity. We follow the Bayesian model selection procedure described in Ref. [22], and use cross-validation to estimate the models' likelihoods. This analysis, detailed in Methods, supports the resulted presented here.

The shape of the best-fitting noise function $s(x)$ for each prior is shown in Figure 3a. In each case, we find that it is U-shaped: minimal at a value close to the mean of the prior (but slightly below it), and increasing as a function of the distance from this minimum. In the Downward-prior condition, larger numbers (which have a low probability) are perceived with a much larger amount of noise than small numbers (which have a high probability). The noise function in the Upward condition approximately mirrors that in the Downward condition, with small numbers perceived with more noise than large numbers. In other words, the least likely numbers are perceived with the greatest randomness, suggesting that the process by which subjects estimate numbers is adapted to the prior distribution from which they are drawn.

The finding that the variance of $\hat{x}$ differs for different numbers x is a natural one if we suppose that, rather than summing transformed values of the individual numbers and adding noise only at the comparison stage, each individual number is encoded with noise, with the estimate $\hat{x}$ then representing an *inference* about the likely value of x based on its noisy internal representation r. In general, an optimal rule of inference will make the estimate $\hat{x}(r)$ a nonlinear function of r, so that the variance of $\hat{x}$ should depend on x even if we suppose that the variance of r does not. At the same time, an encoding-decoding model of this kind, in which decoding is assumed to be optimal given the nature of the noisy encoding, will imply that the function $m(x)$, indicating the bias in the average decoded value, will not be independent of $s(x)$. Thus this class of models represents a special case of the general specification $N(m(x), s^2(x))$, though a different restricted class than any of those considered above.

The encoding-decoding model that we consider, based on a proposal of Morais and Pillow [23], combines a noisy but efficient coding of the presented number, with a Bayesian-mean decoding of the noisy internal representation. More precisely, suppose that a presented number, x, elicits in the brain a series of n signals, $r = (r_1, \ldots, r_n)$, each drawn independently from a distribution conditioned on the presented number, $p(r_i|x)$. Suppose that this encoding

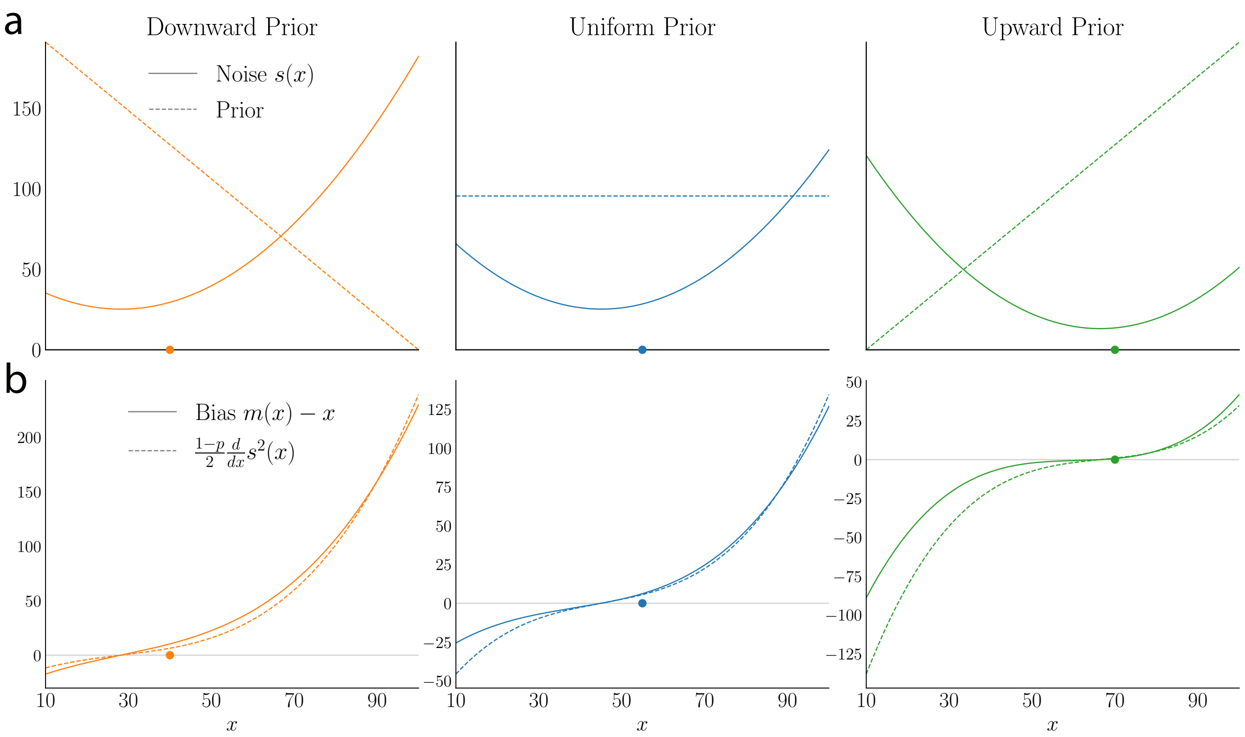

Figure 3: Best-fitting noise and bias: subjects encode less frequent numbers with greater noise. a. The best-fitting noise function $s(x)$ in the $N(m(x), s^2(x))$ model (solid line), and prior distribution (dashed line), in the Downward (left), Uniform (middle) and Upward (right) conditions. The scale refers to the noise (scale of the prior pdf not shown.) **b.** The bias (solid line) and the derivative of the noise variance (multiplied by $\frac{1-p}{2}$, dashed line) for the best-fitting $N(m(x), s^2(x))$ model, as a function of x, in the three prior conditions. The efficient-coding, Bayesian-decoding model requires these two functions to be the same (Eq. (4)).

is efficient, in the sense that the Fisher information of r, $I(x)$, solves the optimum problem

$$\min \int \tilde{\pi}(x) I(x)^{-p/2} \mathrm{d}x \qquad \text{s.t.} \quad \int \sqrt{I(x)} \mathrm{d}x \leq K, \tag{3}$$

where $p > 0$, and where $\tilde{\pi}(x)$ is a subjective prior about the distribution of numbers, which may differ from the prior actually used in the experiment, $\pi(x)$. The constraint in Eq. (3) is the same as in [24], though we consider a more general objective. The rationale for minimizing the objective assumed in Eq. (3) is that, as shown in Ref. [23], it provides a lower bound on the mean L_p error of any point estimator of x. In the limiting case $p \to 0$, minimizing this quantity is equivalent to maximizing an approximation of the mutual information between the number and its internal representation, which is the objective assumed by Wei and Stocker [24]. An objective with $p > 0$ is more plausible, however, in our setting, because in our task, larger estimation errors reduce a subject's reward to a greater extent than small errors do. The optimal Fisher information, under this more general efficient coding problem, is proportional to the prior raised to the power $\frac{2}{p+1}$.

As for the decoding, we assume that the estimate of subjects, $\hat{x}$, is the mean of the Bayesian posterior over the numbers, x, given the noisy signal, r. This decoding rule, combined with the efficient coding scheme just presented, results in a variance of the estimate equal to the inverse of the Fisher information, $1/I(x)$ (to order $1/n$), and implies a simple relation between the bias of the estimate and the derivative of its variance. We find:

$$\text{Bias} \;\equiv\; \mathbb{E}(\hat{x} - x) \;=\frac{1-p}{2}\frac{\mathrm{d}}{\mathrm{d}x}\Big(\frac{1}{I(x)}\Big). \tag{4}$$

Thus we approximate the distribution of estimates as

$$\hat{x}|x \sim N\left(x + \frac{1-p}{2}\frac{\mathrm{d}}{\mathrm{d}x}\Big(\frac{1}{I(x)}\Big),\; \frac{1}{I(x)}\right). \tag{5}$$

As in the one-stage models considered above, the estimate is normally distributed around a transformation of the number. This is a special case of the general model considered above ($N(m(x), s^2(x))$), with the added constraint that both the nonlinear transformation $m(x)$ and the noise $s(x)$ are here determined by a single function, the Fisher information $I(x)$. In particular, the efficient-coding, Bayesian-decoding model predicts that the bias, $m(x) - x$, is proportional to the derivative of the noise variance, $s^2(x) = 1/I(x)$ (Eq. (4)).

