

Peer Review Information

Journal: Nature Human Behaviour

Manuscript Title: Efficient coding of numbers explains decision bias and noise

Corresponding author name(s): Arthur Prat-Carrabin

Reviewer Comments & Decisions:

Decision Letter, initial version:
--

22nd May 2020

Dear Dr. Prat-Carrabin,

Thank you once again for your manuscript, entitled "Efficient coding of numbers explains decision bias and noise", and for your patience during the peer review process.

Your Article has now been evaluated by 2 referees. We also solicited signed comments from Christopher Summerfield, the senior author of Ref [6], to which your work is directly related. You can find all comments at the end of this letter. Please note that signed comments do not function as reviews. The authors of signed comments are asked to comment solely on methodological aspects of your work, as well as on the extent to which their work is accurately represented in your manuscript. Decisions on whether to reject a manuscript or invite a revision are based on the feedback of our independent reviewers. However, if a decision to invite a revision is made, signed comments inform our requests for necessary revisions.

Although there is disagreement among our two independent reviewers about the suitability of the paper for publication, there is consensus that the topic is of high importance. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to (1) identifying key priorities that should be addressed in revision and (2) overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritised set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

First, Reviewer 1 raises quite substantial concerns about the accuracy of the model and its interpretation, the empirical fit of the experimental data to the model's predictions, and the strength of the experimental design. Your revision must carefully address each of these points. Importantly, Reviewer 1 requests that you complete a simulation that comprises the complete encoding-decoding process, in order to be able to provide sufficiently strong evidence for your conclusions. We ask that you carry out this additional work.

Second, Reviewer 2 also questions the interpretation of the results, as well as the strength of the motivation for the experiment and the fit between experimental data and the model. Your revision must address each of these concerns, including new analyses as necessary.

We also request that you incorporate the suggestions made in the signed comments for enhancing the justification for the modeling approach, the assessment of the experimental results (including considering alternative measures of model comparison) and the interpretation of the model's predictions.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. I have also attached a template manuscript file that exemplifies our policies and formatting requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. We understand that the COVID-19 pandemic is causing significant disruptions which may prevent you from carrying out the additional work required for resubmission of your manuscript within this timeframe. If you are unable to submit your revised manuscript within 6 months, please let us know. We will be happy to extend the submission date to enable you to complete your work on the revision.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.
- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,
Aisha

Aisha Bradshaw
Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: computational neuroscience, decision models, perception

Reviewer #2: computational neuroscience, decision models

Signed comments: Christopher Summerfield

REVIEWER COMMENTS:

Reviewer #1:

Remarks to the Author:

The paper presents experimental and modeling results of how humans estimate the average of a sequence of symbolic number stimuli. Using a similar psychophysical task, previous work has shown that humans are overweighing larger numbers when making a binary decision about the averages of two symbolic number streams. That study argued that overweighing occurs because it is beneficial with regard to late noise in the decision process. In the current paper, the authors alternatively argue that the underlying non-linear transformation is due to efficient coding, i.e. d to the statistics of the numbers stream. By experimentally varying these statistics (i.e. the prior distributions), the authors show - via a specific model analysis - that the encoding accuracy changes in a way that is loosely consistent with an efficient coding prediction.

While testing averaging behavior for different prior distributions is a welcome addition to the empirical body of studying number perception and representation, (a) the encoding-decoding model analysis is partially incorrect and misleading in suggesting that it can account for the data, and (b) the experimental support for the paper's main argument of efficient number coding is not particularly strong. As a result, the scientific evidence provided by the paper is limited. What follows is a more detailed discussion of these concerns.

(a) Encoding-decoding model analysis

Two types of models are tested. The first class is essentially a signal detection theory (SDT) description of the data by means of a simple Gaussian decision model (Eq. 2 in its most general form). Such analysis is useful because it allows to test how noise and bias parameters would need to co-vary in order to explain the data. Consistent with previous work, the authors find that bias and noise need to vary with the absolute value of the number but also that they are different under the different prior conditions. One might question the strong Gaussian assumptions that are made, which are a particular problem for the efficient coding hypothesis as this assumes a warped representation, but as a first approximation it is acceptable.

My main concerns are with the second type of model, the encoding-decoding model. First of all, the authors repeatedly refer to this model as combining efficient coding with optimal inference (decoding). This is not correct because their chosen maximum-likelihood estimator (MLE) is *not* an optimal decoding model (see Simoncelli 2009 for a short but insightful discussion about optimal estimation). A statistically optimal estimator is necessarily defined with regard to an optimization criterion (i.e. cost function). In contrast to a Bayesian decoder, however, the MLE is not specified with regard to any aspect of the performance; it is a functional estimator very much like a linear estimator.

Secondly, the formulation of the MLE model is misleading and would not hold up in a simulation validation attempt. What I mean is that the authors derive a relationship between bias and variance of an MLE at very low n (number of samples, i.e. very high noise) that is specified *up to a constant*. The author treat that constant as a free parameter, which it is not because the bias and the variance of an MLE are defined on an absolute level for any specific encoding model. As a result, the free parameter seemingly allows the MLE decoder to fit the data while in reality, operating on an actual encoding stage, I strongly suspect that the MLE will either require too much noise in the encoder in order to fit the bias or will be able to fit the noise but not the bias.

The authors may disagree. If so, running a simulation of their proposed encoding-decoding model in its *entirety* should be conclusive. That is a simulation that comprises the complete encoding-decoding process. That starts with assuming a noise level according to efficient coding or whatever one likes to assume, then computing the likelihoods and the MLEs for both number streams on *every trial*, and finally comparing them and making a binary decision. Repeating this 10000 times should provide enough synthetic data to be compared with the experimental data. This is a simple exercise that does not require much work.

It would be very surprising if this simulation would match the data because the bias magnitudes one will get with the MLE for the level of noise that is necessary to match the psychometric curve of the decision behavior will likely be far too small compared to the experimentally observed biases. Or in other words, there is no noise function that can jointly explain bias and variance with the MLE decoding model.

It is thus no coincidence either that, to the best of my knowledge, there are no examples in the literature that explain perceptual and cognitive biases with a MLE decoding model whereas the Bayesian estimation/decision framework has been so successful in that regard. The experimentally measured bias and noise levels strongly suggest that humans use prior information in their decoding process in an (approximately) Bayesian way.

This, ultimately, also raises the question of why the authors felt compelled to use the MLE decoder instead of an optimal Bayesian estimator, in particular given that the latter already has demonstrated that, in combination with an efficient encoder, it can explain perceptual averaging behavior (line orientations, Wei/Stocker 2015). At least, the authors should provide some insight into their choice.

(b) Limited empirical support

Figure 4 seems to provide the main support for the paper's conclusion that encoding precision of numbers follows an efficient coding scheme. However, the extracted Fisher information does not particularly well match the experimental prior as would be expected under an efficient coding regime, in particular for the uniform prior condition. While the authors mention in passing a couple of reasons (Discussion) a more convincing attempt to explain the discrepancy would have been desirable.

Furthermore, inter-subject variability often provides insight with regard to certain assumptions of the model. Here, for example, the general assumption was that subjects quickly learned the overall prior distributions. However, that might not have been the case for all subjects. Showing and analyzing individual subjects data would reveal this.

The particular choice of the psychophysical experiment also requires some motivation and rationale. As the apparent intention was to test the efficient coding hypothesis for number representations, would a more direct estimation task not have been a better choice as it would have allowed the direct observation of estimation biases and encoding distortions? Why this task as it also requires to assume a decision stage?

Overall, the presentation of the work is disconnected from most of the well established experimental and theoretical literature. Together with the main concerns listed above, this makes judging the scientific contribution of the work difficult.

References:

Simoncelli, E.P. (2009), Optimal Estimation in Sensory Systems, in: The New Cognitive Neuroscience, Gazzaniga, M., MIT Press

Reviewer #2:

Remarks to the Author:

Suppose you're given two sets of 5 numbers, and you're asked to report which has the higher sum. Suppose also that the numbers appear rapidly, so you don't have enough time to memorize them, let alone perform the sum perfectly. Instead, you have to store a noisy estimate of the numbers, and then estimate their sum. This paper focuses on the question: how much bias and noise should you add when you store the numbers?

The answer they propose is that you base the noise you add on the probability of a number appearing, which is the optimal strategy given a finite capacity. Surprisingly (at least to me), subjects do something like that (see Fig. 3). However, there are important deviations from this strategy. In particular, the most striking thing about Fig. 3a is that the subjects add the least amount of noise at the empirical mean of the numbers. Moreover, for the bias, there tends to be an inflection point at the empirical mean.

Overall, I believe this paper is an interesting addition to the literature. My main quibble is with the authors' interpretation: the two observations in the previous paragraph suggest that the subjects noise model is driven by the empirical mean, in a way that may be even more important than the distribution of the numbers.

I also liked the authors' hypothesis that the bias is driven by the derivative of the noise (Eq. 6). However, to me the fit is not that good; compare solid and dashed lines in Figs. 3b and 4. Finally, the authors state that the curves in Fig. 2C are nonlinear. Given this is noisy data from human experiments, I would categorize them as very linear. So this does not seem like great motivation for the study.

I have several suggestions:

1. The authors need to admit that their data is highly consistent with the hypothesis that the subjects are coding numbers near the empirical mean with the highest precision. It would be nice if they could come up with a potential explanation for this. But even if they can't, it's important to point this out as a possibility (unless they have a way to rule it out). It may be one of the most interesting aspect of the data.

2. The authors imply that data supports their model in Eq. 6. I think they need to be clear that there's qualitative, but not quantitative, agreement.
3. Using the nonlinearity of the curves in Fig. 2C seems like a bad way to motivate the work. And it's not needed; there are a lot of things to explain besides that. I would suggest downplaying this aspect of the data.
4. The authors don't comment on the fact that after coding the numbers, computing the sum is a noisy operation. This would have the effect of adding more or less constant noise to, e.g., the denominators on the right hand sides of Eqs. 1 and 2. I would think it should be taken into account in the model, although I suspect it will have a small effect.

Signed Comments from Christopher Summerfield

Remarks to the Author:

This is a clever, elegant and well-written paper that offers a mathematically interesting account of a behavioural phenomenon recently described by [Spitzer et al 2017; I was the senior author on that paper]. The work builds off the earlier finding (replicated here) that when averaging and comparing interleaved streams of numbers, humans tend to exhibit a curious nonlinear weighting of the information that is nevertheless reward-maximising under the assumption that decisions are irreducibly noisy. The previous paper thus emphasises that transforming input information in a particular way renders it robust to decision noise; the current paper reframes this argument, suggesting that the same effect could be due to a combination of noise and bias in numerical encoding, with a specific relationship between noise and bias that can be understood if participants are optimally decoding each numerical estimate using MLE with appropriate bias correction. Not only is a very interesting and elegant explanation for the effect described in Spitzer [2017], but the authors also take their theory a step further, testing it under manipulations of the prior density of number. They find that (broadly) their model accounts for the new data. In my view, the experiments are rigorously designed and executed, the model is interesting and fits the data well, and the paper explains complex concepts in a way that will be approachable for a generalist audience. As such, it is work which meets the high standards expected at NHB. I have a few questions and caveats:

1/ My first question was one of clarification. In the first modelling section, in which the authors simply try to fit the data, the authors chose polynomials as linking functions $m(x)$ and $s(x)$ to describe the relationship between numerical magnitude and the bias/variance of the encoding distribution. This choice follows naturally from the form of the weighting function both here and in Spitzer (2017) but I wondered how critical it was for the author's findings, in particular in the case of the MLE model in which there is a (lawful) restriction imposed on the relationship between $m(x)$ and $s(x)$. Spitzer and colleagues used a sign-conserving power function instead. I thought that perhaps other readers would be curious as to whether the reported finding depended on this choice.

2/ The authors report BIC as a metric of model comparison. Whilst I know this is standard in the literature I (personally) always feel more comfortable with a random- rather than fixed-effects analysis of model evidence, such as that provided by Bayesian model comparison [e.g. Stephan et al

2009, Neuroimage], and b) with crossvalidation as a way to penalise complexity. I do not disbelieve the authors, but I would be even more convinced if their model comparison exercise survived this additional test.

3/ A discussion point. I agree that the model is elegant and nicely motivates the observed form of the weighting function via a (potentially) different (and in my view equally compelling) interpretation as that provided in Spitzer (2017). I should point out, however, my very strong suspicion (backed up by simulations conducted alongside that paper but not published) that the Spitzer account – in which the transducer adapts to account for noise at decoding – makes exactly the same prediction for the variable priors experiment. This set me wondering what (if any) result would actually allow us to distinguish the two accounts, except by appeal to parsimony and/or some ad hoc notion of a priori plausibility. I have to say that if pushed, I think it very likely that there is noise both at encoding and at integration/decoding. The latter is motivated by the fact that we know well that the length, timing, and variability of the sequence strongly impact performance over and above the encoding of any particular item (see, for example, Drugowitsch 2016 Neuron; Herce-Canstanon 2019 Nat Comm). These variables seem incontrovertibly to impact the cost of *combining* information that is already encoded rather than the cost of encoding information per se, i.e. to push around noise that follows encoding (as would reasonably be captured by the noise term in Spitzer). In the authors' data, this would be shown (for example) by differences in weighting attributable to samples where there was a switch from one category to the other, or samples that differed from their predecessor by a greater magnitude, all effects which we have described before in previous papers. I would be astonished if these effects were not present in the current paper. That being said, Spitzer et al assume that encoding is noiseless. This is a convenience, but it is almost certainly not the case – we know that there is encoding noise even for symbolic number, because (for example) reaction times vary with numerical proximity in magnitude comparison tasks.

