nature human behaviour

SUPPLEMENTARY INFORMATION

DOI: 10.1038/s41562-017-0215-1

In the format provided by the authors and unedited.

The idiosyncratic nature of confidence

Joaquin Navajas[1,2*], Chandni Hindocha[1,3], Hebah Foda[1], Mehdi Keramati[4], Peter E. Latham[4] and Bahador Bahrami[1]

[1]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, WC1N 3AZ, UK. [2]Universidad Torcuato Di Tella, Av. Figueroa Alcorta 7350, Buenos Aires, C1428BCW Argentina. [3]Clinical Psychopharmacology Unit, University College London, Gower Street, London WC1E 6BT, UK. [4]Gatsby Computational Neuroscience Unit, University College London, 25 Howland Street, London W1T 4JG, UK. *e-mail: joaquin.navajas@utdt.edu

The idiosyncratic nature of confidence

Supplementary Information

Joaquin Navajas[1,2], Chandni Hindocha[1,3], Hebah Foda[1], Mehdi Keramati[4], Peter E Latham[4], & Bahador Bahrami[1]

[1]Institute of Cognitive Neuroscience, University College London, London, UK

[2]Universidad Torcuato Di Tella, Buenos Aires, Argentina

[3]Clinical Psychopharmacology Unit, University College London, London, UK

[4]Gatsby Computational Neuroscience Unit, University College London, London, UK

Supplementary Notes

Does goodness of fit explain our findings?

We asked if individual differences in how well our model fit the decisions could explain the inter-individual variability in the parameters β_p, β_I, and α_3. To do this, we correlated these values with the deviance[1], a standard metric of quality of the fit,

$$D = -2(\mathcal{L} - \langle\mathcal{L}\rangle), \qquad \text{[S1]}$$

where $\mathcal{L}$ is the log likelihood of the data, obtained through Equation [3] and $\langle\mathcal{L}\rangle$ is the log likelihood of data that perfectly fits the model (often referred to as a saturated model). In our case, $\langle\mathcal{L}\rangle$ is found by replacing the decision dependent terms in Equation [3] (those that depend on d_k) by their probability under the model, leading to

$$\langle\mathcal{L}\rangle = \sum_{k=1}^{N_{tr}} \Phi\left(\frac{\bar{\mu}_{30,k}(\lambda)}{\sigma_{30,k}(\lambda,\gamma)}\right)\log\left[\Phi\left(\frac{\bar{\mu}_{30,k}(\lambda)}{\sigma_{30,k}(\lambda,\gamma)}\right)\right] + \left[1-\Phi\left(\frac{\bar{\mu}_{30,k}(\lambda)}{\sigma_{30,k}(\lambda,\gamma)}\right)\right]\log\left[1-\Phi\left(\frac{\bar{\mu}_{30,k}(\lambda)}{\sigma_{30,k}(\lambda,\gamma)}\right)\right]. \qquad \text{[S2]}$$

Our three parameters, β_p, β_I and α_3, were uncorrelated with the deviance, D (r=0.22, p=0.24 for β_p; r=-0.12, p=0.54 for β_I; r=0.24, p=0.19 for α_3), and D was uncorrelated with average performance (r=0.22, p=0.23). This indicates that individual differences in β_p, β_I, and α_3 are not explained by inter-individual variability in the goodness of the fit.

Do our findings depend on the assumptions of the stochastic updating model?

We assumed that subjects were able to compute the mean and variance following Equations [1] and [2]. To evaluate whether or not the idiosyncrasies in confidence depended on these assumptions, we considered a different model, one without the subject-to-subject distortions (introduced by λ and γ) in the computation of $\hat{p}(\text{correct})$ and Fisher information. We set the mean value of μ_{30} on trial k (Equation [4]) to the true average orientation, and the perceived variance (Equation [2]) to the true variance. We took the inverse of the true variance to obtain trial-to-trial estimates of Fisher information, and used Equations [5-9] to compute $\hat{p}(\text{correct})$.

We regressed these estimates against confidence (Equation [10]) and obtained very similar results to our main study. Both β_p (r=0.95, p=10^{-16}) and β_I (r=0.98, p=10^{-20}) were highly correlated across models.