We can test whether our data are consistent with the additional restriction given by Eq. (4). In Figure 3b, we plot the functions corresponding to the two sides of this equation, implied by our best-fitting estimates of $m(x)$ and $s(x)$ for each of the three conditions, when the theoretical restrictions implied by the efficient-coding, Bayesian-decoding model are *not* imposed in the estimation. Here it is important to note that the transformation $m(x)$ can be identified from choice data only up to an arbitrary affine transformation; this gives us two free parameters to choose in plotting the left-hand side of the equation. As explained in Methods, we choose these parameters to make the implied function more similar to the right-hand side of the equation. When we do so, we obtain the functions plotted in Figure 3b. As noted above, the noise $s(x)$ is U-shaped; thus the derivative of the noise variance, $\frac{d}{dx}s^2(x)$, increases as a function of x; it vanishes at a value close to the prior mean, is negative below this value, and is positive above it (Fig. 3b, dashed line). With an appropriate choice of normalization for $m(x)$, the bias vanishes at the same value; in addition, we note that for all three priors the bias is also negative below this value and positive above it, and increases as a function of x (Fig. 3b, solid line). In other words, numbers below the prior mean are underestimated, while numbers above the prior mean are overestimated. Thus we find that the two functions, when fitted to subjects' data, are qualitatively consistent with the predicted relation (Eq. (4)).

We can also estimate directly the model of estimation bias implied by Eq. (5), finding the parameter p and the Fisher information function $I(x)$ that minimize the BIC. (Here also, we allow the noise function to be a low-order polynomial; in Methods, we consider a different functional form, closer to that used in Ref. [6], with similar results.) We assume that, regardless of the distribution of numbers, the subjects optimize the same efficient coding criterion, determined by p; thus even when fitting with prior-specific parameters, we maintain the parameter p constant across conditions. As with our one-stage models, prior-specific parameters yield a lower BIC than homogeneous ones. Furthermore, the BIC is

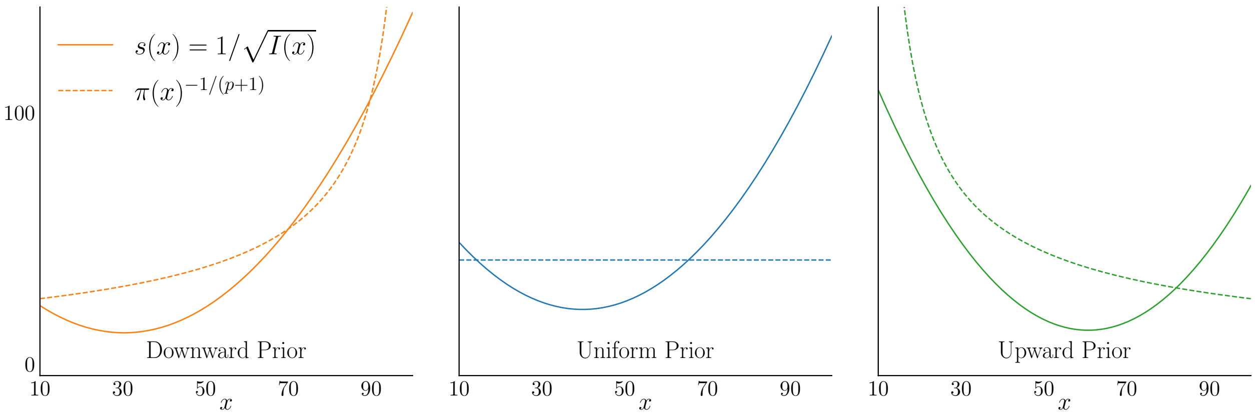

Figure 4: Efficient-coding, Bayesian-decoding model: the Fisher information, fitted to subjects' data, is adapted to the prior. Noise function, $s(x)$, equal to the inverse square-root of the Fisher information, fitted to subjects' data (solid lines), and prior $\pi(x)$ raised to the power $-1/(p+1)$, with the best-fitting value $p = 0.7$ (dashed lines), in the Downward (left), Uniform (middle) and Upward (right) conditions. The ordinate scale refers to the noise function (scale for the prior not shown).

lower than that of the model in which the functions $m(x)$ and $s(x)$ are unrestricted (Fig. 2b, $\Delta\text{BIC} = 71$). We conclude that the model implied by Eq. (5) captures more parsimoniously the behaviour of subjects.

In this model, a single function, the Fisher information, $I(x)$, or equivalently the noise function, $s(x) = 1/\sqrt{I(x)}$, together with the value of p completely determines the statistics of responses. Moreover, the assumption that the encoding is optimized, in our model (Eq. (3)), results in the prediction that the noise function should be proportional to the inverse of the subjective prior, $1/\tilde{\pi}(x)$, raised to the power $1/(p+1)$. As we do not have access to the subjective prior, we show, in Figure 4, the noise functions alongside the correct priors raised to the power $-1/(p+1)$, with p equal to its best-fitting value, 0.7. The best-fitting noise function differs under the three priors. In the Downward condition, the noise is an increasing function of the number over most of the range of presented numbers (Fig. 4, left panel). In the Upward condition, conversely, the fitted noise function decreases with the number (except at large numbers; Fig. 4, right panel). Hence, in both the Downward and Upward conditions, the noise function is higher for numbers that are less likely under the prior. In the Uniform condition, the noise reaches a minimum around 40, and large numbers are observed with somewhat more noise than small numbers (Fig. 4, middle panel).

The behavioral implications of these fitted functions are shown in Figure 5. The estimation bias is proportional to the derivative of the squared noise function (Eq. (4)): it vanishes where the noise reaches a minimum; for smaller x, it is negative, while for higher x it is positive (Fig. 5b). Consequently the efficient-coding, Bayesian decoding model reproduces the unequal weighting of numbers in subjects' decisions (Fig. 5c,d) and the sigmoid shape of the choice probability curve (Fig. 5e).

Our estimates best fit subjects' data (even penalizing the additional free parameters) when the Fisher information is allowed to differ depending on the prior distribution from which numbers are sampled. Context-dependent encoding of this kind is predicted by theories