So here is my request: that the authors either 1) give more consideration to discussing these matters; 2) highlight manipulation or manipulations that they believe might distinguish these accounts, and/or 3) consider capitalising on their dataset to ask whether it is really the case that there is zero cost to combining potentially disparate numbers (in fact, to combining 4 digits every 500ms).

Author Rebuttal to Initial comments

Response to the Reviewers' Comments

We thank the reviewers, and Christopher Summerfield, for their attention to the manuscript, and their useful feedback. Following the comments and recommendations of the reviewers, we have conducted several new analyses, and we have substantially revised the manuscript. A major improvement to our study is the introduction of a model featuring Bayesian decoding (in lieu of the maximum-likelihood decoding used previously), and in which the encoding is derived from principles of efficient coding. In addition, we ran a new model-selection analysis that takes into account the heterogeneity of subjects; the results of this analysis support and strengthen our main conclusions.

We have modified the manuscript to reflect these new results, and to address the various points raised by the reviewers. We regard these revisions as substantial improvements to our manuscript, for which we are thankful to the reviewers. Our modifications, in the revised manuscript, appear in red. Below, we present, first, the details of the new model we examine, which makes use of Bayesian decoding, and second, we address, point by point, the comments made by the reviewers. Our answers appear in bold font. In our responses, we refer, where relevant, to the corresponding modifications in the manuscript.

In addition, we have made the following two changes to the manuscript. First, we concluded that the information in the Supporting Information that accompanied the previous version of the manuscript would be more appropriately incorporated into the Methods section of the manuscript, allowing us to have a single, standalone document. Thus in the new draft, the content of the previous Supporting Information has been included in the Methods section (p. 19-20, 25-26). (We can of course move some details back to a Supporting Information file, if the editors feel that the new Methods section is too long.) Second, we have added an Acknowledgments section to the revised manuscript, in order to thank a research assistant for his help, and the National Science Foundation for grant support (p. 30).

To all reviewers: New model

In the previous version of the manuscript, we examined a model based on the Maximum-Likelihood Estimator (MLE). One of the concerns of the reviewers was that the MLE was not an estimator derived from the optimization of a criterion related to the performance in the task, and that other contributions to the literature had shown the relevance of a Bayesian framework for perceptual models. Following the remarks of the reviewers, a major change in the revised manuscript is our use, instead of the MLE model considered previously, of an encoding-decoding model in which the decoding step now yields an estimate of the magnitude of the encoded number given by the Bayesian posterior mean. Apart from consistency with a number of other papers, this model of decoding has the advantage of meaning that, taking as given the nature of the encoding process, the

assumed rule for producing a decision on the basis of the encoded values is optimal (i.e., is the rule that maximizes the subject's expected reward in our task). We present this model below.

The model that we investigate is based on an analysis of efficient coding by Morais and Pillow (2018), that generalizes the earlier results of authors such as Wei and Stocker (2012). Our proposed model combines noisy but efficient encoding of the presented number with a Bayesian-mean decoding of the noisy internal representation. (Crucially, we depart from Morais and Pillow in using the posterior-mean estimate rather than the MAP estimate emphasized in their paper.) More precisely, suppose that a presented number, x , elicits in the brain a series of n signals, $r = (r_1, \dots, r_n)$, each drawn independently from a distribution conditioned on the presented number, $p(r_i|x)$. Suppose that this encoding is efficient, in the sense that the Fisher information $I(x)$ solves the optimum problem:

$$\min \int \tilde{\pi}(x) I(x)^{-p/2} dx \quad \text{s.t.} \quad \int \sqrt{I(x)} dx \leq K$$

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. The constraint in this optimization problem is the same as in Wei and Stocker (2012), though we consider a more general objective. The rationale for minimizing the objective assumed here is that, as shown by Morais and Pillow (2018), it provides a lower bound on the mean L_p error of any point estimator of x (when this mean error is evaluated using the subjective prior). (Both the value of p and the subjective prior $\tilde{\pi}(x)$ are estimated so as to provide the best fit to our experimental data. But we then check whether the purportedly efficient coding rule implied by the estimated model actually improves average performance in our task, given subjects' actual reward function and the true prior.)

In the limiting case $p \rightarrow 0$, minimizing this quantity is equivalent to maximizing an approximation to the mutual information between the number and its internal representation, which is the objective assumed by Wei and Stocker (2012). 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. (Again, we check whether our estimated encoding rule, optimized for a value $p > 0$, actually increases average performance in our task.) The optimal Fisher information, under this more general efficient coding problem, is proportional to the prior density raised to the power $2/(p+1)$.

As for the decoding, we assume that subjects' estimate of an individual number, $\hat{x}$, is the mean of the Bayesian posterior over the possible 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. (Note that all of our calculations are based on a large- n asymptotic approximation, also used in the work of Wei and Stocker as well as that of Morais and Pillow.) We find:

$$\text{Bias} \equiv E(\hat{x} - x) = \frac{1-p}{2} \frac{d}{dx} \frac{1}{I(x)}.$$

Thus, we approximate the predicted distribution of estimated values as

$$\hat{x}|x \sim N\left(x + \frac{1-p}{2} \frac{d}{dx} \frac{1}{I(x)}, \frac{1}{I(x)}\right).$$

The predictions of this model are thus similar to those of the MLE model used in our previous draft, in that the bias is proportional to the derivative of the variance. However, the proportionality constant is determined, here, by the free parameter p , while in the previous version of the manuscript it was derived from the properties of the MLE estimator, and was equal to 0.25. The best-fitting value of p , 0.70, implies a proportionality constant equal to 0.15, thus not greatly different from the value implied by the MLE estimator. For this reason, the figures in our new draft are not greatly different from our previous ones, though we use a different model and have re-done all of the calculations.

The new model, however, is more satisfactory in several regards. First, it yields a lower BIC (in spite of the penalty assigned to the additional parameter, p), and it is thus the best-fitting model under the criterion for model comparisons emphasized in our previous draft. Second, the Bayesian-mean decoding rule is the one that results in a decision rule that maximizes a subject's expected reward. (Reward is an increasing affine function of the sum of the numbers of the color chosen on a given trial; thus expected reward, under the Bayesian posterior, is an increasing affine function of the Bayesian posterior means of the individual numbers of the color that is chosen. Hence expected reward is maximized by choosing the color for which the sum of the Bayesian posterior-mean estimates is larger. This addresses a point raised by Reviewer #1.) Moreover, with this choice of a decoding rule for production of the estimated magnitudes, the decision rule that we use (comparing the sums of the estimated magnitudes) can also now be shown to be optimal for our task. Third, the functions $m(x)$ and $s(x)$, when fitted separately, are more closely consistent with the hypothesis that the bias is driven by the derivative of the variance, under this model, than under the MLE model (see Figure 3b; this relates to a point raised by Reviewer #2). Finally, this approach is more coherent with the rich literature supporting Bayesian models of perceptual biases, and with the theoretical framework developed by Wei and Stocker (2012) and Morais and Pillow (2018) in particular.

In the revised manuscript, many changes relate to this new model and to the ways in which it captures the behavioral data (p. 1-2, 7-19). We present the model on p.9, detail its behavior on pp. 10-11, and discuss several of its aspects on pp. 14-17.

References:

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.

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.

REVIEWER COMMENTS:

Reviewer #1:

Remarks to the Author:

The paper presents experimental and modeling results of how humans estimate the average of a sequence of symbolic number stimuli. Using a similar psychophysical task, previous work has shown that humans are overweighing larger numbers when making a binary decision about the averages of two symbolic number streams. That study argued that overweighing occurs because it is beneficial with regard to late noise in the decision process. In the current paper, the authors alternatively argue that the underlying non-linear transformation is due to efficient coding, i.e. d to the statistics of the numbers stream. By experimentally varying these statistics (i.e. the prior distributions), the authors show - via a specific model analysis - that the encoding accuracy changes in a way that is loosely consistent with an efficient coding prediction.

While testing averaging behavior for different prior distributions is a welcome addition to the empirical body of studying number perception and representation, (a) the encoding-decoding model analysis is partially incorrect and misleading in suggesting that it can account for the data, and (b) the experimental support for the paper's main argument of efficient number coding is not particularly strong. As a result, the scientific evidence provided by the paper is limited. What follows is a more detailed discussion of these concerns.

(a) Encoding-decoding model analysis

Two types of models are tested. The first class is essentially a signal detection theory (SDT) description of the data by means of a simple Gaussian decision model (Eq. 2 in its most general form). Such analysis is useful because it allows to test how noise and bias parameters would need to co-vary in order to explain the data. Consistent with previous work, the authors find that bias and noise need to vary with the absolute value of the number but also that they are different under the different prior conditions. One might question the strong Gaussian assumptions that are made, which are a particular problem for the efficient coding hypothesis as this assumes a warped representation, but as a first approximation it is acceptable.

My main concerns are with the second type of model, the encoding-decoding model. First of all, the authors repeatedly refer to this model as combining efficient coding with optimal

inference (decoding). This is not correct because their chosen maximum-likelihood estimator (MLE) is *not* an optimal decoding model (see Simoncelli 2009 for a short but insightful discussion about optimal estimation). A statistically optimal estimator is necessarily defined with regard to an optimization criterion (i.e. cost function). In contrast to a Bayesian decoder, however, the MLE is not specified with regard to any aspect of the performance; it is a functional estimator very much like a linear estimator.

As discussed above, we no longer base our analysis on an assumption of MLE decoding. We agree, moreover, that a Bayesian decoder is preferable. As explained above, our new model of decoding results in a decision rule that is optimal, in the sense of maximizing the subject's expected reward in our task (under the subjective prior that is implicit in both the encoding and decoding rules, according to our model). For this reason (and for other reasons explained in the response to all of the reviewers, above), we consider the new model to be superior to our previous one, and we are grateful to the reviewer for pressing us on this issue. We have accordingly revised the manuscript considerably, to reflect the introduction of this new model (pp. 9-11, 14-17).

Secondly, the formulation of the MLE model is misleading and would not hold up in a simulation validation attempt. What I mean is that the authors derive a relationship between bias and variance of an MLE at very low n (number of samples, i.e. very high noise) that is specified *up to a constant*. The author treat that constant as a free parameter, which it is not because the bias and the variance of an MLE are defined on an absolute level for any specific encoding model. As a result, the free parameter seemingly allows the MLE decoder to fit the data while in reality, operating on an actual encoding stage, I strongly suspect that the MLE will either require too much noise in the encoder in order to fit the bias or will be able to fit the noise but not the bias.

The authors may disagree. If so, running a simulation of their proposed encoding-decoding model in its *entirety* should be conclusive. That is a simulation that comprises the complete encoding-decoding process. That starts with assuming a noise level according to efficient coding or whatever one likes to assume, then computing the likelihoods and the MLEs for both number streams on *every trial*, and finally comparing them and making a binary decision. Repeating this 10000 times should provide enough synthetic data to be compared with the experimental data. This is a simple exercise that does not require much work.

It would be very surprising if this simulation would match the data because the bias magnitudes one will get with the MLE for the level of noise that is necessary to match the psychometric curve of the decision behavior will likely be far too small compared to the experimentally observed biases. Or in other words, there is no noise function that can jointly explain bias and variance with the MLE decoding model.

We thank the Reviewer #1 for his/her attention to the details of the MLE model. We agree that any encoding scheme results in a bias and a variance of the MLE that

are defined “on an absolute level”, and not only up to a proportionality constant. In fact, the relation between the bias and the noise in the MLE decoding model presented in the previous draft was characterized by a proportionality constant which was not a free parameter, but instead was necessarily equal to 0.25. The free parameters, in that model, were only those that determined the Fisher information, $I(x)$, and thus the variance. The bias was equal to one fourth of the derivative of the variance, as prescribed by the relation just mentioned, and there was no additional free parameter in the work reported in our previous draft.

But in any event, in the revised manuscript, we introduce a different model, which features Bayesian decoding instead of MLE decoding. As discussed in the comments directed to all reviewers, this new model also implies a proportionality relation between the bias and the derivative of the variance, although in this case the proportionality constant *does* depend on a free parameter, p . The best-fitting value of p implies a proportionality constant equal to 0.15, i.e., a value slightly smaller than that implied by the MLE model, 0.25. We now obtain an even better fit to our data with the Bayesian-decoder model. A single noise function, with this model, jointly explains the bias and the variance of the estimates. We note that as the bias is a function of the derivative of the noise, one can increase the amount of noise by a constant (the same for all x), without impacting the predicted bias.