We also tested an alternative model, in which we relaxed the assumption of an ideal observer, and instead assumed that subjects computed the variance the same way they computed the mean,

$$\sigma_i^2 = (1-\lambda)\,\sigma_{i-1}^2 + \lambda\,\theta_i^2. \qquad [S3]$$

We computed $\hat{p}(\mathrm{correct})$ and Fisher information using Equations [4-9] and [S3], and regressed these values against confidence. Again, our findings were very consistent across models (r=0.99, p=10^{-26} for β_p and r=0.98, p=10^{-20} for β_I). This analysis confirms that our findings did not depend on the specific assumptions of the stochastic updating model.

Neuronal encoding of all functions of variance are fundamentally indistinguishable

In our analysis, we quantified participants' certainty in the estimate of the mean using the observed Fisher information. We used Fisher information, rather than standard deviation or variance, only because it provided the best linear fits to confidence reports in our Control Experiment (see Methods). Is there a more principled way to choose a function of uncertainty? For instance, could we determine which one is used by the brain? The answer to the latter question turns out to be no: even with neuronal recordings, it would be impossible to distinguish which function is encoded by the brain. Indeed, if the brain encodes one function of variance, it automatically encodes all functions of variance. For example, if a neuronal population encodes Fisher information, it automatically encodes variance,

$$p(I_{30}|\mathbf{r}) = p(\sigma_{30}^2|\mathbf{r})\left|\frac{d\sigma_{30}^2}{dI_{30}}\right| = p(\sigma_{30}^2|\mathbf{r})\,\sigma_{30}^4, \qquad [S4]$$

where $\mathbf{r}$ is the population response. Equation [S4] implies that even if we recorded the population activity, $\mathbf{r}$, we would be unable to distinguish whether the brain encodes Fisher information or variance. The same analysis applies to all functions of variance.

Correlation with objective performance

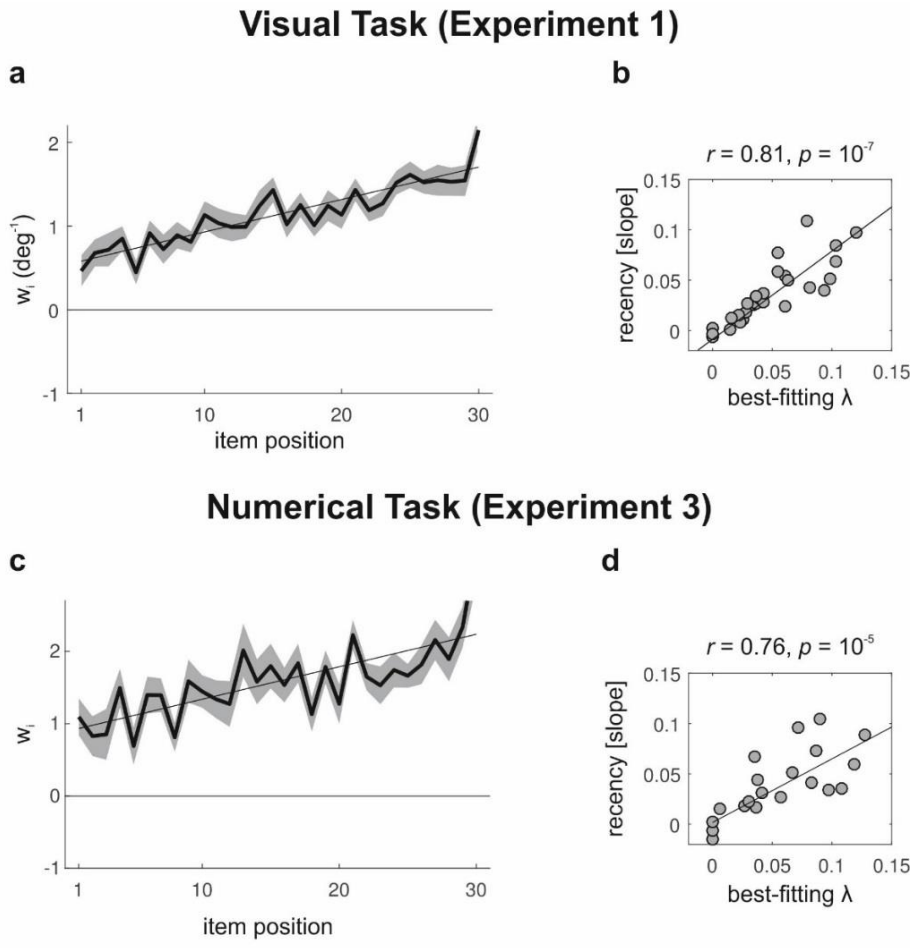
We asked if our three model parameters (β_p, β_I and α_3) were correlated with the average task performance. We did not find any correlation for β_I (r=0.25, p=0.18) or α_3 (r=0.21, p=0.27), but we found that β_p was correlated with task performance (r=0.55, p=0.002). This is consistent with previous studies showing that participants with larger objective performance typically show larger correlation between confidence and their probability of being correct[2].