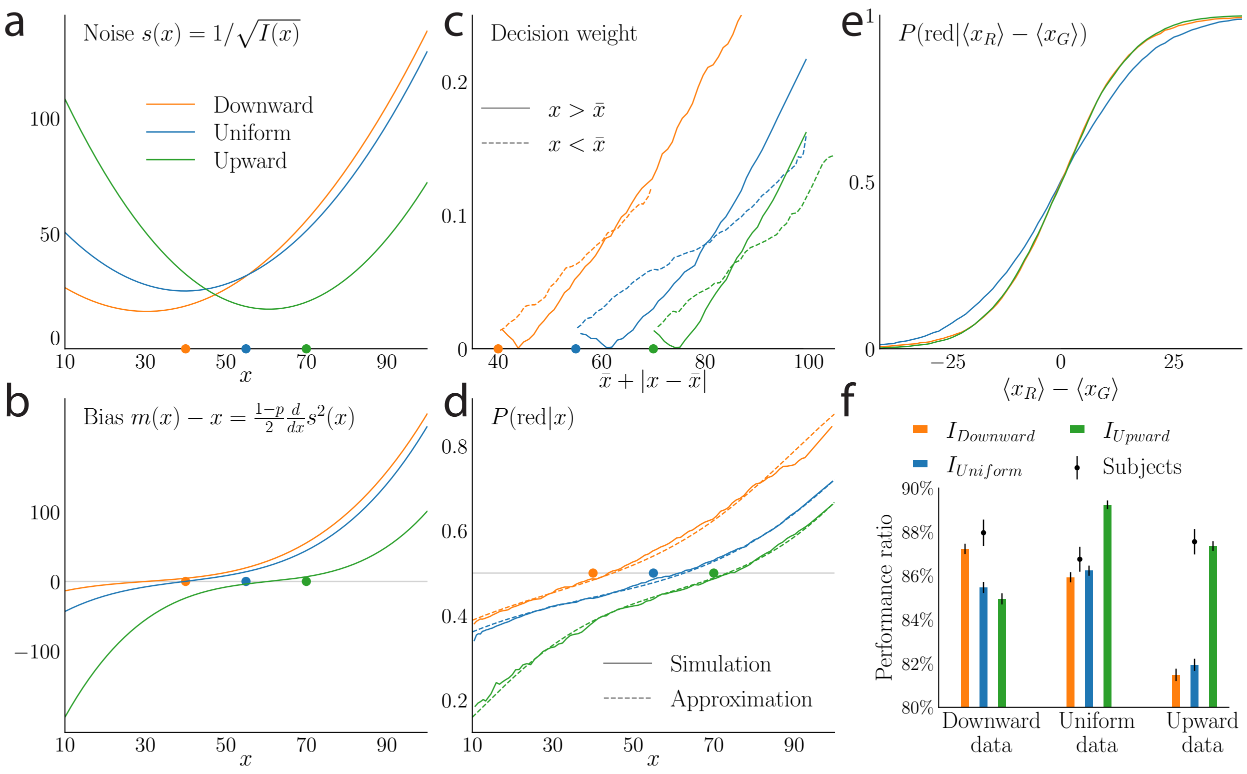

Figure 5: The efficient-coding, Bayesian decoding model reproduces subjects' behaviour, and the Fisher information functions fitted to subjects' data improve the performance ratio, in the Upward and Downward conditions. **a.** Noise $s(x)$ implied by the Fisher information, in the three prior conditions. **b.** Bias $m(x) - x$ in the three prior conditions, also implied by the Fisher information. **c.** Decision weights $|P(\text{red}|x) - 0.5|$ predicted by the efficient-coding, Bayesian decoding model. **d.** Probability of choosing 'red' conditional on a red number x being presented, in the model (solid line), and our approximation to the model prediction (dashed line; see Methods). **e.** Choice probabilities implied by the model as a function of the difference between the averages of the numbers presented. **f.** Performance ratios, over numbers sampled from the Downward (left), Uniform (middle) and Upward (right) priors, for the subjects (black dots) and implied by the estimated encoding rules (bars) for each of the three prior conditions. With numbers sampled from the Downward prior, the Downward encoding rule, $I_{Downward}$, yields the best performance (left orange bar), whereas with numbers sampled from the Upward prior, the Upward encoding rule, I_{Upward}, results in the best performance (right green bar). The black vertical lines represent the standard deviations of the performance ratios (for the subjects, this quantity is bootstrap-estimated from the data).

of efficient coding; this leads us to ask whether the differing encoding rules that we observe represent efficient adaptations to the different priors, in the sense of maximizing average reward in our task. We investigate this hypothesis by examining the performance, over numbers sampled from a given prior (say, Downward), of a model subject equipped with the Fisher information fitted to subjects' data in the context of another prior (say, Upward).

In other words, we look at how successful the "Upward encoding" (and associated Bayesian decoding) would be on "Downward data" (we use these shorthands below). We measure performance as follows. In each trial of our task, the score is augmented by the chosen average, so that the subject is guaranteed to receive at least the minimum of the two averages. The additional value to be captured in a trial i is thus the absolute difference between the two, i.e., $\Delta_i = |\langle x^R \rangle_i - \langle x^G \rangle_i|$. We define the "Performance ratio" as the fraction of this value captured by a model subject, i.e., $\sum \delta_i \Delta_i / \sum \Delta_i$, where δ_i is 1 if the model subject makes the correct choice in trial i and 0 otherwise.

With Downward data, the Downward encoding yields the largest performance ratio (87.2%, s.d.: 0.47%), whereas the Upward encoding results in the lowest ratio (84.9%, s.d.: 0.51%, Fig. 5f, left bars). Conversely, with Upward data the Upward encoding outperforms the Downward encoding (87.4% vs. 81.5%, s.d.: 0.44% and 0.56%, Fig. 5f, right bars). In other words, the encoding rule fitted to the behaviour of subjects in the context of a given prior, Upward or Downward, results in higher rewards in the context of numbers sampled from this same prior, suggesting that the choice of encoding rule is efficient. With Uniform data, the performance ratio of the Uniform encoding is slightly better than that of the Downward encoding rule, but not quite as good as that of the Upward encoding rule (86.2% vs. 85.9% and 89.2%, sd: 0.47%, 0.47%, and 0.39%). However, the performance ratio of the subjects in the Uniform condition (86.8%, s.d.: 1.1%) is lower than that in the Upward (87.6%, s.d.: 1.2%) condition. Consistent with this, the fitted Uniform encoding rule is noisier than the Upward one; and it is primarily this fact (rather than the way that the precision of encoding varies for different numbers) that makes the fitted Uniform encoding less efficient even for Uniform data.

Discussion

We designed an average-comparison task in which we changed the prior distribution from which numbers were sampled across blocks of trials. In their choices, subjects seem to differentially weight numbers that should be equally relevant to the correct decision. Subjects' behaviour can be characterized by a model of noisy perception of the size of numbers, in which both the average estimation bias and the variability of estimates depends on the magnitude of the number. Furthermore, we introduced an encoding-decoding model, in which the variable precision of encoding is specified by a Fisher information function that is efficiently chosen with regard to a subjective prior, and the decoded value of each number is the Baysian-mean estimate based on the encoded evidence. This is equivalent to a constrained version of the model of noisy perception (Eq. (5)), in which both the bias and the variable noise are determined by a single function, the Fisher information. This implies a relation between the bias and the derivative of the noise variance (Eq. (4)), for which we find qualitative support in data. The efficient-coding, Bayesian-decoding model yields the lowest BIC among the models considered, and reproduces the patterns observed in behaviour. Furthermore, the Fisher information fitted to subjects' data varies depending on the prior distribution. With the two skewed priors (Upward and Downward priors,) the Fisher information is lower for numbers less likely to appear under the prior (Fig. 4). This resulted in a higher performance in the task (Fig. 5f), suggesting that subjects efficiently adapted their decision-making process to the prior distribution of presented numbers. Our results are supported by a Bayesian random-effect analysis, which takes into account the subjects' heterogeneity (see Methods).

Reference [6] had previously found that a model featuring a nonlinear transformation of presented numbers (i.e., a bias) reproduced the apparent unequal weighting of numbers in subjects' decisions, and interpreted the bias as a strategy to compensate for decision noise. We provide a different account, that relies on two further observations. First, the noise in the perceived value of a number seems to vary with the number, and to depend on the prior (Fig. 3a; see also the Methods, in which we exhibit behavioral patterns that are captured by the models featuring both nonlinear transformation and varying noise, but not by the model that features a nonlinear transformation only). Second, the way that the bias and the noise function vary from one condition to another is consistent with the theoretical relation between these two functions predicted by our model of efficient coding and Bayesian decoding. Under this account, estimation noise, estimation bias, and thus the unequal weighting of numbers in average comparisons and the sigmoid-shape choice probability function, all derive from the form of the Fisher information function. In turn, the Fisher information is an increasing function, at least roughly, of the prior density from which numbers are sampled, suggesting that the encoding is efficiently adapted to the prior.

We do not, however, reject the hypothesis that there may also be noise in the decision process. Our models, in fact, are not inconsistent with this hypothesis. Decision noise could be included in our models by adding a noise term to the decision variable (defined as the difference between the estimated empirical averages), as in Ref. [6]. This would add a constant to the expression inside the square-root sign in Eq. (2), but the same effect would result from adding a constant to the function $s^2(x)$ in our model. The noise functions that we fit can be understood as implicitly already including this constant; note that if we assume that $s^2(x)$ is equal to $1/I(x)$ plus a constant (rather than simply $1/I(x)$, as stated above), the relationship between the bias and the variance of estimates implied by Eq. (4) remains the same, and our conclusions about the best-fitting function $I(x)$ for each prior will remain the same. We find it plausible that some amount of noise perturbs the decision, and not only the estimation of the numbers (see Refs. [25, 26]). But as introducing decision noise would not change our conclusions, we have focused, in the work reported above, on estimation noise, and on how it adapts efficiently to the prior.