As we had done for the MLE model in our previous draft, we have run trial-by-trial simulations of the new model, with its best-fitting parameters, using the same numbers as were actually presented to the subjects in the experiment. For each trial, we simulate the ten (random) estimates resulting from the ten presented numbers, and compute the decision based on these ten estimates, thus obtaining a synthetic dataset immediately comparable to that of our subjects. As we do with the subjects’ data, we compute the choice probabilities, the decision weights, and the probabilities of choosing ‘red’ conditional on a red number x being presented, on the basis of this simulated dataset. We show these quantities in the first row (A) of the figure below. We also used the model’s probability of choosing red conditional on the ten presented numbers, specified by Eq. (2) in the manuscript. For each trial, we thus obtain a (deterministic) probability of choosing red, instead of a (stochastic) binary decision. We use this dataset to compute the same three quantities, which we show in the second row (B) in the figure below. This second method, naturally, yields smoother curves. In the manuscript, we choose to use these smoother curves, in Fig. 5c,d,e (p.12), because they abstract from the variation resulting from a specific random sample, while capturing the important stylized facts about the model.

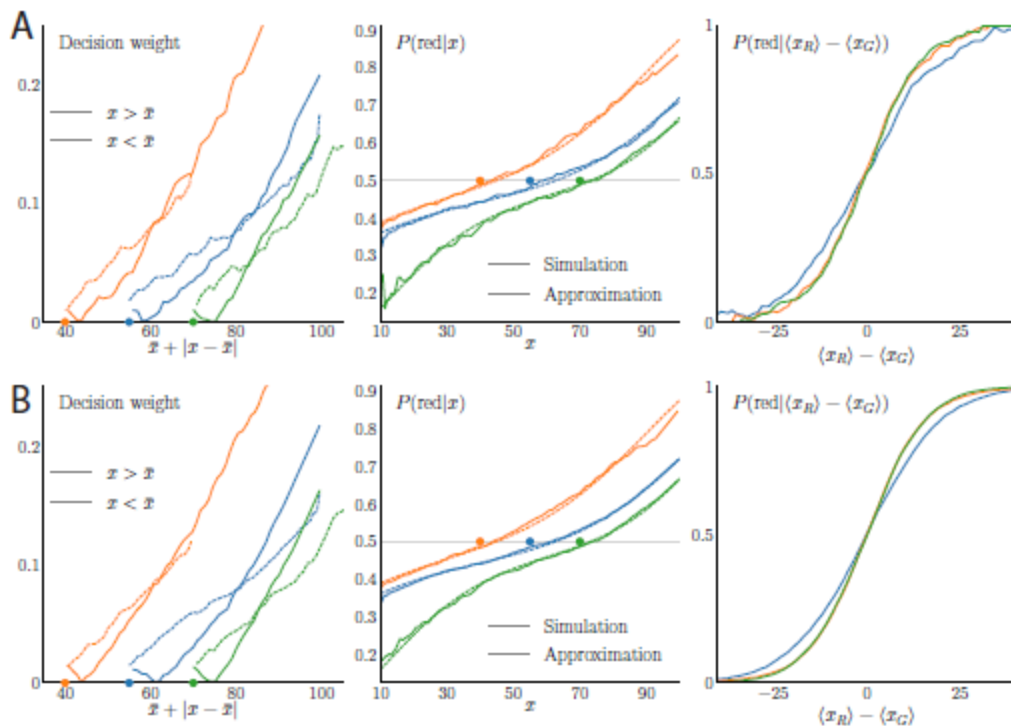


Figure: Decision Weights, Probabilities of choosing ‘red’ conditional on a red number x , and Choice Probabilities, resulting (A) from a trial-by-trial simulation of the decisions, and (B) from a trial-by-trial computation of the probability of choosing ‘red’.

It is thus no coincidence either that, to the best of my knowledge, there are no examples in the literature that explain perceptual and cognitive biases with a MLE decoding model whereas the Bayesian estimation/decision framework has been so successful in that regard. The experimentally measured bias and noise levels strongly suggest that humans use prior information in their decoding process in an (approximately) Bayesian way.

This, ultimately, also raises the question of why the authors felt compelled to use the MLE decoder instead of an optimal Bayesian estimator, in particular given that the latter already has demonstrated that, in combination with an efficient encoder, it can explain perceptual averaging behavior (line orientations, Wei/Stocker 2015). At least, the authors should provide some insight into their choice.

We agree with Reviewer #1 that the Bayesian framework, in combination with efficient coding, has been successful in accounting for behavioral and perceptual data, in addition to being well grounded from a theoretical perspective. Following the Reviewer’s comments, we have reconsidered our approach, and chose to examine a Bayesian-decoder model, instead of an MLE-decoding model. We believe that our results contribute to the literature on optimal encoding and

decoding; they suggest, in particular, that this theoretical framework also applies to the semantic representation of numerosity carried by Arabic numerals, and not only to the perception of physical stimuli (we discuss this point, in the revised manuscript, on pp.14-15).

(b) Limited empirical support

Figure 4 seems to provide the main support for the paper's conclusion that encoding precision of numbers follows an efficient coding scheme. However, the extracted Fisher information does not particularly well match the experimental prior as would be expected under an efficient coding regime, in particular for the uniform prior condition. While the authors mention in passing a couple of reasons (Discussion) a more convincing attempt to explain the discrepancy would have been desirable.

There are indeed discrepancies between the Fisher information function as fitted to our subjects' data, and the one that would be implied by efficient coding using the true prior distribution. We show, below, the revised Figure 4 (p. 11), which corresponds to the new model featuring Bayesian decoding, and in which we now show the noise function $s(x)$ (instead of the Fisher information $I(x)$) in order to be more consistent with the Figure 3a (p. 8) in which we show the best-fitting noise for the unconstrained model $(N(m(x), s^2(x)))$. Discrepancies remain, in the new model, between the theoretical predictions and the function fitted to data.

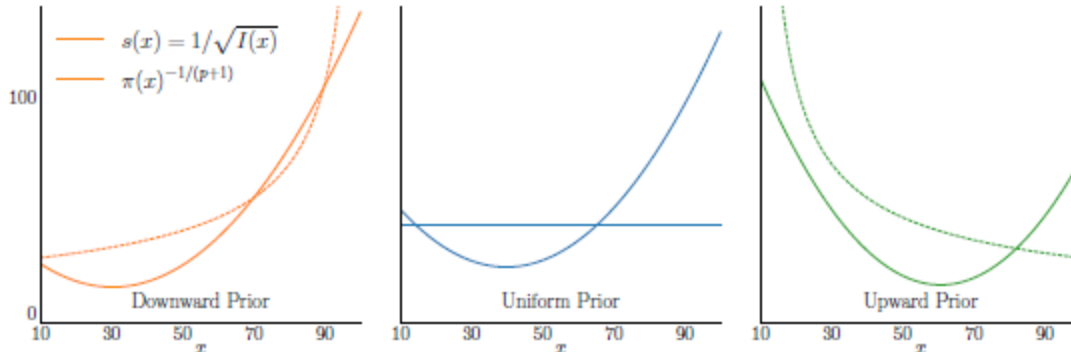


Figure: Best-fitting noise function for the Bayesian-decoder model, and predicted noise function using the correct priors.

In our new model, we assume that the Fisher information is chosen so as to solve a minimization problem that involves a subjective prior, $\tilde{\pi}(x)$, that may differ from the correct prior. A possible explanation of the observed discrepancies is that subjects may not have been able to learn the correct prior, in an experiment of the length that we conducted. 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 (detailed in Methods). We emphasize that most results on the efficient tuning of perception to a prior distribution (e.g.

Girshick et al., 2011; Wei and Stocker, 2015) have been obtained in the context of unchanging distributions of stimuli (e.g., line orientations and spatial frequencies in natural images), to which the perceptual system has presumably been able to adapt over very long timescales. By contrast, in our experiment, the subjects had to learn a new prior very quickly; it is quite plausible to suppose that they do not learn it perfectly.

The frequency of numerical quantities typically observed, outside the context of our experiment, has been argued to follow approximately a power law, i.e., a distribution skewed toward small numbers. In the Uniform condition, if the subjective prior is in-between the natural, power-law prior, and the correct, Uniform prior under the experimental condition, 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. Incorrect subjective priors (reflecting imperfect learning) 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. We suspect that this discrepancy between our theoretical predictions and what we observe may relate to the small-noise approximation that we rely upon in deriving the theoretical implications of efficient coding (following the methods also used by Wei and Stocker and by Morais and Pillow, as discussed above). In particular, the quantity minimized in the optimum problem, above, is a lower bound on the mean L_p error only in the limit of a vanishing noise (Morais and Pillow, 2018). It is already known for the special case of mutual information maximization (the $p \rightarrow 0$ limit of our model) that when noise is large, the optimal Fisher information differs from a simple power of the prior density, and this is particularly true at the boundaries (Wei and Stocker, 2016). The optimal Fisher information $I(x)$ at some number x does not depend only on the local density of the prior at this number, $\pi(x)$, but depends also on the density over a larger neighborhood around it. This makes the small-noise solution inaccurate, especially at endpoints where the prior vanishes abruptly after reaching its highest value. We suspect that a solution method that was more accurate in the case of larger noise might show that the predictions of efficient coding are not as far from our subjects' behavior as our small-noise approximation implies. However, undertaking such an analysis would mean no small increase in the numerical difficulty of the calculations, since our parameter estimation procedure requires that we be able to compute the efficient encoding rule for each of many different possible assumptions about the value of p and the degree of unavoidable encoding noise (as well as a large number of different possible priors, if we allow the subjective prior to differ from the correct one).

A different possible explanation for the observed discrepancies is that the objective for which subjects' encoding rules are optimized might not correspond to any member of the family of objectives (indexed by p) that we consider, following Morais and Pillow (2018). For example, maximization of subjects' expected financial reward in our task probably does not correspond exactly to any objective in this family, though we believe that the more realistic objective would have at least some similarities to the mean L_p reconstruction error, for a value $p > 0$. An alternative hypothesis, suggested by the Reviewer #2, 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, though this would not correspond to the solution of an optimum program of the kind described above. We discuss this possibility in our response to Reviewer #2's comment, below. In the revised manuscript, we discuss these points on pp.15-17, in the Discussion section.

Furthermore, inter-subject variability often provides insight with regard to certain assumptions of the model. Here, for example, the general assumption was that subjects quickly learned the overall prior distributions. However, that might not have been the case for all subjects. Showing and analyzing individual subjects data would reveal this.

We thank Reviewer #1 for suggesting this analysis. We conducted a Bayesian Model Selection (BMS) analysis (Stephan et al., 2009), which takes into account subjects' heterogeneity, in combination with cross-validation to evaluate the likelihoods of the different models. This method assumes, first, that there is a multinomial distribution over the 10 models we consider (with or without nonlinear transformation of the number (bias); with or without varying noise $s(x)$; or with bias constrained by the noise; and with homogeneous or prior-specific parameters). Second, it prescribes a way to obtain a Dirichlet posterior over the parameters of this multinomial distribution. In the revised manuscript, we describe this procedure in more detail, in the Methods section (pp. 19-21).

For 75.7% of subjects, the highest likelihood, computed with cross-validation, is obtained with prior-specific parameters, i.e., when the parameters of the model are allowed to depend on the prior condition. With the BMS analysis, 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.

The figure below (Figure 6, p. 21, in the revised manuscript) shows the best-fitting noise functions for each subject, in the three prior conditions. Consistently with the functions fitted on pooled data, 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.

We believe that the results of this investigation of the individual behavior of subjects strengthens 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 in addition, 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 to be adapted to the priors, in the sense that 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, and thus that this relation not only applies to sensory perception, but also to computations using quantities presented as symbols.

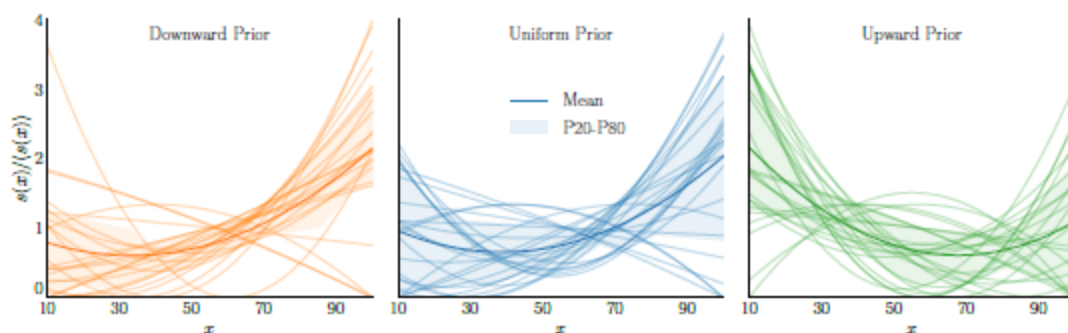


Figure: Individual subjects' noise functions, in the three prior conditions.