This raises a potential concern: the stability of β_p over time and across tasks might simply reflect the stability of performance. To evaluate this possibility we computed the partial correlation of β_p across experiments after controlling for the mean performance on each task and observed that β_p was still stable over time (r=0.63, p=0.025) and across domains (r=0.63, p=0.005). This finding suggests that even though β_p correlates with performance, it still reflects an idiosyncratic property of confidence reports that is stable over time and across tasks involving uncertainty in different domains.

Controlling for individual differences in eye movement

We analysed electrooculography (EOG) data collected on 20 subjects while they performed Experiment 1. To measure individual differences in the amount of eye movement, we computed the EOG power (mean squared amplitude) on each trial and averaged this quantity across trials. We found that the EOG power did not correlate with β_p (r=0.11, p=0.63), β_I (r=-0.07, p=0.75) or α_3 (r=0.35, p=0.12), nor was it correlated with average performance in the task (r=0.09, p=0.70).

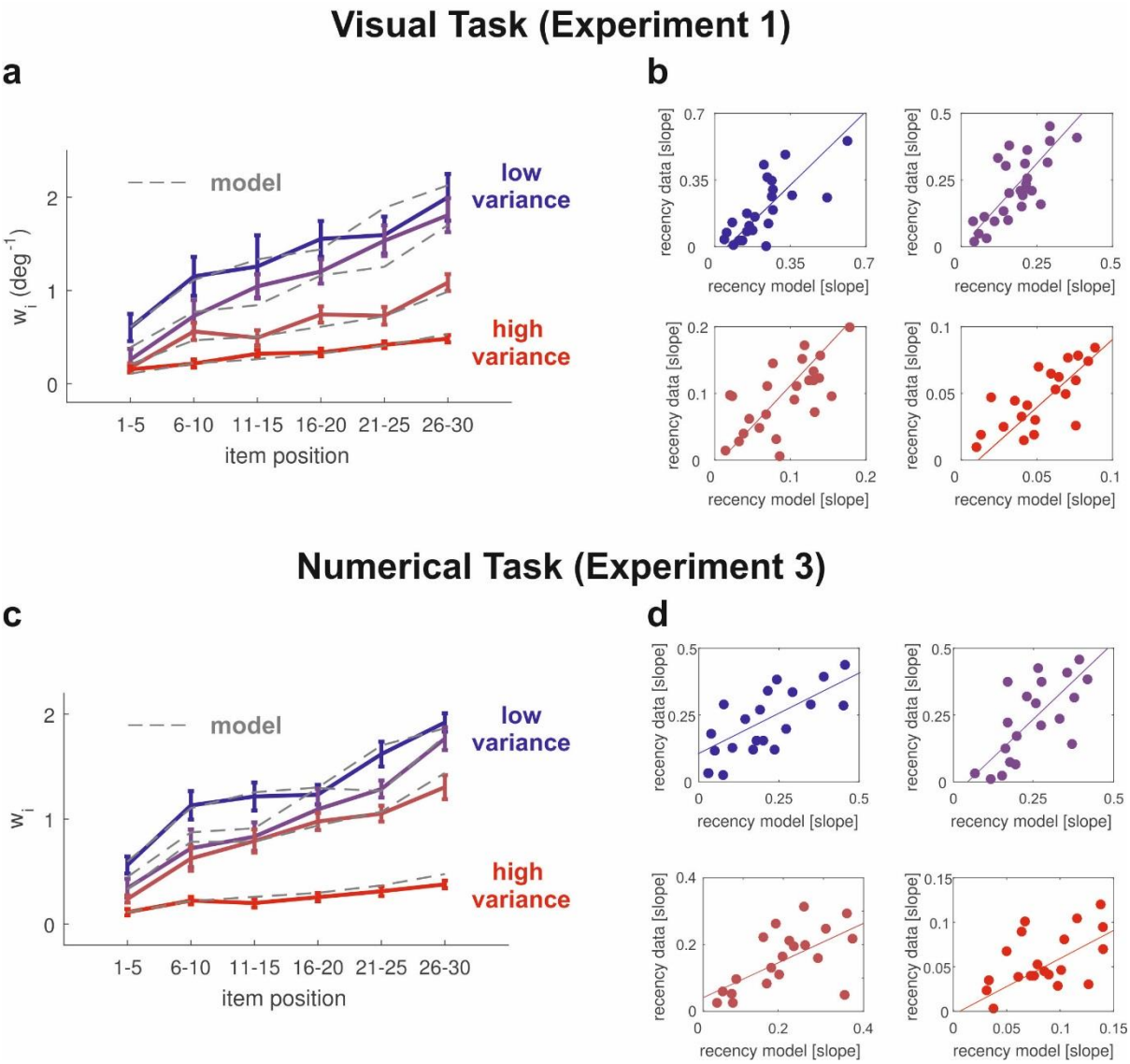
Supplementary Figures



Supplementary Figure 1. Recency effect. To test the influence that each Gabor patch (Experiment 1) or number (Experiment 3) exerted on choice, we implemented a multivariate logistic regression where the independent variables were the orientations/numbers presented at each position in the sequence (with positive items favouring the clockwise/greater option and negative items favouring the counter-clockwise/lower option), and the dependent variable was the probability of giving a clockwise/greater answer (for consistency with our notation, we define a variable, d, that is equal to 1 in clockwise/greater decisions and -1 in counter-clockwise/lower choices),

$$\log\left(\frac{p(d=1)}{1-p(d=1)}\right) = w_0 + \sum_{i=1}^{30} w_i\,\theta_i\,, \qquad \text{[S5]}$$