The idea that the nervous system encodes stimuli efficiently was first proposed by H. Barlow [27], and finds a modern formulation in the recent literature on neural coding [17, 18, 19, 20, 23, 28]. Previous studies have provided experimental evidence that the encoding of stimuli is efficiently adjusted to the stimuli statistics [29, 18, 13, 24, 15, 19, 21]. With the exception of the 'contextual modulation' presented in Ref. [21], a common assumption is that the encoding strategy is permanently adapted to some unchanging distribution of stimuli, encountered in nature over long timescales. Here we show, conversely, that subjects are able to adapt their encoding over a relatively short time-scale, the one-hour timeframe of the experiment. This raises the question, which we leave for future investigations, as to how the brain dynamically tunes its representational capacity to changing environmental statistics.

Our study furthermore differs from these psychophysics results in that the stimuli presented are Arabic numerals. The variables relevant for decision are thus not physical magnitudes (such as the lengths of bars), but magnitudes represented in symbolic form, which one could expect to be processed in a different way (e.g., ‘25’ and ‘52’ have the same size and luminosity, but carry different semantic meaning). Studies on numerosity perception, however, suggest that there are important analogies with sensory perception [30, 31, 32, 33, 34]. For instance, subjects exhibit scalar variability in various numerosity tasks [35, 36, 37], which has also been interpreted as resulting from an efficient representation of numbers, adapted to a ‘natural’ number distribution [38]. Our results supports the hypothesis that number processing can be understood using the same theoretical framework as accounts for sensory perception. Likewise, Ref. [39] have used this framework to account for the subjective valuation of food items. In addition, we show that the encoding-decoding of numerals can quickly adapt to positively and negatively skewed distributions of numbers.

The efficient-coding, Bayesian-decoding model we examine predicts that estimates should be biased, and that a simple mathematical relation exists between the bias and the derivative of the estimation variance (Eq. (4)). A quantity related to the estimation variance and often measured in perceptual tasks is the discrimination threshold, $DT(x)$, which quantifies the sensitivity of an observer to small changes in the stimulus. From Eq. (4) we can derive, as shown in Methods, a relation between the bias and the discrimination threshold, as

$$\mathbb{E}(\hat{x} - x) \propto \frac{d}{dx} DT^2(x). \tag{6}$$

Wei and Stocker [21] derive a relationship of this form, but under an assumption that the encoding maximizes the mutual information between the number and its internal representation. They show that it is supported by numerous empirical results obtained in perceptual tasks, and thus they call it a “law of human perception”. Morais and Pillow [23] obtain this relation when encoding solves a Fisher-information optimization problem like the one that we consider (Eq. (3)), but assuming that estimates are given by a maximum-a-posteriori (MAP) estimator, while we assume a Bayesian-mean estimator. Moreover, in the version of Eq. (6) that they derive from the MAP estimator, the constant of proportionality is necessarily negative in sign, whereas a positive sign is implied by the Bayesian-mean estimator (when $p < 1$), and this is required in order to fit our empirical results.

An assumption of the efficient-coding, Bayesian-decoding model is that subjects hold a subjective prior about the numbers. If the subjects’ prior matched the actual distribution of numbers used in each of the three prior conditions, the noise function, $s(x)$, should equal the inverse prior raised to the exponent $1/(p+1)$. Although in the Downward and Upward conditions the two functions are qualitatively comparable, discrepancies remain; and in the Uniform condition, the prior is flat while the noise function is U-shaped, and larger for large numbers (Fig. 4). A possible explanation is that subjects may not hold the correct prior. They might, for instance, start the experiment with a prior that corresponds to the frequencies of numbers that one usually encounters, and not fully update these prior beliefs during the task, in spite of our efforts to familiarize them with the correct distribution (see Methods). The frequency of numeric quantities typically observed follows approximately a power law [38, 40, 41], i.e., a distribution skewed towards small numbers. In the Uniform condition, if the subjective prior is in-between the natural, power-law prior, and the correct, Uniform prior, it should also be skewed, to some extent, towards small numbers. The efficient noise function, in this case, would be larger for large numbers, which is precisely what we observe in subjects' data (Fig. 4). Incorrect subjective priors are thus a plausible source of the observed discrepancies between the subjects' noise functions and those predicted by the correct priors.

A deviation of the noise functions away from theoretical predictions is found near the endpoints of the interval of presented numbers: the noise increases, despite the rising prior density, near the lower endpoint in the Downward condition, and near the upper endpoint in the Upward condition. The model we investigate relies on a small-noise assumption (in particular, the quantity minimized in Eq. (3) is a lower bound on the mean L_p error only in the limit of a vanishing noise [23]). When noise is large, the optimal Fisher information differs from a simple power of the prior — in particular at the boundaries [20], if the mutual information is maximized ($p \to 0$). The optimal Fisher information, for a number x, presumably does not depend only on the local value of the prior at this number, $\pi(x)$, but depends instead on a larger neighborhood around it. This should distort the small-noise solution, especially at the endpoints where the prior vanishes abruptly after reaching its highest value. Subjects' data seem consistent with this account, but it calls for further theoretical investigation away from the small-noise hypothesis.

A different account of the observed discrepancies is that the subjects do not apply exactly the efficient-coding optimization program we assume, in which a lower bound on the mean L_p error is minimized (or an approximation of the mutual information is maximized, in the limiting case $p \to 0$; Eq. (3)). In our task, subjects might instead maximize the financial reward they can expect; the optimal Fisher information is presumably somewhat different for a different objective. An alternative hypothesis is that subjects adapt their encoding strategy to the prior, but choose to encode with higher precision the numbers that are close to the prior mean, instead of optimizing the program described by Eq. (3). The three best-fitting noise functions, indeed, reach their minima near the priors' means (Figs. 3a, 5a). This could be a general strategy, used for stimulus encoding in various situations, or it could stem from the specifics of our tasks, in which subjects are asked to compare empirical averages. In any case, we note that the proportionality relation between the bias and the derivative of the noise variance (Eq. (4)) relies on the choice of a noise function that is efficient in the sense of our optimization program (Eq. (3)). A different choice of encoding, such as one that minimizes the noise near the mean, would result in a different relation (if any), and presumably not one of proportionality. However, when fitting the bias and the noise separately, we find that they are consistent with the proportionality relation (Fig. 3b); and the model that directly implies this relation is our best-fitting model. (Besides, in sensory domains, the proportionality relation is supported by many experimental results, as mentioned above.) We cannot, however, reject this hypothesis; it calls for finer investigations on the influence of task reward and prior distribution on the choice of an encoding strategy.

Methods

Experiment, subjects, and reward. 37 subjects, 18 female and 19 male, aged 24.5 on average (s.d.: 8.6), participated in the experiment. All but one subject participated in two blocks of trials, in each of which all numbers were samples from a single distribution (Uniform, Upward or Downward), so that each subject (except one) experienced two of the three conditions. Each session lasted about one hour. Subjects were explicitly told the current distribution, and at the beginning of each new block they were presented with a series of random samples, in order to familiarize them with the distribution. On screen, the numbers were presented with two decimal digits, and for a duration of 500ms. In each trial, the color of the first presented number was chosen between red and green with equal probability. The score of the subject was augmented at each trial by the chosen average, and thus increased over the course of the experiment. At the end of the experiment, the subject received a financial reward, which is a linear function of the total score, with a $10 "show-up" minimum. The expected reward, not taking into account the $10 minimum, for a hypothetical subject providing random responses (i.e., choosing "red" with probability 0.5 at all trials) was $10, and an accuracy of 80% of correct responses yielded an average of $25. The average reward over the 37 subjects was $28 (s.d.: 3.8). 33 subjects experienced two blocks of 200 trials (so that 2x200=400 decisions were collected per subject), three subjects experienced two blocks of 210 trials, and one subject experienced one block of 210 trials (totaling 14,670 trials). Three trials were excluded from the analysis because the number 100.00 erroneously appeared in these trials. A total of 4610 responses were collected in the Downward condition, 5040 in the Uniform condition, and 5017 in the Upward condition. All participants provided informed consent. The procedures of this experiment comply with the relevant ethical regulations and were approved by the Institutional Review Board of Columbia University (Protocol Number: IRB-AAAR9375). The task was coded in the Python programming language and with the help of the PsychoPy library [42].