We include these new results, this discussion, and a more detailed presentation of the BMS analysis, in a new section in the Methods of the revised manuscript (pp. 20-24), and we refer to it in the main text on pp. 6 and 13. In addition, we report in more detail on the results of the model comparison in our response to Prof. Summerfield, below.

Reference:

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, 46(4), 1004–1017. <https://doi.org/10.1016/j.neuroimage.2009.03.025>

The particular choice of the psychophysical experiment also requires some motivation and rationale. As the apparent intention was to test the efficient coding hypothesis for number representations, would a more direct estimation task not have been a better choice as it would have allowed the direct observation of estimation biases and encoding distortions? Why this task as it also requires to assume a decision stage?

One motivation for the choice of this task was to further investigate the results of Spitzer et al., 2017 (who implemented a similar average-comparison task with numbers), in particular by manipulating the prior. (Like us, those authors argued that the encoding rule could be viewed as having been optimized for the distribution of decision problems faced in their task; but an obvious test of this hypothesis is to see if encoding changes across different prior distributions for the stimuli.) Moreover, by contrast with the physical stimuli typically used in psychophysics (e.g., bars of given lengths), estimation tasks with Arabic numerals bring additional problems: indeed, presenting a symbolic number to a subject and then asking them for an estimate of this number poses the risk that the subject simply identifies and names the symbol, instead of estimating its semantic content. We present, instead, ten numbers in rapid succession, so that it is unlikely that the subjects can remember each number exactly in their symbolic form and use rules of arithmetic to compute the two empirical sums or averages. It is more plausible that their so-called “number sense” (Dehaene, 2011) provides them with an “analog” estimate of the quantity represented by the symbol — in a way that might well be amenable to modeling through the theoretical framework developed in the context of perceptual studies.

Reference:

Dehaene, S. (2011). *The Number Sense : How the Mind Creates Mathematics*, Revised and Updated Edition.

Overall, the presentation of the work is disconnected from most of the well established experimental and theoretical literature. Together with the main concerns listed above, this makes judging the scientific contribution of the work difficult.

The new model that we use in the revised manuscript is arguably more in line with the established literature on perceptual biases resulting from efficient coding and Bayesian decoding. In terms of presentation, we have taken care to mention the previous studies that inspired our work. In particular, we cite several seminal studies that developed the theoretical framework which helped formalize the idea that the encoding by neurons should be adapted to the statistics of the environment, and that established its empirical relevance; these include, for instance, Ganguli and Simoncelli (2010, 2016), Girshick et al. (2011), and Wei and Stocker (2016, 2017), which we cite in the introduction (p. 2) and in the Discussion (p. 14). In the revised manuscript, our new model is based on a proposal by Morais and Pillow (2018), which we cite upon our presentation of the model (p. 9), and in the Discussion (p. 14, 15, and 16). Our experiment, moreover, is directly inspired

by that of Spitzer et al. (2017), to which we refer many times throughout the manuscript.

In the Discussion section of the revised manuscript, we discuss the connection of our work to the broader literature on efficient neural coding, and mention how our results provide evidence that subjects can adapt their encoding over a relatively short time-scale (p. 14). Our study also contributes to the extensive literature on number perception; in the revised manuscript, we now discuss this connection as well, on p. 15.

References:

Simoncelli, E.P. (2009), Optimal Estimation in Sensory Systems, in: The New Cognitive Neuroscience, Gazzaniga, M., MIT Press

Reviewer #2:

Remarks to the Author:

Suppose you're given two sets of 5 numbers, and you're asked to report which has the higher sum. Suppose also that the numbers appear rapidly, so you don't have enough time to memorize them, let alone perform the sum perfectly. Instead, you have to store a noisy estimate of the numbers, and then estimate their sum. This paper focuses on the question: how much bias and noise should you add when you store the numbers?

The answer they propose is that you base the noise you add on the probability of a number appearing, which is the optimal strategy given a finite capacity. Surprisingly (at least to me), subjects do something like that (see Fig. 3). However, there are important deviations from this strategy. In particular, the most striking thing about Fig. 3a is that the subjects add the least amount of noise at the empirical mean of the numbers. Moreover, for the bias, there tends to be an inflection point at the empirical mean.

Overall, I believe this paper is an interesting addition to the literature. My main quibble is with the authors' interpretation: the two observations in the previous paragraph suggest that the subjects noise model is driven by the empirical mean, in a way that may be even more important than the distribution of the numbers.

I also liked the authors' hypothesis that the bias is driven by the derivative of the noise (Eq. 6).

We thank Reviewer #2 for the positive comments on our work.

However, to me the fit is not that good; compare solid and dashed lines in Figs. 3b and 4.

Finally, the authors state that the curves in Fig. 2C are nonlinear. Given this is noisy data from human experiments, I would categorize them as very linear. So this does not seem like great motivation for the study.

We have tried to make clearer the most obvious ways in which subjects' decision weights differ from those of an ideal observer, and from those predicted by simpler models of imprecise decision making, rather than simply stating that the subjects' curves are "nonlinear". (See the new passage in red on p. 3, explaining the notable differences between subjects' curves, shown in Fig. 1, and those of an ideal observer. The discussion of Fig. 2C then builds upon this discussion.)

Our main point (which is hopefully clearer in the new draft) is not so much that the curves are very nonlinear, as that decision weights are not the same, for x a given distance away from the prior mean, regardless of whether x is above or below the prior mean. This asymmetry seems to us a clear regularity, and it is not predicted by the models in which $m(x) = x$ (top two panels of Fig. 2C), regardless of whether one allows the variance to vary with x .

I have several suggestions:

1. The authors need to admit that their data is highly consistent with the hypothesis that the subjects are coding numbers near the empirical mean with the highest precision. It would be nice if they could come up with a potential explanation for this. But even if they can't, it's important to point this out as a possibility (unless they have a way to rule it out). It may be one of the most interesting aspect of the data.

We thank Reviewer #2 for pointing out this aspect of the behavioral data. The three best-fitting noise functions do indeed reach their minima near the prior mean under each of the three priors considered (this is still true for the new, Bayesian-decoding model; see Figs. 3a, 5a in the revised manuscript). 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.

While this observation is interesting, it is not clear how it would follow from a more general hypothesis that can account for the nature of the bias and noise variance functions simultaneously. One of our reasons for interest in the kind of efficient coding theory that we pursue is that it makes tightly linked predictions about both functions $m(x)$ and $s(x)$, and the theoretically predicted proportionality relation between the bias and the derivative of the noise variance (Eq. (4)) holds fairly well in our data, even when the bias and noise variance are estimated without imposing this theoretical restriction. (See the figure below.)

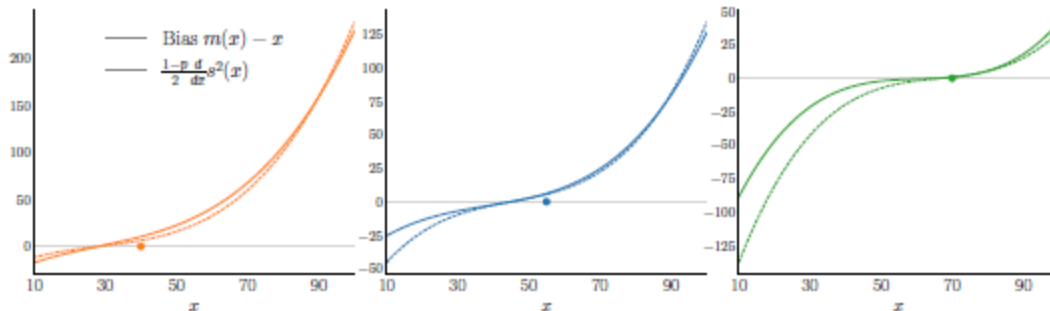


Figure: Bias and derivative of the noise variance multiplied by $(1-p)/2$, for the best-fitting $N(m(x), s(x))$ model, in which the two functions $m(x)$ and $s(x)$ are fitted separately; with the Downward (left), Uniform (middle), and Upward (right) priors.

We cannot, however, reject the hypothesis that subjects minimize the noise at a value close to the mean, as the reviewer suggests; this would call for more detailed investigation of the influence of task reward and prior distribution on the choice of encoding strategy. In the revised manuscript, we have added a discussion of this aspect of the data, and of this hypothesis (pp. 16-17).

2. The authors imply that data supports their model in Eq. 6. I think they need to be clear that there's qualitative, but not quantitative, agreement.

We agree with Reviewer #2 that the relation between bias and noise (Eq. (6) in the previous version of the manuscript, and Eq. (4) in the revised manuscript) is not verified precisely. However, as noted above, the new model featuring Bayesian decoding yields a bias function and noise function that satisfy more closely this relation (than was true with our previous MLE model), in particular in the Downward and Uniform conditions (see Figure above, left and middle panel; Fig. 3b in the manuscript). In the revised manuscript, we have taken care to indicate that the support for the relation was “qualitative” (see on p.10, and in the Discussion on p. 13).

3. Using the nonlinearity of the curves in Fig. 2C seems like a bad way to motivate the work. And it's not needed; there are lot of things to explain besides that. I would suggest downplaying this aspect of the data.

We agree with Reviewer #2 that the data present many regularities worthy of discussion besides the nonlinearity of subjects' decision weights. Our intention in presenting the figures with decision weights was, first, to demonstrate that small and large numbers seemed to be treated differently by subjects (as pointed out also by Spitzer et al., 2017). One way to show this (pioneered by Spitzer et al.) was by the asymmetry of the decision weights above and below the prior mean (Fig.

1e). Second, we show that this difference is captured only by models that feature a nonlinear transformation $m(x)$ of the number (Fig. 2c). In turn, the modulations of this transformation are accounted for, in our best-fitting model, as being driven by the modulations of the noise function. In the revised manuscript, we have reworded the corresponding paragraph, in the Results section, to take into account the comment made by the Reviewer #2 (red passage on p. 3).

4. The authors don't comment on the fact that after coding the numbers, computing the sum is a noisy operation. This would have the effect of adding more or less constant noise to, e.g., the denominators on the right hand sides of Eqs. 1 and 2. I would think it should be taken into account in the model, although I suspect it will have a small effect.

We thank Reviewer #2 for pointing out the possibility of noise arising in the decision process, and not only in the estimation of the numbers. Our new draft explicitly discusses this possibility, and makes clear that we do not reject this hypothesis. Our preferred models, in fact, are not inconsistent with it. 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 Spitzer et al., 2017. As noted by Reviewer #2, 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 $1/l(x)$ plus a constant (rather than simply $1/l(x)$, 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 $l(x)$ 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 Drugowitsch et al., 2016; Herce Castañón et al., 2019). But as introducing decision noise would not change our conclusions, we have focused, in our work, on estimation noise, and on how it adapts efficiently to the prior. We have added a new paragraph, in the Discussion section of the revised manuscript, to comment on this hypothesis (p. 14).

References:

- Drugowitsch, J., Wyart, V., Devauchelle, A., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, 92(6), 1398–1411. <https://doi.org/10.1016/j.neuron.2016.11.005>
- Herce Castañón, S., Moran, R., Ding, J., Egner, T., Bang, D., & Summerfield, C. (2019). Human noise blindness drives suboptimal cognitive inference. *Nature Communications*, 10(1), 1–11. <https://doi.org/10.1038/s41467-019-09330-7>

Signed Comments from Christopher Summerfield

Remarks to the Author:

This is a clever, elegant and well-written paper that offers a mathematically interesting account of a behavioural phenomenon recently described by [Spitzer et al 2017; I was the senior author on that paper]. The work builds off the earlier finding (replicated here) that when averaging and comparing interleaved streams of numbers, humans tend to exhibit a curious nonlinear weighting of the information that is nevertheless is reward-maximising under the assumption that decisions are irreducibly noisy. The previous paper thus emphasises that transforming input information in a particular way renders it to robust to decision noise; the current paper reframes this argument, suggesting that the same effect could be due to a combination of noise and bias in numerical encoding, with a specific relationship between noise and bias that can be understood if participants are optimally decoding each numerical estimate using MLE with appropriate bias correction. Not only is a very interesting and elegant explanation for the effect described in Spitzer [2017], but the authors also take their theory a step further, testing it under manipulations of the prior density of number. They find that (broadly) their model accounts for the new data. In my view, the experiments are rigorously designed and executed, the model is interesting and fits the data well, and the paper explains complex concepts in a way that will be approachable for a generalist audience. As such, it is work which meets the high standards expected at NHB. I have a few questions and caveats:

We thank Professor Summerfield for the positive comments on our work.

1/ My first question was one of clarification. In the first modelling section, in which the authors simply try to fit the data, the authors chose polynomials as linking functions $m(x)$ and $s(x)$ to describe the relationship between numerical magnitude and the bias/variance of the encoding distribution. This choice follows naturally from the form of the weighting function both here and in Spitzer (2017) but I wondered how critical it was for the author's findings, in particular in the case of the MLE model in which there is a (lawful) restriction imposed on the relationship between $m(x)$ and $s(x)$. Spitzer and colleagues used a sign-conserving power function instead. I thought that perhaps other readers would be curious as to whether the reported finding depended on this choice.