where w_i measures the weight that an item presented at position i had over choice. We ran this regression for all subjects separately. **a)** Average weights across subjects for the visual task (Experiment 1); the shaded area is the s.e.m. We observed that all items had a significant effect on choice ($t(29)>3.17$, $p<0.003$ for all item positions). We also observed a significant recency effect, which we quantified by fitting a line to the weights of each individual and comparing the distribution of slopes against zero ($t(29)=4.70$, $p=10^{-6}$). **b)** This recency effect is captured by our model and modulated by the parameter λ in Equation [1]; larger values of λ (x-axis) lead to a larger influence of recent items (slope of the regression, y-axis). Each grey dot is a different participant of Experiment 1. We observed that subjects with a larger recency effect (quantified by the slope in the regression) had a larger best-fitting λ ($r=0.81$, $p=10^{-7}$). Importantly, the extent to which people focus on recent items, quantified by λ, does not correlate with the overall performance in the task ($r=-0.28$, $p=0.13$), and it was also uncorrelated with the best-fitting parameters of our model of confidence ($r=0.25$, $p=0.17$ for β_p, $r=-0.03$, p=0.85 for β_I, and $r=0.15$, $p=0.42$ for α_3). **c-d)** Same as **a-b)** but for the numerical task performed in Experiment 3. **c)** All items had a significant effect on choice ($t(19)>2.4$, $p<0.03$ for all item positions). The recency effect was also significant, as quantified by the distribution of best-fitting slopes ($t(19)=3.81$, $p=10^{-3}$). **d)** The parameters λ correlate with the recency effect quantified by the best-fitting slope of the regression weights ($r=0.76$, $p=10^{-5}$).

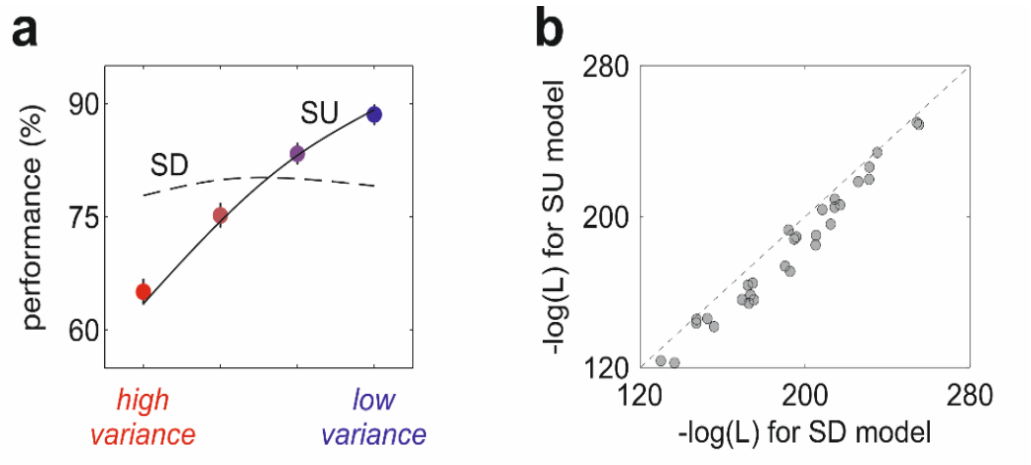


Supplementary Figure 2. Influence on choice for different variance conditions. To test if subjects integrated items differently depending on the variance of each trial, we implemented a multivariate logistic regression separately for each variance condition. To prevent overfitting, we considered a regression where the weights changed every 5 items,