The transformation $m(x)$ and decision weights. To shed light on the relation between the transformation $m(x)$ and the decision weights, we derive an approximation to the probability $P(red|x)$ of choosing 'red' conditional on a red number x being presented. Consider first the model with constant noise ($N(m(x), s^2)$). The probability can be computed by marginalization of the probability of choosing 'red' conditional on ten numbers as

$$\begin{aligned} P(red|x) &= \int \dots \int P(red|x, x^R_{2:5}, x^G_{1:5}) \pi(x^R_2) \dots \pi(x^G_5) dx^R_2 \dots dx^G_5 \\ &= \int \Phi\Big(\frac{m(x) + \Delta_m}{s\sqrt{10}}\Big) \pi_\Delta(\Delta_m) d\Delta_m, \end{aligned} \tag{7}$$

where

$$\Delta_m \equiv \sum_{i=2}^{5} m(x^R_i) - \sum_{i=1}^{5} m(x^G_i) \tag{8}$$

and π_Δ is the prior density for this random quantity. We approximate π_Δ by a Gaussian distribution with the same mean and variance, so that

$$\pi_\Delta \approx N(-\bar{m}, 9\text{Var}m), \tag{9}$$

where $\bar{m}$ is mean of $m(x)$ under the prior and Var$m$ is the variance. Substituting this approximation in Eq. (7) results in

$$P(\text{'red'}|x) \approx \Phi\left(\frac{m(x) - \bar{m}}{\sqrt{10s^2 + 9\text{Var}m}}\right). \tag{10}$$

We find this approximation to be fairly close to the conditional probabilities obtained through simulations of the model. In case of variable noise ($N(m(x), s^2(x))$), we replace in Eq. (10) the noise variance s^2 by its average under the prior, $\mathbb{E}s^2(x)$, and despite this coarse approximation, we also obtain a close match to simulated data. Figure 5d provides an example of the quality of these approximations, in the case of the efficient-coding, Bayesian-decoding model.

Equation (10) implies that the decision weight, $|P(\text{‘red’}|x) - 0.5|$, vanishes at approximately the number whose transformation equals the average transformation. In the absence of bias ($m(x) = x$), this number would be the prior mean; but it is slightly greater in the case of the fitted transformations. Hence the decision weights at the prior mean do not fall to zero, and consequently numbers above and below the prior mean have different weights (Fig. 2c).

Model fitting and identification. When fitting subjects' data to the general one-stage model $N(m(x), s^2(x))$, it is important to note that the transformation $m(x)$ can be identified from choice data only up to an arbitrary affine transformation: a model with transformation $\alpha m(x) + \beta$ and noise $\alpha s(x)$ makes the same predictions as the model with transformation $m(x)$ and noise $s(x)$, for any non-zero α and any β (see Eq. (2)). In calculating the bias plotted in Figure 3b, we choose the ‘scale and location’ parameters α and β to satisfy two additional desiderata. First, for any choice of α, we choose β so that the left- and right-hand sides of Eq. (4) are exactly equal at the particular value of x where the right-hand side is equal to zero. And second, given this, we choose α so as to minimize

$$\frac{\int (LHS(x) - RHS(x))^2 dx}{\langle |RHS| \rangle^2},$$

a measure of the relative difference between the two functions $LHS(x)$ and $RHS(x)$ defined by the two sides of the equation.

Behaviour captured only by models with both transformation and varying noise. Here we exhibit patterns in the behaviour of subjects that are not captured by the one-stage noisy-estimation model with general transformation $m(x)$ but constant noise s, $N(m(x), s^2)$, but that can be captured by a model with both a nonlinear transformation and variable noise, $N(m(x), s^2(x))$, or by the efficient-coding, Bayesian-decoding model. The choice probability implied by the constant-noise model can be written as $\Phi(z)$, where z is

$$\frac{\sum_{i=1}^{5} m(x_i^R) - m(x_i^G)}{s\sqrt{10}},$$

whereas the choice probability implied by a model that additionally features a variable noise $\tilde{s}(x)$ is given by $\Phi(z)$ with z now defined more generally as

$$\frac{\sum_{i=1}^{5} \tilde{m}(x_i^R) - \tilde{m}(x_i^G)}{\sqrt{\sum_{i=1}^{5} \tilde{s}^2(x_i^R) + \tilde{s}^2(x_i^G)}}.$$

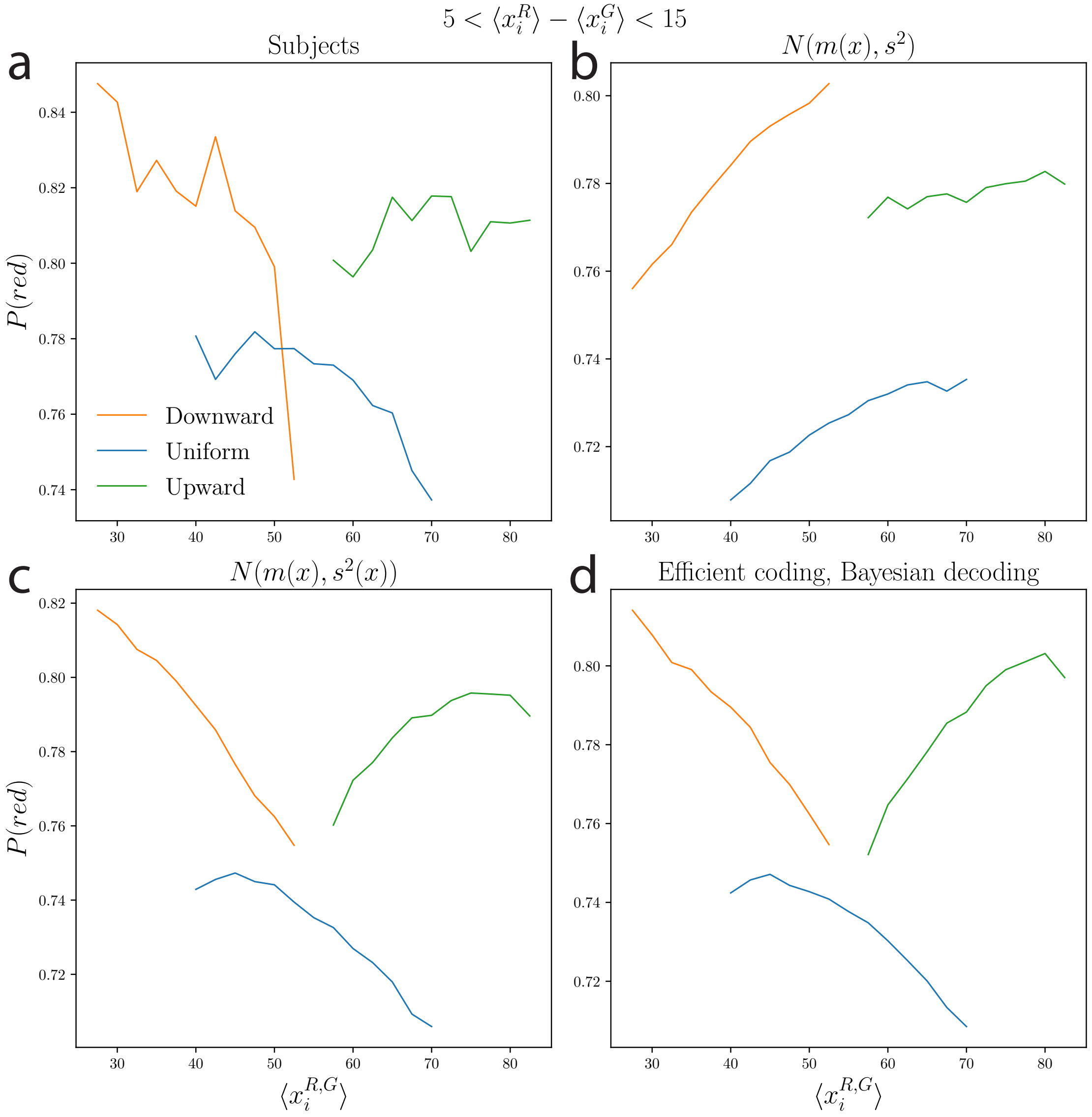

Figure 6: Choice probability as a function of overall average magnitude of numbers. Considering only the trials in which the constraint (11) is verified, the probability of choosing 'red', as a function of the average (12) of the ten numbers presented, for subjects (**a**), for the model with a transformation $m(x)$ but constant noise ($N(m(x), s^2)$, **b**), for the model with a transformation and variable noise ($N(m(x), s^2(x))$, **c**), and for the efficient-coding, Bayesian-decoding model (**d**).