We have indeed chosen polynomials to fit these functions flexibly with a small number of parameters. We agree with Prof. Summerfield that one might well ask whether another functional form would yield similar results. In response, we have investigated another functional form, in which the bias is a sign-conserving power function — similar to the non-linear transformation used in Spitzer et al., 2017. More precisely, we study the model featuring the relation between the bias and the noise, this time assuming that the noise variance, $s^2(x)$, is written as

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

The implied bias is then a sign-conserving power function, namely

$$E(\hat{x} - x) = \frac{1-p}{2} a \operatorname{sign}(x - b) |x - b|^k.$$

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 manuscript, 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. In the revised manuscript, we have added a section (p. 24), in Methods, to discuss this point, and also we refer to it in the Results section, on p. 10.

2/ The authors report BIC as a metric of model comparison. Whilst I know this is standard in the literature I (personally) always feel more comfortable a) with a random- rather than fixed-effects analysis of model evidence, such as that provided by Bayesian model comparison [e.g. Stephan et al 2009, Neuroimage], and b) with crossvalidation as a way to penalise complexity. I do not disbelieve the authors, but I would be even more convinced if their model comparison exercise survived this additional test.

We thank Prof. Summerfield for suggesting this analysis, which enriches our study. Following this recommendation, we conducted a Bayesian Model Selection (BMS) analysis (Stephan et al., 2009), which takes into account subjects' heterogeneity, in combination with cross-validation to evaluate the likelihoods of the models. In our response to Reviewer #1, above, we have already presented this analysis, and showed that a large majority of subjects adapted their encoding strategy to the statistics of the numbers, in the different prior conditions. Here, we report in more detail the results of the model comparison.

For 83.8% of subjects, the highest likelihood (computed through cross-validation) is obtained with the efficient-coding, Bayesian-decoding model, in which the bias is 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.

These results 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. The BMS procedure, however, makes use of these likelihoods, and thus provides a finer look on models. We implement this procedure and obtain the parameters $\vec{\alpha}$ of the BMS Dirichlet posterior distribution. Following the literature,

we report, in the Table below (Table 1, p. 22, in the revised manuscript), the parameters $\vec{\alpha}$, the expected probabilities $\vec{r} = \vec{\alpha}/\|\vec{\alpha}\|$ 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 that of the other models: 68.8%, with prior-specific parameters, and 10.6% with homogeneous parameters. 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.

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, Bayesian decoding	Homogeneous	4.97	10.58%	.00008%
	Prior-specific	32.35	68.82%	99.99992%

Table: BMS Results. Parameter vector of the Dirichlet posterior distribution over the multinomial distribution of the ten models, implied expected probabilities, and exceedance probabilities.

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)). In the revised manuscript, we included a new section in Methods, to report these results (p. 20-24), and we refer to it in the main text (p.6 and 13).

3/ A discussion point. I agree that the model is elegant and nicely motivates the observed form of the weighting function via a (potentially) different (and in my view equally

compelling) interpretation as that provided in Spitzer (2017). I should point out, however, my very strong suspicion (backed up by simulations conducted alongside that paper but not published) that the Spitzer account – in which the transducer adapts to account for noise at decoding – makes exactly the same prediction for the variable priors experiment. This set me wondering what (if any) result would actually allow us to distinguish the two accounts, except by appeal to parsimony and/or some ad hoc notion of a priori plausibility. I have to say that if pushed, I think it very likely that there is noise both at encoding and at integration/decoding. The latter is motivated by the fact that we know well that the length, timing, and variability of the sequence strongly impact performance over and above the encoding of any particular item (see, for example, Drugowitsch 2016 Neuron; Herce-Canstanon 2019 Nat Comm). These variables seem incontrovertibly to impact the cost of *combining* information that is already encoded rather than the cost of encoding information per se, i.e. to push around noise that follows encoding (as would reasonably be captured by the noise term in Spitzer). In the authors' data, this would be shown (for example) by differences in weighting attributable to samples where there was a switch from one category to the other, or samples that differed from their predecessor by a greater magnitude, all effects which we have described before in previous papers. I would be astonished if these effects were not present in the current paper. That being said, Spitzer et al assume that encoding is noiseless. This is a convenience, but it is almost certainly not the case – we know that there is encoding noise even for symbolic number, because (for example) reaction times vary with numerical proximity in magnitude comparison tasks.

So here is my request: that the authors either 1) give more consideration to discussing these matters; 2) highlight manipulation or manipulations that they believe might distinguish these accounts, and/or 3) consider capitalising on their dataset to ask whether it is really the case that there is zero cost to combining potentially disparate numbers (in fact, to combining 4 digits every 500ms).

We thank Prof. Summerfield for raising these points. We agree that a model with a transducer, but without varying noise (i.e., with a constant degree of noise, whether its origin is in the estimation or in the decision), can capture the observed form of the decision weight functions. In fact, in Fig. 2c, bottom left panel, the model with transformation of the number and constant noise ($N(m(x), s^2)$) reproduces qualitatively the decision weight functions observed in behavioral data. But the BIC of the model with transformation and varying noise ($N(m(x), s^2(x))$), and the BIC of the model featuring a relation between the two, are both lower than that of the model with constant noise, suggesting that both a transformation and a varying noise are necessary to account for data. We also provide a qualitative, behavioral discussion of the aspects of subjects' responses that cannot be fit well without allowing for variable noise. This discussion was in the Supporting Information of the previous version of the manuscript; we now include it in the Methods section of the revised manuscript.

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

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

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 $\sum_{i=1}^5 m(x_i^R) - m(x_i^G) = k$, for some constant k , 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 k we have at best a few points verifying this constraint. We therefore choose a coarser version of the constraint:

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

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,

$$\langle x_i^{R,G} \rangle = \frac{1}{10} \sum_{i=1}^5 \sum_{c \in \{R,G\}} x_i^c$$

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 (see figure below, panel a; Fig. 6a, p. 21, in the revised manuscript). 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.

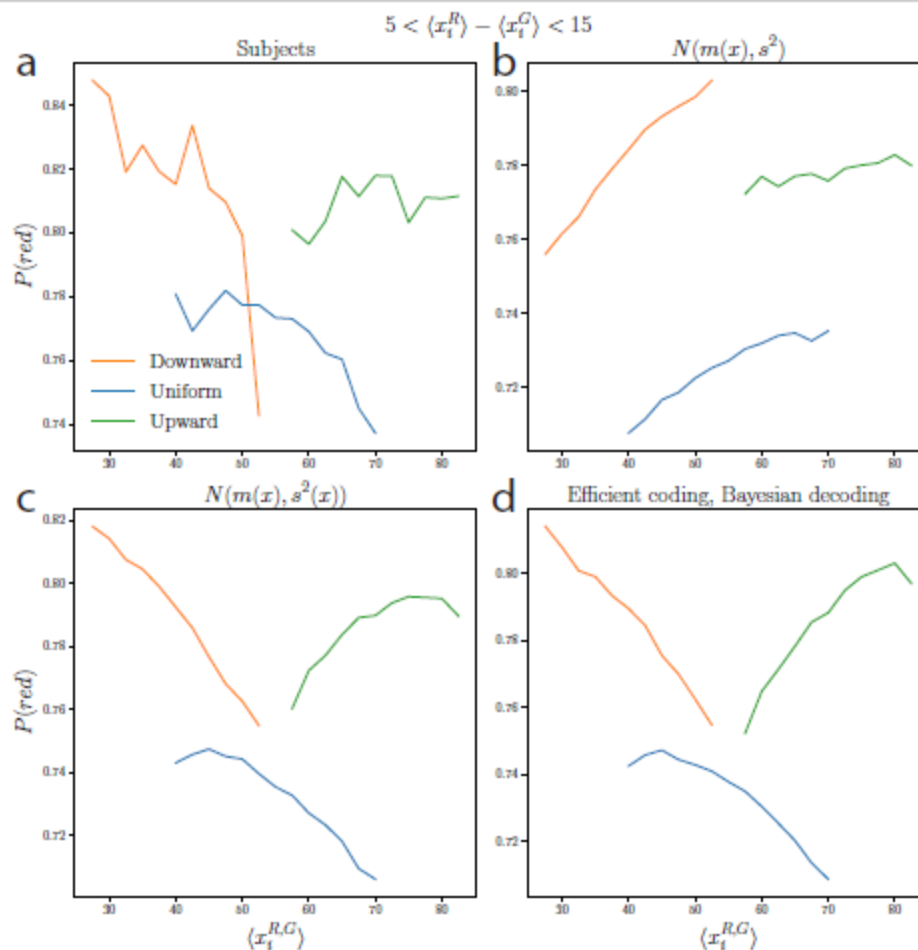


Figure: Choice probability as a function of overall average magnitude of numbers, for trials in which the difference in the averages is between 5 and 15.

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 (panel b). 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 (panels c,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.

We included these results in the Methods section of the revised manuscript (p. 19-21), and refer to it in the Results section on p. 6 and 14.

As for the noise in the decision process, we agree that it is plausible that there is noise arising at this stage as well, and not only in the estimation of the numbers. In fact, we do not reject this hypothesis, and our models are not inconsistent with such a hypothesis. We refer to our response to question N°4 of Reviewer #2 for our answer to this question. In short, introducing decision noise in our models would not change our conclusions. Thus we have focused, in our work, on estimation noise, and on how it adapts efficiently to the prior. We have added a new paragraph, in the Discussion section of the revised manuscript, to comment on this hypothesis (p. 14).

Decision Letter, first revision:

13th April 2021

Dear Arthur,

Thank you once again for your revised manuscript, entitled "Efficient coding of numbers explains decision bias and noise," and for your patience during the peer review process.

Your manuscript has now been evaluated by the same 2 reviewers who saw the original version of your work. You will find their comments included at the end of this letter. Although the reviewers continue to find your work to be of interest and indicate that it is in many ways improved over the original version, they also raise some important remaining concerns. We remain very interested in the possibility of publishing your study in Nature Human Behaviour, but would like to consider your response to these concerns in the form of a revised manuscript before we make a decision on publication.

In particular, the reviewers highlight several technical concerns that must be addressed, including Reviewer 1's comments about the parameterization of the estimate distributions and Reviewer 2's concerns about figure construction/presentation and the number of samples in Equation 3. Reviewer 2 also highlights a general concern about the clarity and organization of your manuscript. As you revise, we ask that you ensure that the aims and logic of your study are clearly presented from the beginning.

Finally, your revised manuscript must continue to comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which

will delay its consideration. To assist you in this process, I have attached a copy of our checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within four to eight weeks. We understand that the COVID-19 pandemic is causing significant disruption for many of our authors and reviewers. If you cannot send your revised manuscript within this time, please let us know - we will be happy to extend the submission date to enable you to complete your work on the revision.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.
- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,
Aisha

Aisha Bradshaw
Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: computational neuroscience, decision models, perception

Reviewer #2: computational neuroscience, decision models

REVIEWER COMMENTS:

Reviewer #1:

Remarks to the Author:

The manuscript has substantially improved. The change to a Bayesian decoder model led to an improved model account of the data, and the individual subject analysis provided statistical support for the proposed model hypothesis. Also it is indeed remarkable that the efficient-coding / Bayesian-decoding model originally introduced by Wei and Stocker seems to generalize to the perception of higher-level, symbolic variables such as numbers.

One remaining concern is that for the parametrization of the estimate distributions the authors assume the variance of the estimates to be inversely proportional to Fisher information (Eq.5). That contradicts the Cramer-Rao bound for biased estimators. Although it might be that treating the estimator like an unbiased one is not a terrible approximation in the particular case, it is a theoretical mismatch given that the paper is about estimation biases. At least, this needs to be discussed (see also the discussion in Wei/Stocker 2017).

References:

Wei, X.X. and Stocker, A.A. (2017), Lawful relation between perceptual bias and discriminability, PNAS

Reviewer #2:

Remarks to the Author:

I'm mainly happy with the paper, although I would say that it could be a lot clearer -- it took several reads to figure out what was going on. As far as I can tell, the authors do the following:

They consider two models for $P(\hat{x}|x)$, the distribution of the encoded numbers given the true numbers. The first is

$$P(\hat{x}|x) \sim N(m(x), s^2(x)).$$

The second is given in Eq. 5.

It would be very nice if they told us that from the start; starting on line 170, I was mystified until I went back and forth several times. It would have been a lot nicer if they had said: "we're going to derive $P(\hat{x}|x)$ from an optimal coding perspective, and here's how we're going to do it".

So my main comment is: be a lot more clear about the organization of the paper! Also, presumably the main point of the paper is that both m and s^2 are functions of x , and presumably the fact that the model in Eq. 5 fits better is a secondary point? Which is pretty much what the author say in the intro, but a lot of the paper is devoted to the efficient coding model, and gradually my view shifted so I thought that was the main point. Which was confusing. The authors might want to keep this in mind when writing the paper. Or not -- up to them.

And now some more minor comments.

1. I'm guessing all the figures are averaged over all the subjects. Was the data fit to individual subjects, and then averaged? Or was all data combined and then fit? This should be stated up front, in the main text (I think it's implied that they're combined in Methods, but I'm not sure). It should also be in every figure caption, as it's important. (Sorry if this was mentioned and I missed it.) And if the data was fit to individual subjects and averaged, it would be nice to see individual subject data.