$$\log\left(\frac{p(d=1)}{1-p(d=1)}\right) = w_0 + \sum_{i=1}^{6} w_i \left(\sum_{j=5(i-1)+1}^{5i} \theta_j \right). \qquad [S6]$$

This is very similar to Equation [S5]; the main difference (besides the grouping into 5 weights) is that we estimated the weights, w_i, for each variance condition separately. **a)** Visual Task (Experiment 1): Weights for each variance condition, averaged over subjects; error bars are

s.e.m., and colours code for different variance conditions as in the main figures. Presentations in the low-variance condition had larger influence over choice, and, as in Supplementary Figure 1, later items had larger weights than early items (2-way repeated measures ANOVA; effect of item position, $F(5,29)=16.19$, $p=10^{-12}$; effect of variance condition $F(3,29)=57.8$, $p\sim0$). We asked if these findings were consistent with our model. To test this, for each subject we found the best-fitting parameters λ and γ, as described in Methods, and used those to compute, on each trial, the probability of a clockwise option, $p(d=1)$. We then used that in the left-hand side of Equation [S6], and ran standard linear regression to find the model weights. The grey dashed lines show the model weights averaged across subjects. **b)** Recency effect estimated by the best-fitting slopes of the weights obtained from data versus model for each variance condition. Colours code as in panel **a**. Each dot is a different subject. The model weights matched well the weights computed from data ($r>0.71$, $p<10^{-5}$ for all four conditions). **c-d)** Same as **a-b** but for the numerical task (Experiment 3). **c)** We observed that later items had larger influence on choice ($F(5,19)=18.4$, $p=10^{-12}$) and that items had less influence if they had higher variance ($F(3,19)=19.4$, $p=10^{-8}$). **d)** The model captured individual differences in recency, quantified by the slope of the regression weights for each variance condition ($r>0.63$, $p<0.003$ for all conditions). This finding suggest that the last term in Equation [1], noise that scales with the size of the upcoming sample relative to the decision boundary (modulated by parameter γ), is not a property of the visual task but of the serial integration of items. To provide further support for this idea, we compared the best-fitting γ in both tasks (using the data collected in Experiment 3 and comparing the visual and numerical sessions) and observed a positive correlation ($r=0.80$, $p=10^{-5}$). This suggests that the subjects who had larger integration noise in one task also had larger integration noise in the other.



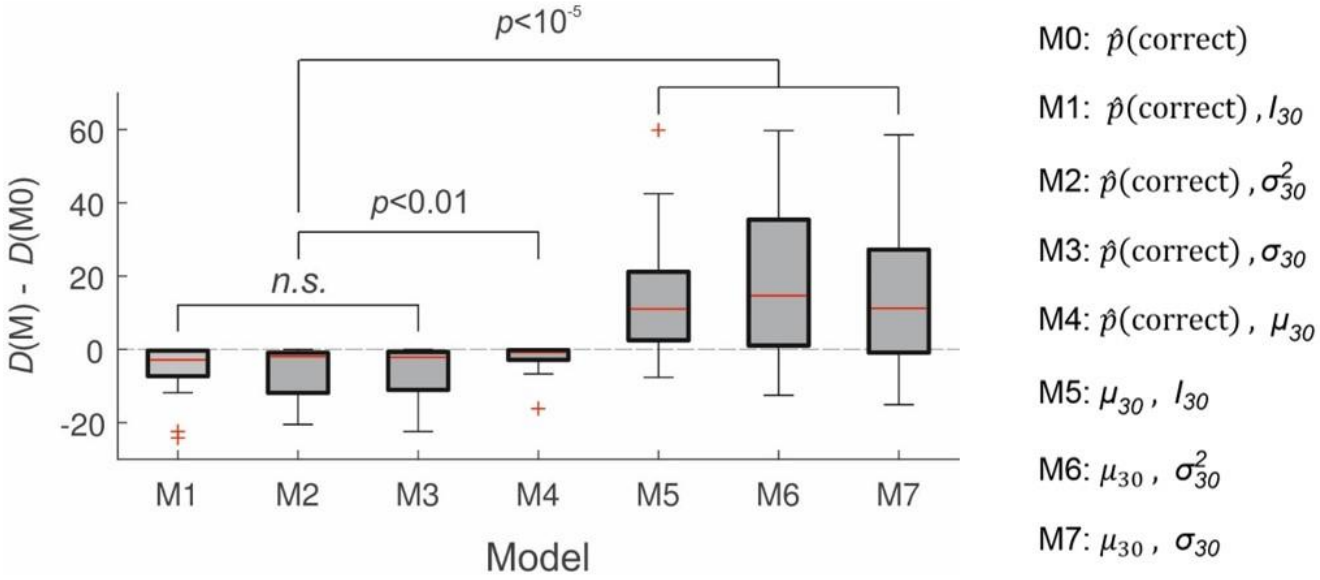
Supplementary Figure 3. Model fitting results in Experiment 1. We fit two probabilistic models that make different assumptions about how decisions are made. The stochastic updating (SU) model is described in the main text (Equations [1] and [2]). In the stochastic decision (SD) model, the agent makes deterministic updates,