Thus, over a set of presented numbers $x_{1:5}^R, x_{1:5}^G$ that all imply the same value for the numerator of z, the predicted choice probability in the constant-noise case is constant, while it varies with the values of the presented numbers in the variable-noise case. Thus an examination of the choice probabilities of the subjects over different sets of presented numbers satisfying the constraint $\sum_{i=1}^{5} m(x_i^R) - m(x_i^G) = \kappa$, for some constant κ, should allow us to discriminate between the predictions of these different model classes.

However, we observe subjects' behaviour for only a finite number of points $x_{1:5}^R, x_{1:5}^G$ in the space of $[10.00, 99.99]^{10}$ possibilities, and thus for any value of κ we have at best a few points verifying this constraint. We therefore choose a coarser version of the constraint:

$$5 \leq \frac{1}{5}\sum_{i=1}^{5} x_i^R - \frac{1}{5}\sum_{i=1}^{5} x_i^G \leq 15, \tag{11}$$

i.e., we consider the trials in which the difference in the averages of the presented numbers is between 5 and 15. We look at how the choice probabilities in these trials varies as a function of the average value of the ten numbers,

$$\langle x_i^{R,G} \rangle = \frac{1}{10}\sum_{i=1}^{5} \sum_{C \in \{R,G\}} x_i^C. \tag{12}$$

In short, we examine, for trials in which the difference in the averages is around 10, how the choice probability depends on the overall average magnitude of the presented numbers. We find that in the Uniform and Downward conditions, subjects' choice probabilities decrease as a function of the overall average, whereas they increase in the Upward condition (Fig. 6a). In other words, in the Uniform and Downward conditions the choice probability is closer to 0.5 (incorrect decisions are more frequent) when the presented numbers are large, than when they are small, while the opposite occurs in the Upward condition.

A model with a nonlinear transformation $m(x)$ but constant noise s does not capture this behaviour: in this kind of model, the choice probability is predicted to increase as a function of the overall average in all three conditions (Fig. 6b). In contrast, the model with both a transformation and variable noise ($N(m(x), s^2(x))$), and the efficient-coding, Bayesian-decoding model (which implies a particular kind of variable noise), exhibit the same patterns in behaviour as those of the subjects: the choice probability increases in the Upward condition, and decreases in the Uniform and Downward conditions (Fig. 6c,d). In these latter two conditions, the noise $s(x)$ at large numbers is appreciably larger than at small numbers (see Fig. 5a, main text), which results in choice probabilities closer to 0.5 at large numbers. These results support the hypothesis that the models which feature both a transformation and variable noise capture more accurately the behaviour of subjects, as also indicated by the model comparison statistics reported in the main text.

Bayesian model selection using individual data We conduct a model-fitting and model-selection procedure which differs from that used in the main text in two ways. First, cross-validation, instead of the Bayesian Information Criterion, is used to evaluate the likelihoods of the models while penalizing model complexity. Second, we implement a Bayesian random-effects analysis, which takes subjects heterogeneity into account.

For each subject and each of our 10 models (five estimate models, with either homogeneous or prior-specific parameters), we estimate the likelihood of the model, for that subject, through 10-fold cross-validation. Specifically, we split the subject's data in 10 interleaved subsets. For each subset, we fit the model by maximizing the likelihood of the 9 complementary subsets, and then use the best-fitting parameters to compute the 'out-of-sample' likelihood of the subset. With this procedure, for 83.8% of subjects, the highest total likelihood is obtained with the efficient-coding, Bayesian-decoding model, in which the bias is

Model	Parameters	Dirichlet $\vec{\alpha}$	Expected prob. $\langle\vec{r}\rangle$	Exceedance prob.
$N(x, s^2)$	Homogeneous	1.00	2.13%	0%
	Prior-specific	1.00	2.14%	0%
$N(x, s^2(x))$	Homogeneous	1.00	2.13%	0%
	Prior-specific	1.00	2.13%	0%
$N(m(x), s^2)$	Homogeneous	1.00	2.13%	0%
	Prior-specific	2.60	5.51%	0%
$N(m(x), s^2(x))$	Homogeneous	1.01	2.16%	0%
	Prior-specific	1.06	2.26%	0%
Efficient coding,	Homogeneous	4.97	10.58%	.00008%
Bayesian decoding	Prior-specific	32.35	68.82%	99.99992%

Table 1: Bayesian Model Selection favors the efficient-coding, Bayesian-decoding model with prior-specific parameters. Parameter vector $\vec{\alpha}$ of the Dirichlet posterior distribution over the multinomial distribution of the ten models, computed following Ref. [22], along with the implied expected probabilities of the multinomial distribution and the exceedance probabilities.

proportional to the derivative of the variance. 13.5% of subjects are best fitted by the model with transformation of the number, and constant noise ($N(m(x), s^2)$). Only 2.7% of subjects are best fitted by the model with both transformation and varying noise ($N(m(x), s^2(x))$), in spite of the absence of constraints relating these two functions, in this model. Furthermore, for 75.7% of subjects the highest total likelihood is obtained with prior-specific parameters, i.e., when the parameters of the model are allowed to depend on the prior condition (except the parameter p, in the efficient-coding, Bayesian-decoding model).

The results just presented relate to which model yields the largest likelihood, for each subject, but neglect the relative values of the likelihoods of the various models, and thus provide no indication on the weight of the evidence favoring one model over another. We conduct a finer analysis through the 'Bayesian model selection' (BMS) procedure described in Ref. [22]. This method assumes that there is a multinomial distribution over the 10 models, parameterized by the vector of probabilities of the models, $\vec{r}$, and that the behaviour of each subject is randomly chosen among the 10 models, according to this distribution. The vector $\vec{r}$ is unknown, but the BMS method prescribes a way to obtain a posterior over the vector $\vec{r}$, as a Dirichlet distribution, parameterized by a vector $\vec{\alpha}$, of size equal to that of $\vec{r}$.

We implement this procedure and obtain the parameters $\vec{\alpha}$ of the Dirichlet posterior distribution. Following the literature, we report, in Table 1, the parameters $\vec{\alpha}$, the expected probabilities $\langle\vec{r}\rangle = \vec{\alpha}/||\vec{\alpha}||$ of the models, and the 'exceedance probabilities', i.e., the probability that each model is more likely than the other models. The expected probability of the efficient-coding, Bayesian-decoding model is substantially larger than the other models: 68.8%, with prior-specific parameters, and 10.6% with homogeneous parameters. One advantage of the BMS method is that it allows to compute the expected probability of hypotheses that correspond to classes of models; for instance, in our case, the expected probability that the parameters are prior-specific, by simply adding the expected probabilities of the

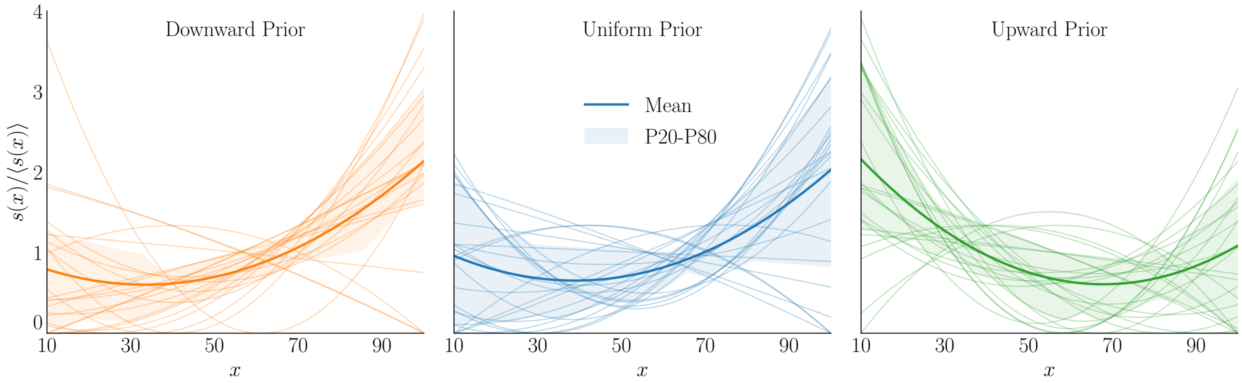

Figure 7: Subjects' noise functions are adapted to the prior distributions. For each subject and in each prior condition (Downward, *left*, Uniform, *center*, and Upward, *right*), best-fitting noise function of the subject, $s(x)$, divided by its average over the interval of numbers, $\langle s(x) \rangle$, so as to obtain comparable curves across subjects (thin lines); mean function across subjects (thick lines); and 20th and 80th percentiles (shaded areas).

corresponding models. We find that the expected probability that the parameters depend on the prior is 80.9%, thus substantially larger than the probability that they are homogeneous across prior conditions. Finally, we compute numerically the exceedance probabilities, through Monte-Carlo estimation, using 10 million samples from the Dirichlet distribution. Out of these samples, in eight cases the most likely model was the efficient-coding, Bayesian-decoding model with homogeneous parameters, and for all the other samples the most likely model was the efficient-coding, Bayesian-decoding model with prior-specific parameters.