2. The relationship between $I(x)$ and $\tilde{\pi}(x)$ should be derived, as should Eq. 5. Neither are hard, but Eq. 5 appears especially mysterious until you write down some equations. I personally would put them in the main text, but at the very least they should be in Methods.

3. For the model starting on line 170, it looks like n samples r_i are drawn. This increases the Fisher information by a factor of n . That factor should be included in the equations. Presumably it's an unimportant scale factor, but I believe the equations are wrong without it.

In Fig. 3a and b, the numbers on the y-axis seem way too big. For instance, in the left panel b, the bias is 200 when $x = 100$, meaning 100 comes out as 300 in the subjects' heads. I would be shocked if that were true. $s(x)$ also seems too big. It's between 50 and 100, but, from Fig. 1, I get, under a constant variance model,

$\text{Var}(\langle X_R \rangle - \langle X_G \rangle) \approx 100$ (from Fig. 1)
 $\text{Var}(\langle X_R \rangle - \langle X_G \rangle) \approx (10/25) s^2$ (constant variance model)
 $\Rightarrow s \approx \sqrt{25 \cdot 100 / 10} \approx 15$.

However, in the Fig. 3a, $\sigma(x)$ is never below 15. Is it a units problem? Or did I make some kind of mistake?

Author Rebuttal, first revision:

Response to the Reviewers' Comments

We are grateful to the Reviewers for their attention to our manuscript and for their comments. From the points raised by the reviewers, it appeared that the manuscript could, first, explain more clearly our general approach in this study, and second, provide more detail on the efficient-coding, Bayesian-decoding model (Eq. (5)). In order to address these points, we have brought, in the revised manuscript, several modifications to the Abstract, to the Introduction and to the Results section, and we have added two new sections in Methods. These changes make clearer that the encoding-decoding model is a special specification of the conditional distribution of estimates, $p(\hat{x}|x)$, and that this specification is derived from asymptotic approximations to the bias and to the variance of the Bayesian-mean decoder, which does not violate the Cramér-Rao bound. We regard these modifications as welcome clarifications that should help the reader, and we thank the Reviewers for their remarks and suggestions. We have also addressed the other points raised by the Reviewers.

Below, we answer, point by point, to the comments made by the Reviewers, and we refer, wherever relevant, to the corresponding modifications in the manuscript. Our answers appear in bold font.

Reviewer #1:

Remarks to the Author:

The manuscript has substantially improved. The change to a Bayesian decoder model led to an improved model account of the data, and the individual subject analysis provided statistical support for the proposed model hypothesis. Also it is indeed remarkable that the efficient-coding / Bayesian-decoding model originally introduced by Wei and Stocker seems to generalize to the perception of higher-level, symbolic variables such as numbers.

We thank Reviewer #1 for the positive comments on our work.

One remaining concern is that for the parametrization of the estimate distributions the authors assume the variance of the estimates to be inversely proportional to Fisher information (Eq.5). That contradicts the Cramer-Rao bound for biased estimators. Although it might be that treating the estimator like an unbiased one is not a terrible approximation in the particular case, it is a theoretical mismatch given that the paper is about estimation biases. At least, this needs to be discussed (see also the discussion in Wei/Stocker 2017).

References:

Wei, X.X. and Stocker, A.A. (2017), Lawful relation between perceptual bias and discriminability, PNAS

We thank Reviewer #1 for raising the important question of the consistency of our model of encoding-decoding with the Cramér-Rao lower bound on the variance of biased estimators. This bound is

$$\text{bound} = \frac{(1 + b'(x))^2}{I(x)},$$

where $b(x)$ is the bias in the estimation of x (and $b'(x)$ its derivative), and $I(x)$ is the Fisher information. In our model, the variance of the estimates is

$$\text{variance} = \frac{1}{I(x)}.$$

This variance is obtained as an asymptotical approximation to the variance of the Bayesian-mean decoder, when the stimulus is encoded through a vector $r = (r_1, \dots, r_n)$ of n independent signals, whose total Fisher information is $I(x)$ (Prat-Carrabin and Woodford, 2021). The Fisher information $I(x)$ is of order n , and thus the variance is of order $1/n$.

Although the two above quantities are different, and may seem to allow for a variance in the estimates that is lower than the bound, we do not suggest that the subjects' estimates of the stimuli can violate the Cramér-Rao bound. The bias in our model,

$$b(x) = \frac{1-p}{2} \frac{d}{dx} \frac{1}{I(x)},$$

also derives from an asymptotical approximation (obtained for large n), and it is of order $1/n$. With this bias, the Cramér-Rao bound, above, is asymptotically dominated by $1/I(x)$, i.e.,

$$\text{bound} \approx \frac{1}{I(x)}.$$

Therefore, the difference between the Cramér-Rao bound and the variance of the estimator in the model is within the error of our approximations; and asymptotically, it is vanishing. The distribution of estimates, in Eq. (5), is thus an approximation of an encoding-decoding estimation mechanism that is consistent with the Cramér-Rao bound. In the revised manuscript, we have reformulated the presentation of the model, so as to make clearer that its bias and its variance result from an asymptotic approximation of an encoding-decoding scheme (p. 9-10). In Methods, moreover, we further detail these points in a new section (p. 25-26), where we explicitly discuss the consistency of our results with the Cramer-Rao bound.

Reference:

Prat-Carrabin, A. and Woodford, M. (2021), Bias and variance of the Bayesian-mean decoder. *arXiv:2105.13705 [q-bio.NC]*.

Reviewer #2:

Remarks to the Author:

I'm mainly happy with the paper, although I would say that it could be a lot clearer -- it took several reads to figure out what was going on. As far as I can tell, the authors do the following:

They consider two models for $P(\hat{x}|x)$, the distribution of the encoded numbers given the true numbers. The first is

$$P(\hat{x}|x) \sim N(m(x), s^2(x)).$$

The second is given in Eq. 5.

It would be very nice if they told us that from the start; starting on line 170, I was mystified until I went back and forth several times. It would have been a lot nicer if they had said: "we're going to derive $P(\hat{x}|x)$ from an optimal coding perspective, and here's how we're going to do it".

So my main comment is: be a lot more clear about the organization of the paper! Also, presumably the main point of the paper is that both m and s^2 are functions of x , and presumably the fact that the model in Eq. 5 fits better is a secondary point? Which is pretty much what the author say in the intro, but a lot of the paper is devoted to the efficient coding model, and gradually my view shifted so I thought that was the main point. Which was confusing. The authors might want to keep this in mind when writing the paper. Or not -- up to them.

We thank Reviewer #2 for suggesting to clarify the main points of the paper. In our study, we indeed consider several models for the conditional distribution of estimates, $p(\hat{x}|x)$. The first four models are purely descriptive statistical models of noisy estimation: with no bias and with constant variance, i.e., $\hat{x}|x \sim N(x, s)$; with no bias and with the variance that changes with the number, i.e., $\hat{x}|x \sim N(x, s(x))$; with bias and a constant variance, i.e., $\hat{x}|x \sim N(m(x), s)$; and with bias and a variance that changes with the number, i.e., $\hat{x}|x \sim N(m(x), s(x))$.

The fifth model is indeed different in that it is an encoding-decoding model, in which the estimation is decomposed in these two stages. We assume in addition that the encoding is optimal (under a constraint) and that the decoding is Bayesian; as summarized by Reviewer #2, we derive $p(\hat{x}|x)$ from this perspective. The resulting distribution (Eq. (5)) is a special case of the $N(m(x), s(x))$ distribution, in which $m(x)$ and $s(x)$ are both defined by the Fisher information $I(x)$.

That both the bias and the variance are functions of the presented number, x , is indeed an important point in our study. That the model defined by Eq. (5) provides a better fit of the behavioral data, as measured by the BIC, is also a very important point. Conceptually, this finding suggests that the brain is able to adapt its encoding rule to the statistics of the numbers, and within a relatively short duration. In addition, not only is this model more parsimonious than the

unrestricted, $N(m(x), s(x))$ model, but it also shows that principles of efficient coding and Bayesian decoding are successful in describing the behavior of subjects. For these reasons, we dedicate a large fraction of the manuscript to this model.

Following the suggestions of Reviewer #2, we have improved the presentation of our results, in the revised manuscript. In the Introduction, we have clarified the question we address, and our approach, and we make explicit the progression from the purely descriptive statistical models to a model in which the estimation is decomposed into an encoding stage and a decoding stage (p. 1-2). In the Results section, when introducing this latter model, we explain clearly that, in contrast with the preceding models, we now derive the conditional distribution of estimates, $p(\hat{x}|x)$, from an encoding-decoding scheme (p. 9). Finally, we have enhanced the clarity of the Abstract, while respecting the 150-word constraint (p. 1).

And now some more minor comments.

1. I'm guessing all the figures are averaged over all the subjects. Was the data fit to individual subjects, and then averaged? Or was all data combined and then fit? This should be stated up front, in the main text (I think it's implied that they're combined in Methods, but I'm not sure). It should also be in every figure caption, as it's important. (Sorry if this was mentioned and I missed it.) And if the data was fit to individual subjects and averaged, it would be nice to see individual subject data.

It is indeed correct that in our analyses of the subjects' data, in the main text, we combine the responses from all the subjects, and use this combined dataset to compute quantities such as the choice probability and the decision weight, and to fit the models. However, we have also examined individual data, and this analysis substantiates our main results: a majority of subjects, indeed, appear to estimate with greater variability the numbers that are less frequent under the prior; and 83.8% of subjects are best fit by the efficient-coding, Bayesian-decoding model. (A finer analysis, the Bayesian Model Selection procedure, further supports these results.) These results can be found in Methods (p. 22-24). We agree that it is desirable to state clearly that we combine the subjects' data to obtain the figures in the main text and the fits of the models. In the revised manuscript, following the recommendation of Reviewer #2, we now state this at the very beginning of the presentation of the results (p. 3), and in every figure caption (p. 4, 7, 8, 11, 12, and 21).

2. The relationship between $I(x)$ and $\tilde{\pi}(x)$ should be derived, as should Eq. 5. Neither are hard, but Eq. 5 appears especially mysterious until you write down some equations. I personally would put them in the main text, but at the very least they should be in Methods.

We thank Reviewer #2 for pointing out the points that require more explanation. As for the relationship between the Fisher information $I(x)$ and the prior $\tilde{\pi}(x)$, it results from the optimum problem in Eq. (3), whose solution is obtained through variational calculus. Indeed, when taking the functional derivative of

$$\int \tilde{\pi}(x) I(x)^{-\frac{p}{2}} dx + \lambda \int \sqrt{I(x)} dx$$

with respect to $I(x)$, and setting to zero, we obtain

$$-\frac{p}{2} \tilde{\pi}(x) I(x)^{-\frac{p}{2}-1} + \frac{\lambda}{2} I(x)^{-\frac{1}{2}} = 0,$$

Which implies the relation of proportionality between the optimal Fisher information and the prior raised to the power $2/(p+1)$, i.e., $I(x) \propto \tilde{\pi}(x)^{2/(p+1)}$. In the revised manuscript, we have included in Methods a new section that provides more details on this derivation (p. 24-25). In the main text, we now mention that this solution is obtained through variational calculus, and we refer to the related new section in Methods (p. 9).

As for Eq. (5), it results from the following asymptotic approximations to the bias and to the variance of the Bayesian-mean decoder (Prat-Carrabin and Woodford, 2021):

$$\begin{aligned} \text{bias} &\approx \frac{1}{I(x)} \left[\frac{\tilde{\pi}'(x)}{\tilde{\pi}(x)} - \frac{I'(x)}{I(x)} \right] \\ \text{variance} &\approx \frac{1}{I(x)} \end{aligned}$$

The choice of a Fisher information that is a solution to the optimum problem in Eq. (3), as just presented (i.e., $I(x) \propto \tilde{\pi}(x)^{2/(p+1)}$), implies the equality

$$\frac{\tilde{\pi}'(x)}{\tilde{\pi}(x)} = \frac{p+1}{2} \frac{I'(x)}{I(x)}.$$

As a result, the bias is approximately

$$\text{bias} \approx \frac{p-1}{2} \frac{I'(x)}{I^2(x)} = \frac{1-p}{2} \frac{d}{dx} \frac{1}{I(x)}.$$

We thus posit a bias of this form, in the efficient-coding, Bayesian-decoding model (Eq. (5)). In the revised paper, we have reworded the corresponding passage in the presentation of the results (p. 9-10), in order to clarify the rationale underlying Eq. (5). In addition, a new section in Methods (p. 25-26) now presents in more detail the derivation of the bias and variance in this model.

Reference:

Prat-Carrabin, A. and Woodford, M. (2021), Bias and variance of the Bayesian-mean decoder. *arXiv:2105.13705 [q-bio.NC]*.

3. For the model starting on line 170, it looks like n samples r_i are drawn. This increases the Fisher information by a factor of n . That factor should be included in the equations. Presumably it's an unimportant scale factor, but I believe the equations are wrong without it.