$$\mu_i = (1-\lambda)\,\mu_{i-1} + \lambda\,\theta_i \qquad [S7]$$

and then makes a softmax decision,

$$p(d=1) = \frac{\exp(-\mu_{30}/\tau)}{\exp(-\mu_{30}/\tau) + \exp(\mu_{30}/\tau)} \qquad [S8]$$

where $p(d=1)$ is the probability of choosing clockwise and τ is the temperature of the *softmax* rule. In this model, the agent updates perfectly and uses a stochastic (and thus suboptimal) rule for action selection; errors are due to noise in the decisional stage. In the SU model, the updating process is stochastic (Equation [1] in the main text), and decisions are optimal based on the perceived estimate; errors are due to uncertainty in the updating process. Both models fit two parameters to the data of each individual. **a)** The SU model (solid line) but not the SD model (dashed line) fits the pattern of increasing performance with decreasing variance. **b)** Model comparison: negative log likelihood of the SU and SD models using the best fitting parameters. Each dot is a different participant. The SU model fits the data significantly better than the SD model ($t_{(29)}$=9.0, $p<10^{-9}$).

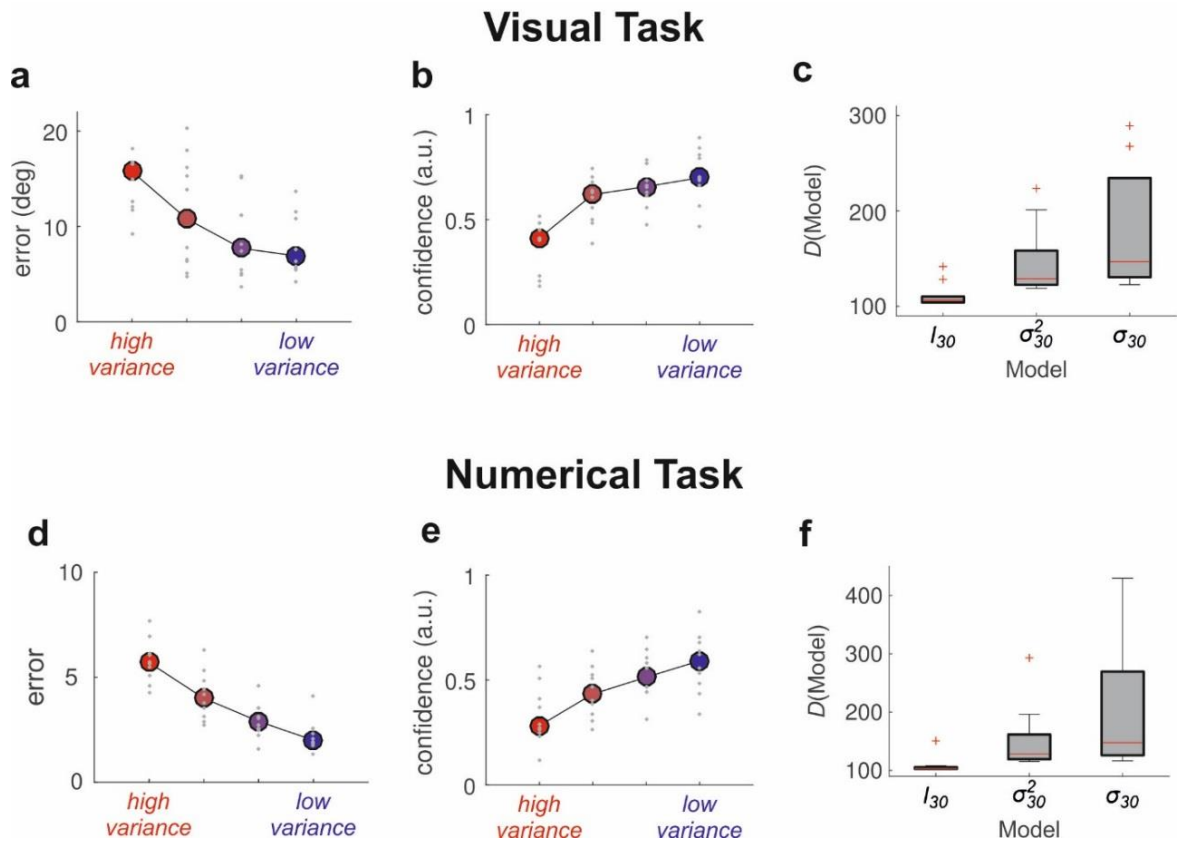