Figure 7 shows the best-fitting noise functions for each subject, in the three prior conditions. Consistently with the functions fitted on pooled data, presented in the main text, for a majority of subjects the noise function is broadly increasing in the Downward-prior condition, and decreasing in the Upward-prior condition. More quantitatively, for instance, in the Upward-prior condition, the average value of the noise function $s(x)$ over the first half-interval (from 10 to 54.995) is greater than its average value over the second half-interval (from 54.995 to 99.99) for 80.0% of subjects, i.e., the encoding of a large majority of subjects is more noisy for the less frequent numbers. In the Downward-prior condition, consistently, the average value of the noise over the first half-interval is lower (higher precision) than the average noise over the second half-interval, for 78.3% of subjects; and in the Uniform condition, for 72.0% of subjects. Finally, the best-fitting values, across subjects, of the parameter p (median: 0.74, mean: 0.60, s.d.: 0.49) are consistent with the value obtained with the pooled data (0.70).

The results of our investigation on the individual behavior of subjects strengthen our main conclusions. First, the responses of a large majority of subjects are better captured by models with prior-specific parameters, rather than parameters that are homogeneous across priors, suggesting that the way subjects encode the presented numbers depends on the statistics of the numbers in each condition — and that the subjects were able to change their encoding within the duration of the experiment (one hour). Second, the subjects' best-fitting noise functions, in the Upward and Downward prior conditions, appear adapted to the priors, i.e., subjects encode the less likely numbers with more noise. Third, the efficient-coding, Bayesian-decoding model is the best-fitting model for most subjects, suggesting that the bias in subjects' estimates is proportional to the derivative of the estimate variance (Eq. (4)), and thus that this relation not only applies to sensory perception, but also to the understanding of quantities presented as symbols.

A different functional form: power function. In the main text, we have chosen polynomials to fit the functions $m(x)$ and $s(x)$ with a small number of parameters. Here, we investigate another functional form in which the bias is a sign-conserving power function, similar to the non-linear transformation used in Ref. [6]. More precisely, we study the model implied by Eq. (5), this time assuming that the noise variance, $s^2(x)$, is written as

$$s^2(x) = s_0^2 + \frac{a}{k+1}|x-b|^{k+1}. \tag{13}$$

The implied bias (Eq. (4)) is then a sign-conserving power function, namely

$$\mathbb{E}(\hat{x}-x) = \frac{1-p}{2}\, a \,\mathrm{sign}(x-b)|x-b|^k. \tag{14}$$

We fit the model with this functional form to our data, and find that its log-likelihood is very close, and slightly higher, than that of the same model fit with a polynomial function (-6750.46 vs. -6751.87). The power-function approach, however, uses four free parameters (a, b, k, and s_0^2), i.e., one more than the polynomial approach (in which $s(x)$ is a quadratic polynomial). As a result, the BIC with the power function is larger than that with the polynomial function (13612 vs. 13589), so that we conclude that the polynomial form is more parsimonious in its account of the data.

The two functional forms yield very similar results. In the Results section, in order to be consistent with the other models, and for the sake of concision, we present only the results obtained with the polynomial approach. Our results, however, do not depend on a specific choice of functional form.

Relations between bias, variance, and discrimination threshold. The discrimination threshold $DT(x)$ is defined as the difference δ in stimulus magnitude for which a subject distinguishes two stimuli x and $x+\delta$ with a given success rate (e.g., 75%.) In a model in which an estimate $\hat{x}$ of presented number x is normally distributed around a transformation $m(x)$ of the number, with varying noise $s(x)$, i.e., $\hat{x} \sim N(m(x), s^2(x))$, the probability of telling $x_1 = x$ and $x_2 = x + \delta$ apart is, assuming δ small,

$$\begin{aligned} P(\hat{x}_2 > \hat{x}_1) &= \Phi\Big(\frac{m(x_2)-m(x_1)}{\sqrt{s^2(x_2)+s^2(x_1)}}\Big) \\ &\approx \Phi\Big(\frac{m'(x)}{s(x)}\delta\Big) \\ &\approx \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\frac{m'(x)}{s(x)}\delta, \end{aligned}$$

This implies

$$DT(x) \propto \frac{s(x)}{m'(x)}, \tag{15}$$

where the proportionality factor depends on the chosen target success rate. In the efficient-coding, Bayesian-decoding model, the bias $m(x) - x$ is proportional to the derivative of the inverse of the Fisher information, therefore the bias is of order $\mathcal{O}(1/n)$, and $m'(x) \approx 1$. Equations (15) and (4) then immediately results in Wei and Stocker's law of human perception [21] (Eq. (6)):

$$\begin{aligned}\frac{d}{dx}DT^2(x) &\propto \frac{d}{dx}s^2(x)\\ &\propto m(x) - x = \mathbb{E}(\hat{x} - x).\end{aligned}$$

We note, in addition, that the converse derivation appears in the Supporting Information of Ref. [21]: from the relation involving the discrimination threshold (Eq. (6)), the authors derive a relation involving the noise variance, as in Eq. (4).

A model with variable noise is distinguishable from a model with constant noise. In addition to the results presented above, here we offer a theoretical discussion of the distinguishability of models with constant as opposed to variable noise. We show that, given a sufficient amount of data, a noisy-estimation model with a transformation $\tilde{m}(x)$ and a variable noise $\tilde{s}(x)$ cannot result in the same choice probabilities as a model with transformation $m(x)$ and *constant* noise s, unless $\tilde{s}(x)$ is constant. In this discussion, we assume that $\tilde{m}(x)$, $m(x)$ and $s(x)$ are smooth functions, and that $\tilde{m}(x)$ and $m(x)$ are monotonically increasing. Equality of the choice probabilities implied by the two models requires that

$$\frac{\sum_{i=1}^{5} \tilde{m}(x_i) - \tilde{m}(y_i)}{\sqrt{\sum_{i=1}^{5} \tilde{s}^2(x_i) + \tilde{s}^2(y_i)}} = \frac{\sum_{i=1}^{5} m(x_i) - m(y_i)}{s\sqrt{10}}, \tag{16}$$

where we denote the red numbers by x_i and the green numbers by y_i. Introducing the notation

$$\tilde{\Delta} = \sum_{i=1}^{5} \tilde{m}(x_i) - \tilde{m}(y_i) \quad \text{and} \quad \Delta = \sum_{i=1}^{5} m(x_i) - m(y_i),$$

we can rewrite Eq. (16) as

$$\tilde{\Delta}^2 = \frac{1}{10s^2}\Delta^2\left(\sum_{i=1}^{5} \tilde{s}^2(x_i) + \tilde{s}^2(y_i)\right). \tag{17}$$

Taking the derivative of Eq. (17) with respect to x_j, we obtain

$$\begin{aligned}2\tilde{m}'(x_j)\tilde{\Delta} = &\frac{1}{10s^2}2m'(x_j)\Delta\left(\sum_{i=1}^{5} \tilde{s}^2(x_i) + \tilde{s}^2(y_i)\right)\\ &+ \frac{1}{10s^2}\Delta^2\frac{d}{dx_j}\tilde{s}^2(x_j).\end{aligned}$$