We agree with Reviewer #2 that the Fisher information of the vector r scales with the number of independent samples r_i . We had defined $I(x)$ as the Fisher information of the vector r (and thus implicitly as n times the Fisher information of one sample r_i), and we had made this choice for the sake of concision, and in order to avoid carrying the symbol " n " in all the subsequent equations, which may confuse the reader. We agree, however, that this should be made clearer. In the revised manuscript, we reformulated this passage to give more clearly the definition of $I(x)$ and explicitly state that it is n times the Fisher information of one sample (p. 9).

In Fig. 3a and b, the numbers on the y-axis seem way too big. For instance, in the left panel b, the bias is 200 when $x = 100$, meaning 100 comes out as 300 in the subjects' heads. I would be shocked if that were true. $s(x)$ also seems too big. It's between 50 and 100, but, from Fig. 1, I get, under a constant variance model,

$\text{Var}(\langle X_R \rangle - \langle X_G \rangle) \approx 100$ (from Fig. 1)
 $\text{Var}(\langle X_R \rangle - \langle X_G \rangle) \approx (10/25) s^2$ (constant variance model)
 $\Rightarrow s \approx \sqrt{25 \cdot 100 / 10} \approx 15$.

However, in the Fig. 3a, $\sigma(x)$ is never below 15. Is it a units problem? Or did I make some kind of mistake?

We thank Reviewer #2 for his/her close attention to our results. The calculations of the reviewer are correct; in fact, when fitting the model with constant noise, and under an assumption that $m(x)$ is the identity mapping, we obtain best-fitting values of s around 20, thus close to the reviewer's estimate. The values of $s(x)$, in Figures 3, 4, and 5, are indeed appreciably greater than this, but these values are obtained with models in which there is a nonlinear transformation of the presented numbers, $m(x)$. As noted by the reviewer, the estimated bias $m(x) - x$ is relatively large; the resulting space of transformed numbers $m(x)$ is roughly three times larger than the original space of presented numbers. It is therefore natural that the amount of noise in these models (given the transformation of the numbers) is roughly three times larger than that in the model with no transformation. The two puzzles for Reviewer #2 (the size of the bias and the size of the noise) are thus directly related.

As for the relatively large bias, we note that the subjects are shown ten numbers, and each for half a second: we therefore find it plausible that the brain encodes each number with a relatively low precision (a much lower precision than if only one number were shown for several seconds), i.e., with a low Fisher information

$I(x)$. In our model, the bias of the optimal Bayesian decoder is inversely related to the encoding precision: so a large bias can result from the decoding of signals that have been encoded with low precision. Furthermore, less probable numbers, under the prior, are encoded with a lower precision, as a consequence of efficient coding; this results in larger biases for those numbers (Fig. 5b). In the revised manuscript, we now comment on the size of the bias, and its connection with the imprecision with which certain numbers are encoded, in the presentation of the results (p. 12).

It is also worth noting that in our estimated model with efficient coding and Bayesian decoding (the results shown in Fig. 5), we require that both the encoding and decoding be optimized for some possible prior (and the same prior in each case), but we do not require the prior to be the correct prior for that block of trials (which is known to the experimenters, but may not be fully learned by the subjects). Our subjects' behavior can only be reconciled with a model based on efficient coding and Bayesian decoding on the assumption that the subjects' implicit priors (the ones for which their encoding and decoding rules are optimized) differ from the correct prior; for example, that these implicit priors allow larger numbers to occur with positive probability than ever occur in the experiment. We nonetheless show that subjects' implicit priors vary across blocks of trials (rather than remaining the same in all three conditions), and that the way in which they vary is consistent with partial learning about the frequency distribution of numbers in each condition; for example, implicit priors in the "upward" condition appear to assign greater prior density to larger numbers than do the implicit priors in the "downward" condition, as would also be true if the correct prior were learned for each condition. In the revised manuscript, we clarify these points, in Methods (p. 24).

Decision Letter, second revision:

8th September 2021

Dear Dr. Arthur,

Thank you once again for your revised manuscript, entitled "Efficient coding of numbers explains decision bias and noise," and for your patience during the peer review process.

Your manuscript has now been evaluated by our 2 reviewers, and their comments are included at the end of this letter. You will see from their comments that the reviewers find your manuscript to be improved over the previous version, but they also highlight potentially substantial remaining issues. We remain interested in the possibility of publishing your study in Nature Human Behaviour, but you must comprehensively address these points in a final opportunity for revision before we can make a decision on publication.

In particular, Reviewer 2 raises ongoing concerns about the bias and its implications for the conclusions drawn. After receiving these reviews, we followed up with the reviewers to request additional information about this concern and detail on the nature of the problem, and you will find additional comments from Reviewer 2 included below. Your revision must thoroughly address these comments and concerns.

Additionally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a copy of our checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within four to eight weeks. We understand that the COVID-19 pandemic is causing significant disruption for many of our authors and reviewers. If you cannot send your revised manuscript within this time, please let us know - we will be happy to extend the submission date to enable you to complete your work on the revision.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.
- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,
Aisha

Aisha Bradshaw, PhD
Senior Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: computational neuroscience, decision models, perception

Reviewer #2: computational neuroscience, decision models

REVIEWER COMMENTS:

Reviewer #1:

Remarks to the Author:

I understand and appreciate the authors' explanations regarding the approximation of the variance. Nonetheless, there is still a somewhat uncomfortable feeling for using large n limit approximations (which is equivalent to an unbiased estimator) when the entire study is about accounting for the reportedly large biases (and thus low n) in human number averaging.

However, I can agree with the current version of the manuscript as it clearly indicates in its main body that the variance is approximated and provides a full explanation and justification of the approximation in its Methods section.

Reviewer #2:

Remarks to the Author:

I'm again pretty happy with the paper. Although I'm still a bit skeptical about the bias. Given that the range is 10-100, I would be shocked if the subjects ever say the average is above 100 or below 10. Especially since they can learn the downward versus upward prior in one session. I'm guessing the subjects weren't asked what the average was, but maybe this is known from other studies? The authors should definitely be clear what this means: the subjects would have to occasionally think the average is as high as 300, and also occasionally think it's negative.

Further comments from Reviewer 2:

The bias they predict is, I think, way too large: if you only saw numbers in the range 1-100, I don't think you would ever say the average is greater than 100 or less than 1, whereas their bias went as high as 300, and it could be negative (I forget by how much).

However, to figure out exactly what that means for the model would require a detailed look at exactly what they did. It's certainly possible that the bias is off but all other conclusions are correct. Alternatively, the fact that the bias is too large may mean that everything they conclude is incorrect. Or somewhere in between.

I suggest you ask the authors to provide a thorough explanation for the bias, and explain why they think their conclusions are correct when the bias is clearly way too large.

Author Rebuttal, second revision:

Response to the Reviewers' Comments

We thank the Reviewers for their attention to our manuscript. Following the comments of the reviewers, we have examined more closely the large biases in our model, and how they relate to the analytic approximation that we use. These results, presented in a new section in Methods, bring important clarifications to the reader, for which we are thankful to the Reviewers.

Below, we answer to the comments made by the Reviewers, and we refer, wherever relevant, to the corresponding modifications in the manuscript. Our answers appear in bold font.

Reviewer #1:

Remarks to the Author:

I understand and appreciate the authors' explanations regarding the approximation of the variance. Nonetheless, there is still a somewhat uncomfortable feeling for using large n limit approximations (which is equivalent to an unbiased estimator) when the entire study is about accounting for the reportedly large biases (and thus low n) in human number averaging.

However, I can agree with the current version of the manuscript as it clearly indicates in its main body that the variance is approximated and provides a full explanation and justification of the approximation in its Methods section.

We have tried to indicate more clearly in the revised manuscript that the calculations of the bias and variance of subjects' estimates are based on an asymptotic approximation (p. 13). We agree that there is reason to doubt the wisdom of using such an approximation in a study in which we conclude that both bias and variability are large (so that encoding noise cannot be too small). Hence the revision includes additional discussion of the conditions under which this approximation should be more or less reliable, including the results of some illustrative numerical simulations based on Bayesian decoding in the presence of noise of the size implied by our best-fitting parameter values (as well as reference to another paper in which we treat this issue further). See the new section p. 26-30 in Methods, and further discussion in the response to Reviewer #2 below.

Reviewer #2:

Remarks to the Author:

I'm again pretty happy with the paper. Although I'm still a bit skeptical about the bias. Given that the range is 10-100, I would be shocked if the subjects ever say the average is above 100 or below 10. Especially since they can learn the downward versus upward prior in one session. I'm guessing the subjects weren't asked what the average was, but maybe this is known from other studies? The authors should definitely be clear what this

means: the subjects would have to occasionally think the average is as high as 300, and also occasionally think it's negative.

Further comments from Reviewer 2:

The bias they predict is, I think, way too large: if you only saw numbers in the range 1-100, I don't think you would ever say the average is greater than 100 or less than 1, whereas their bias went as high as 300, and it could be negative (I forget by how much).

However, to figure out exactly what that means for the model would require a detailed look at exactly what they did. It's certainly possible that the bias is off but all other conclusions are correct. Alternatively, the fact that the bias is too large may mean that everything they conclude is incorrect. Or somewhere in between.

I suggest you ask the authors to provide a thorough explanation for the bias, and explain why they think their conclusions are correct when the bias is clearly way too large.

We are glad that Reviewer #2 is “again pretty happy” with the manuscript and we are grateful for her/his careful attention to our manuscript.

To recall, the numerical analysis of the theoretical model in our paper is based on an asymptotic (small-noise) approximation, in which the predicted magnitude of the bias, with the Bayesian decoder, is inversely related to the precision of the encoding: more specifically, the bias is approximately

$$bias \approx \frac{1}{I(x)} \left[\frac{\tilde{\pi}'(x)}{\tilde{\pi}(x)} - \frac{I'(x)}{I(x)} \right],$$

where $I(x)$ is the Fisher information, which measures the precision of the encoding. Thus the bias scales with the inverse of the Fisher information. It is reasonable to suppose that in our task, the degree of precision is relatively low (so that the predicted bias can be relatively large), because our subjects have to process 10 numbers in just five seconds. It is also worth recalling that our task does not require the subject to have too precise an estimate of the magnitude of each individual number, since they must make only a single binary decision on the basis of their sense of the average magnitudes of the two series of numbers. Nonetheless, the reviewer questions whether the size of the estimation biases implied by our model are plausible. Here we examine this question more closely.

We note first that there was an unfortunate error in the previous version of Figure 5b, that made the bias appear substantially larger than the one actually implied by our calculations. According to our asymptotic approximation (recalled above), the bias in the efficient-coding Bayesian-decoding model is proportional to the derivative of the variance of the estimates, with the coefficient of proportionality equal to 0.15 (which corresponds to $(1-p)/2$, with $p = 0.7$; see Eqs. 4-5 in the manuscript). This contrasts with the coefficient 0.25 implied by the maximum-likelihood-estimation decoding model analyzed in a previous version of

the manuscript. The relation of proportionality between the bias and the derivative of the variance was used to compute the bias shown in this figure (and only in this figure); but instead of the correct coefficient of 0.15, the figure in our previous submission mistakenly used a coefficient of 0.25 (a holdover from the calculations for the MLE decoding model). The bias actually implied by our Bayesian decoding model (under the asymptotic approximation) is thus smaller by 40% than what appeared in Figure 5b of the previous version of the manuscript. Figure (i) below shows the erroneous bias as it appeared in our previous Fig. 5b (dotted lines) together with the smaller, correct bias (solid lines). We emphasize that this error was only made in the context of Figure 5b of the last draft; all other reported calculations used the correct value of the bias. In Figure 5b of the revised version of the manuscript, we now show the correct bias (p. 12).

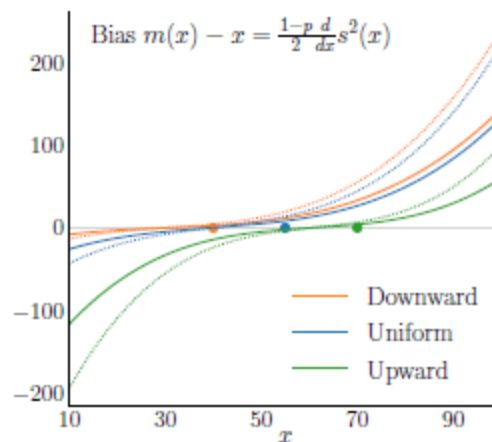


Fig. (i)

Although the magnitude of the bias is thus substantially smaller than it appeared to be in Figure 5b of the previous draft, the bias might still seem implausibly large. Our second observation is that while the corrected figure still shows large biases for some values of x (biases greater than +100 in the Downward and Uniform cases, and below -100 in the Upward case), these large biases are predicted only for extreme values of x that do not occur very often under those conditions. In fact, the biases are smaller than 50 (in absolute value) 88% of the time in the Downward case, 78% of the time in the Upward case, and 68% of the time in the Uniform case, according to our calculations based on the asymptotic approximation.