Supplementary Figure 4. Probing different models of confidence. Normative models propose that confidence should be a function of only $\hat{p}(\text{correct})$. We compared such a model (M0) with 7 alternative models which linearly combine two different probabilistic quantities (ordinal regression, see Equation [10] in **Methods**). Models M1 to M3 are extensions of M0 using a function of variance: they are based on $\hat{p}(\text{correct})$ and a second quantity (M1: Fisher information, M2: variance, M3: standard deviation). Model M4 is a different extension of M0 based on $\hat{p}(\text{correct})$ and the perceived mean. Models M5 to M7 are alternative models to M0 that linearly combine the perceived mean with Fisher information (M5), variance (M6), or s.d. (M7). The y-axis shows the difference in deviance between the extended/alternative models and M0. The difference in deviance is defined as two times the negative log-likelihood ratio,

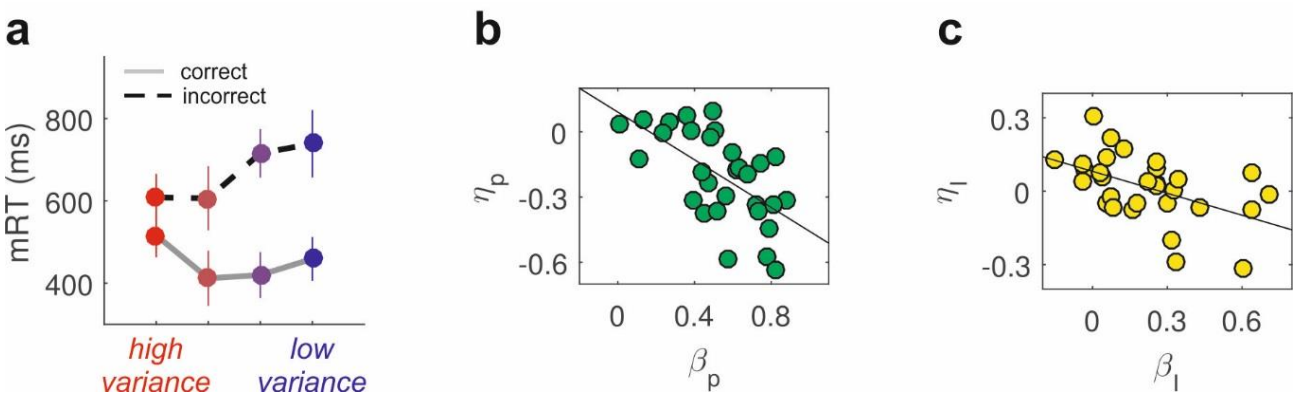
$$D(\mathrm{M}) - D(\mathrm{M0}) = -2\sum_{i=1}^{400}\log\left(\frac{p(d_i|\mathrm{M})}{p(d_i|\mathrm{M0})}\right), \qquad [S9]$$

where $p(d_i|\mathrm{M})$ is the probability of observing decision d_i given model M. More negative values provide stronger support for the extended/alternative model compared to M0. The boxplots show the distribution of difference in deviance for the 30 subjects in Experiment 1 (red line: median; box limits: 25 and 75-percentiles, whiskers at 1.5 times the interquartile range, red