Similarly, taking the derivative with respect to y_k, we obtain

$$\begin{aligned}
-2\tilde{m}'(x_j)\tilde{m}'(y_k) = &- \frac{1}{10s^2} 2m'(x_j)m'(y_k)\left(\sum_{i=1}^{5} \tilde{s}^2(x_i) + \tilde{s}^2(y_i)\right) \\
&+ \frac{1}{10s^2} 2m'(x_j)\Delta \frac{d}{dy_k}\tilde{s}^2(y_k) \\
&- \frac{1}{10s^2} 2m'(y_k)\Delta \frac{d}{dx_j}\tilde{s}^2(x_j).
\end{aligned}$$

Finally, taking the (third) derivative with respect to x_l, we obtain

$$\begin{aligned}
0 = &- \frac{1}{10s^2} 2m'(x_j)m'(y_k)\frac{d}{dx_l}\tilde{s}^2(x_l) \\
&+ \frac{1}{10s^2} 2m'(x_j)m'(x_l)\frac{d}{dy_k}\tilde{s}^2(y_k) \\
&- \frac{1}{10s^2} 2m'(y_k)m'(x_l)\frac{d}{dx_j}\tilde{s}^2(x_j).
\end{aligned}$$

This must be true in particular for $y_k = x_l$. Thus we must have

$$\forall x_j, \frac{d}{dx_j}\tilde{s}^2(x_j) = 0,$$

i.e., $\tilde{s}(x)$ must be a constant function.

Data availability

Requests for the data can be sent via email to the corresponding author.

Code availability

Requests for the code used for all analyses can be sent via email to the corresponding author.

References

[1] Amos Tversky. Elimination by aspects: a theory of choice. *Psychological Review*, 79(4), 1972.

[2] John W Payne, R Bettman, and Eric J Johnson. *The adaptive decision maker*. Cambridge University Press, 1993.

[3] Gerd Gigerenzer and Daniel G Goldstein. Reasoning the Fast and Frugal Way : Models of Bounded Rationality. *Psychological Review*, 103(4):650–669, 1996.

[4] Eric J Johnson and Roger Ratcliff. Computational and Process Models of Decision Making in Psychology and Behavioral Economics. In *Neuroeconomics*, chapter 3, pages 35–48. Elsevier Inc., 2014.

[5] Christopher Summerfield and Konstantinos Tsetsos. Do humans make good decisions ? *Trends in Cognitive Sciences*, 19(1):27–34, 2015.

[6] Bernhard Spitzer, Leonhard Waschke, and Christopher Summerfield. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nature Human Behaviour*, 1(8):1–8, 2017.

[7] Vickie Li, Santiago Herce Castañon, Joshua A Solomon, Hildward Vandormael, and Christopher Summerfield. Robust averaging protects decisions from noise in neural computations. *PLoS Computational Biology*, 13(8):1–19, 2017.

[8] Vincent De Gardelle and Christopher Summerfield. Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, 108(32), 2011.

[9] Konstantinos Tsetsos, Rani Moran, James Moreland, Nick Chater, Marius Usher, and Christopher Summerfield. Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, 113(11):3102–3107, 2016.

[10] David C Knill and Whitman Richards. *Perception as Bayesian Inference*. Cambridge University Press, Cambridge, 1996.

[11] Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, jan 2002.

[12] Alan A Stocker and Eero P Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585, apr 2006.

[13] Ahna R Girshick, Michael S Landy, and Eero P Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7):926–932, jul 2011.

[14] Frederike H. Petzschner, Stefan Glasauer, and Klaas E. Stephan. A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 19(5):285–293, 2015.

[15] Xue-Xin Wei and Alan A. Stocker. A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, 18(10):1509–1517, 2015.

[16] Bertrand S. Clarke and Andrew R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

[17] Nicolas Brunel and J P Nadal. Mutual information, Fisher information, and population coding. *Neural computation*, 10(7):1731–57, 1998.

[18] Deep Ganguli and Eero P Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. *Advances in neural information processing systems*, 2010:658–666, 2010.

[19] Deep Ganguli and Eero P. Simoncelli. Neural and perceptual signatures of efficient sensory coding. *ArXiv e-prints*, pages 1–24, feb 2016.

[20] Xue-Xin Wei and Alan A Stocker. Mutual Information, Fisher Information, and Efficient Coding. *Neural Computation*, 326:305–326, 2016.

[21] Xue-Xin Wei and Alan A Stocker. Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, 114(38):10244–10249, 2017.

[22] Klaas Enno Stephan, Will D Penny, Jean Daunizeau, Rosalyn J Moran, and Karl J Friston. Bayesian model selection for group studies. *NeuroImage*, 46(4):1004–1017, jul 2009.

[23] Michael Morais and Jonathan W Pillow. Power-law efficient neural codes provide general link between perceptual bias and discriminability. *Advances in Neural Information Processing Systems 31*, 2(1):5076–5085, 2018.

[24] Xue-xin Wei and Alan A Stocker. Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. *Advances in Neural Information Processing Systems*, 2012.

[25] Jan Drugowitsch, Valentin Wyart, Anne-dominique Devauchelle, and Etienne Koechlin. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, 92(6):1398–1411, 2016.

[26] Santiago Herce Castañón, Rani Moran, Jacqueline Ding, Tobias Egner, Dan Bang, and Christopher Summerfield. Human noise blindness drives suboptimal cognitive inference. *Nature Communications*, 10(1):1–11, 2019.

[27] H. B. Barlow. Possible Principles Underlying the Transformations of Sensory Messages. In Walter A. Rosenblith, editor, *Sensory Communication*, chapter 13, pages 217–234. The MIT Press, Cambridge, MA, sep 1961.

[28] Mark D. McDonnell and Nigel G. Stocks. Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Physical Review Letters*, 101(5):1–4, 2008.

[29] Alan A. Stocker and Eero P. Simoncelli. Sensory adaptation within a Bayesian framework for perception. *Advances in Neural Information Processing Systems*, 18:1291–1298, 2006.

[30] Robert S. Moyer and Thomas K. Landauer. Time required for Judgements of Numerical Inequality. *Nature*, 215:1519–1520, 1967.

[31] John M. Parkman. Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, 91(2):191–205, 1971.

[32] James V. Hinrichs, Dale S. Yurko, and Jing Mei Hu. Two-digit number comparison: Use of place information. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4):890–901, 1981.

[33] Philippe Pinel, Stanislas Dehaene, Denis Rivière, and Denis LeBihan. Modulation of parietal activation by semantic distance in a number comparison task. *NeuroImage*, 14(5):1013–1026, 2001.

[34] Esther F. Kutter, Jan Bostroem, Christian E. Elger, Florian Mormann, and Andreas Nieder. Single Neurons in the Human Brain Encode Numbers. *Neuron*, 100(3):753–761.e4, 2018.

[35] John Whalen, C. R. Gallistel, and Rochel Gelman. Nonverbal Counting in Humans: The Psychophysics of Number Representation. *Psychological Science*, 10(2):130–137, 1999.

[36] Véronique Izard and Stanislas Dehaene. Calibrating the mental number line. *Cognition*, 106(3):1221–1247, 2008.

[37] Stanislas Dehaene, Véronique Izard, Elizabeth Spelke, and Pierre Pica. Log or linear? Distinct intuitions of the number scale in western and Amazonian indigene cultures. *Science*, 320(5880):1217–1220, 2008.

[38] Samuel J. Cheyette and Steven T. Piantadosi. A unified account of numerosity perception. *Nature Human Behaviour*, 2020.

[39] Rafael Polanía, Michael Woodford, and Christian C. Ruff. Efficient coding of subjective value. *Nature Neuroscience*, 22(1):134–142, 2019.

[40] Stanislas Dehaene and Jacques Mehler. Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1):1–29, 1992.

[41] Steven T. Piantadosi and Jessica F. Cantlon. True Numerical Cognition in the Wild. *Psychological Science*, 28(4):462–469, 2017.

[42] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203, 2019.

Acknowledgments

We thank Brian Ho for his outstanding help as a research assistant, and the National Science Foundation for research support.

Author information

Contributions

M.W. conceptualized the study. M.W. and A.P.C. designed the experiment. A.P.C. implemented the task and collected the data. M.W. and A.P.C. analyzed the data and wrote the computational models. A.P.C. implemented the models. M.W. and A.P.C. interpreted the results and wrote the manuscript.

Ethics declarations

Competing interests

The authors declare no competing interests.