Consequently, it is relatively rare that the estimate of an average of five numbers be extremely biased. For instance, a worry expressed by Reviewer #2 is that the subjects' estimate of the average of the presented numbers could be above 300, or below 0, although all of the numbers are actually positive and below 100. For each prior, we thus look at the distribution of the estimates, by the model, of the averages of five numbers, where each number is sampled from the prior (Fig. (ii) below).

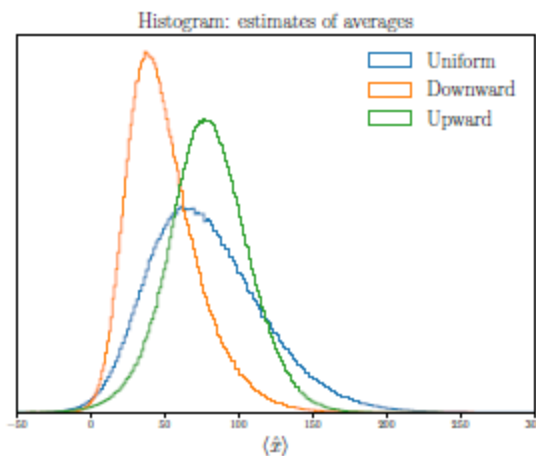


Figure (ii)

These empirical distributions were obtained by simulating 10 million model estimates of the averages of five numbers (for each prior). We find that in the case of the Downward and Upward distributions, in none of these simulations did the estimate of the average exceed 300; with the Uniform distribution, it exceeded 300 in only 114 simulations (i.e., in 0.00114% of cases). As for estimates below zero, we find that this occurs in well below 1 percent of all cases (0.60% in the Uniform condition, 0.25% in the Downward condition and 0.37% in the Upward condition). We conclude that while the model implies that a subject sometimes has an extreme estimate of the average, this is quite rare.

Third, it should also be noted that while a model of Bayesian decoding implies that estimates should always be within the support of the prior distribution, our model does not assume that subjects' estimates represent Bayesian posterior mean estimates based on the *objectively correct prior* (i.e., the distribution of numbers actually used in the experiment in each condition). Instead, we hypothesize that subjects' decisions represent optimal Bayesian decisions under an *implicit prior* distribution, that we need to infer from their behavior. Thus estimates that are occasionally outside the interval [0, 100] are perfectly consistent with Bayesian-mean decoding, if we suppose that the subject's prior has a support wider than the range of numbers actually used in the experiment.

Of course, we argue that our subjects' implicit priors seem to adjust, at least to some extent, to the distribution of numbers actually used; we conclude that the implicit priors are different under the three different conditions, with the mean of the implicit prior highest for the Upward distribution (where the correct prior mean is highest) and lowest for the Downward distribution (where the correct prior mean is lowest). Since we argue that at least some degree of learning appears to occur even over the time scale involved in our experiment, we would be disturbed if the implicit priors required to make sense of the subjects' behavior were very different (for example, with a much higher variance) than the actual distributions used in the

experiment. But there is nothing implausible about assuming that subjects' decision rules are optimal for a prior that assigns some probability to the occurrence of numbers more extreme than any of those actually used.

Finally, we concede that there is some degree of numerical error involved in the use of our asymptotic (small-noise) approximation in an experimental setting where our estimates imply (and indeed it is reasonable to expect) that the noise in the internal representations of the individual numbers is substantial. Small-noise approximations of the same kind as we use are commonplace in the literature on efficient coding (e.g., references [21,28]), and the use of this approximation has important advantages: first, of providing analytic expressions for the bias and for the variance of the estimates (Eq. 5), and second, of allowing to abstract away the specifics of the encoding (such as the detailed properties of encoding neurons). Nonetheless, it is worth considering the errors that this might involve in the present application. It seems likely that our calculations based on the asymptotic approximation exaggerate the size of the bias in the average estimates associated with extreme values of x , especially when these are in a region of low prior density (very large numbers, in the Downward case, or very small numbers, in the Upward case).

In Prat-Carrabin and Woodford (2021), we show (in the context of a Gaussian prior and a Gaussian encoding) that the error is larger — and the magnitude of the bias is overestimated — for the less frequent stimuli. In our model, the largest biases in the Upward and in the Downward conditions are obtained, respectively, for the small numbers and for the large numbers, i.e., precisely for the less frequent numbers, where the bias is likely to be overestimated.

To illustrate this qualitatively, we run simulations of an encoding-decoding process in which we compute the Bayesian posterior mean estimate taking into account the actual (rather large) degree of noise in the internal representations, rather than relying upon the asymptotic approximation. As in Prat-Carrabin and Woodford (2021), in these simulations we assume an encoding model with Gaussian noise, i.e., given a stimulus x the encoded representation r is sampled from a normal distribution with mean $\mu(x)$ and variance v^2 , which are chosen so that the Fisher information of this encoding scheme is precisely that of our best-fitting model (for each of the three conditions). The estimated value $\hat{x}(r)$ is then computed (numerically) as the mean of the Bayesian posterior.

Consistently with our model, the subjective prior $\tilde{\pi}(x)$ is assumed to be the “implicit prior” for which the Fisher information would be optimal under the asymptotic approximation (i.e., $I(x) \propto \tilde{\pi}(x)^{\frac{2}{p+1}}$). Implementing this proposal requires, however, that we specify the precision of the encoding (i.e., the Fisher information) outside the range of numbers actually used in the experiment. We cannot estimate this from our data, but we do not believe there is any reason to assume that the precision of encoding is zero for all numbers outside the true support (given that the prior need not have been learned too precisely). Thus we

choose a Fisher-information function that decreases outside this range as the inverse of the distance to the closest border of the range (i.e., numbers further from the true support are more poorly encoded). These simulations are not meant to constitute a new model (and we do not fit their parameters to behavioral data), but only to give an illustration of the degree to which the bias of a Bayesian decoder in the case of large encoding noise may differ from the bias implied by the asymptotic approximation.

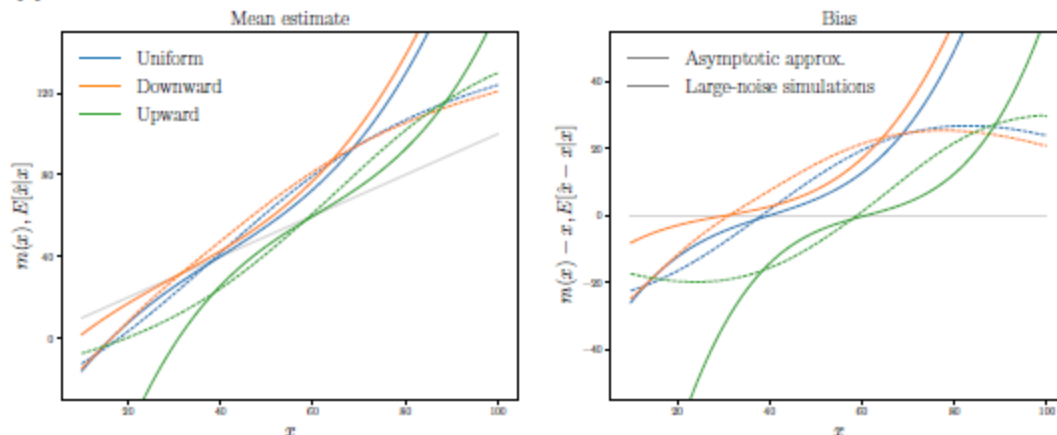


Figure (iii)

The Figure (iii) above shows the estimates (left panel), and the average biases (right panel) resulting from the simulations of the encoding-decoding process taking into account the degree of noise implied by our estimates (dashed lines) and from the asymptotic approximation used in the paper (solid lines). For all three priors, the biases in the two cases follow the same broad pattern: (i) the bias is zero when x is equal to the mean of the implicit prior; (ii) the bias is negative for values of x below the prior mean, and positive for values of x above the prior mean (i.e., the bias is “repulsive”); and (iii) the bias increases with x at about the same rate on average, over much of the range of values occurring in the experiment. The exact shapes of the bias functions are however different, and the predicted values for the bias are quite different in the case of extreme values of x that occur with low frequency (very large values under the Downward condition, very small values under the Upward condition).

This suggests that the largest of the biases shown in Figure 5b (even in its corrected form) are actually artifacts of the inaccuracy of the asymptotic approximation in the case of extreme values of x . For small values under the Upward prior, and large values under the Downward prior (i.e., for infrequent values), the asymptotic approximation implies large biases (as shown in Figure (i) above), while the bias in our simulations based on Bayesian decoding for the actual noise variance is more modest; in fact, it is never larger than 30. This suggests that

in a more exact calculation of the implications of a Bayesian decoding model fit to our subjects' data, the biases in subjects' estimates of individual numbers would also have a maximal magnitude on this order.

One might be concerned that the numerical results in our paper rely upon an approximation that sometimes overestimates the bias in this way; but these overestimations happen infrequently (according to the simulations just discussed), so that overall, they are likely to have only a limited impact on our conclusions (for example, about the Fisher-information function that best fits behavior under each of the three conditions). Specifically, the asymptotic approximation is less accurate when the encoding precision is low; but in our efficient coding model, encoding precision is low precisely for the numbers that are the least frequent (Eq. 16). As a result, the distributions of implied values for the estimation error $\hat{x} - x$ (taking into account the noise in the estimates due to the noise in r) are not very different in our simulation using large-noise Bayesian decoding and the calculations using the asymptotic approximation, as shown in Figure (iv) below. (Note that in this figure the estimation noise is included, while Fig. (iii) shows the average bias).

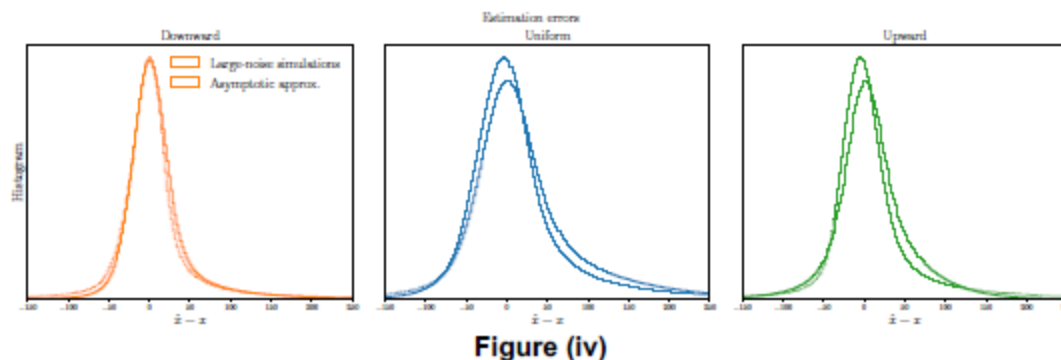


Figure (iv)

In summary, the biases in the estimates of numbers by a Bayesian decoder can be somewhat larger than the referee may have supposed, because our hypothesis allows for decoding using an “implicit prior” that assigns some probability to the occurrence of values more extreme than those used in the experiment. Moreover, the asymptotic approximation used in our paper sometimes overestimates the size of this bias (though these are not as large in the corrected version of Figure 5b as they previously appeared). These overestimations, moreover, are severe mainly in the case of the most infrequent numbers, so that even under the asymptotic approximation, extreme estimates of the average of a sequence of five numbers by our model are very rare, as shown in Figure (ii).

We thank the referee for pressing to clarify these points. We believe that the additional results summarized here should offer important clarifications to the reader. In the revised version of the manuscript, we have therefore included them

in a new section of the Methods (p. 26-30), to which we also refer in the main text (p. 13).

Reference:

Prat-Carrabin, A. and Woodford, M. (2021), Bias and variance of the Bayesian-mean decoder. *arXiv:2105.13705 [q-bio.NC]*.

Decision Letter, third revision:

8th February 2022

Dear Arthur,

Thank you for submitting your revised manuscript "Efficient coding of numbers explains decision bias and noise" (NATHUMBEHAV-20029220C), and for your patience while waiting for a decision. It has now been seen by one of the original referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,
Aisha

Aisha Bradshaw, PhD
Senior Editor
Nature Human Behaviour

Reviewer #2 (Remarks to the Author):

I would like to thank the authors for their detailed response. And now I'm totally happy with the paper. I'm sorry for the extra effort that I caused. But at least you found a mistake, so hopefully it was worth it.

Final Decision Letter:

Dear Arthur,

We are pleased to inform you that your Article "Efficient coding of numbers explains decision bias and noise", has now been accepted for publication in *Nature Human Behaviour*.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors whose manuscript was first submitted on or after January 1st, 2021, may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. IMPORTANT NOTE: Articles submitted before January 1st, 2021, are not eligible for Open Access publication. [Find out more about Transformative Journals](#)

Authors may need to take specific actions to achieve [compliance](#) with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to [Plan S principles](#)) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including [self-archiving policies](#). Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see <http://www.nature.com/nathumbehav/info/gta>). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at <https://www.nature.com/reprints/author-reprints.html>. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

With best regards,
Aisha

Aisha Bradshaw, PhD
Senior Editor
Nature Human Behaviour