crosses: outliers). We observed that models M1 to M3 were significantly more likely than M0 (Wilcoxon sign rank test, $z > 4.7$, $p < 10^{-5}$ for all pairwise comparisons to M0; log likelihood ratio test, $\Delta df = 30$, $p \sim 0$), but not significantly different from each other ($z < 1.7$, $p > 0.1$ for all pairwise comparisons between M1, M2 and M3). The model based on $\hat{p}(\mathrm{correct})$ and the perceived mean (M4) was more likely than M0 ($z = 4.7$, $p = 10^{-5}$, log likelihood ratio test, $\Delta df = 30$, $p = 10^{-14}$) but less likely than M1, M2, or M3 ($z > 2.7$, $p < 0.006$ for all pairwise comparisons to M4). All alternative models based on the perceived mean and a function of variance (M5 to M7) were significantly less likely than M0 ($z > 3.2$, $p < 0.002$ for all pairwise comparisons to M0). This finding indicates that confidence is not well fit by a linear combination of mean and variance (or mean and Fisher information or s.d.). Altogether, this analysis suggests that confidence is better explained by a linear combination of $\hat{p}(\mathrm{correct})$ and a function of variance.



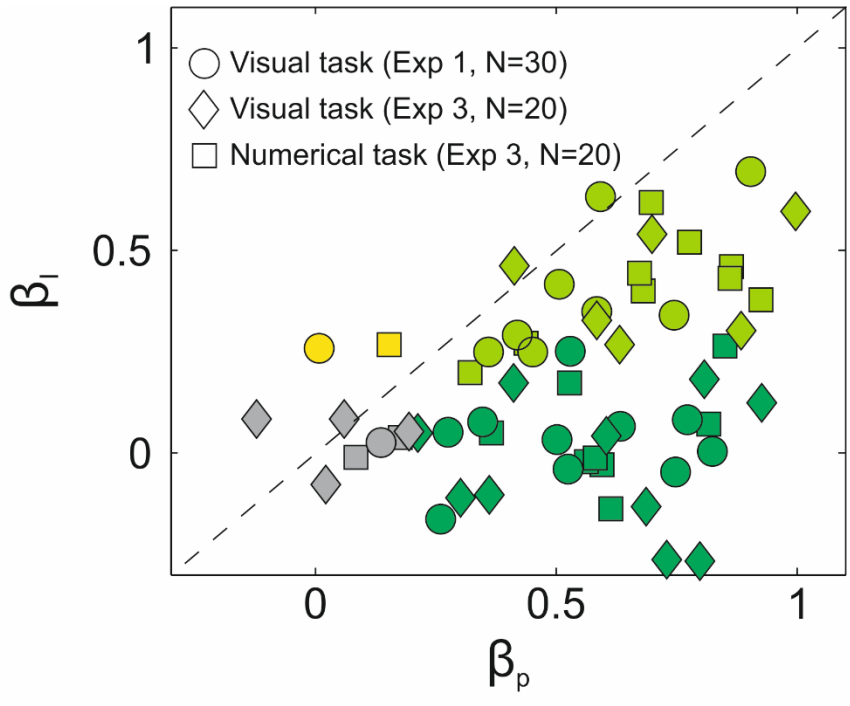
Supplementary Figure 5. Control Experiment. We asked if Fisher information correlates with confidence or other functions of variance. (**a-c**): Visual task. (**d-f**) Numerical task. Participants observed a sequence of items (Gabor patches in the visual task and two-digit numbers in the numerical task) serially flashed at the fovea at 4 Hz, and we asked them to provide their analog estimate of the mean (see **Methods**). We observed that, as we increased the variance in the sequence, responses became more accurate (panel **a** for the visual task ($F(3,9)=13.21$, $p=10^{-5}$), panel **d** for the numerical task $F(3,9)=3.8$, $p=0.003$) and more confident (panel **b** for the visual task, $F(3,9)=37.4$, $p=10^{-9}$, panel **e** for the numerical task, $F(3,9)=7.6$, $p=10^{-4}$). **c** and **f)** We regressed confidence against Fisher information (I_{30}), variance (σ^2_{30}), or standard deviation (σ_{30}) and measured the deviance of each model (see Equation [S9] in Supplementary Figure 4). The boxplots show the distribution of deviances for each model across subjects. In both tasks, the winning model was the one in which linear changes of Fisher information modulated confidence ratings (Wilcoxon sign-rank test, $z > 2.8$, $p < 0.005$ for both pairwise comparisons in the visual task, $z > 2.7$, $p < 0.006$ for the numerical task).



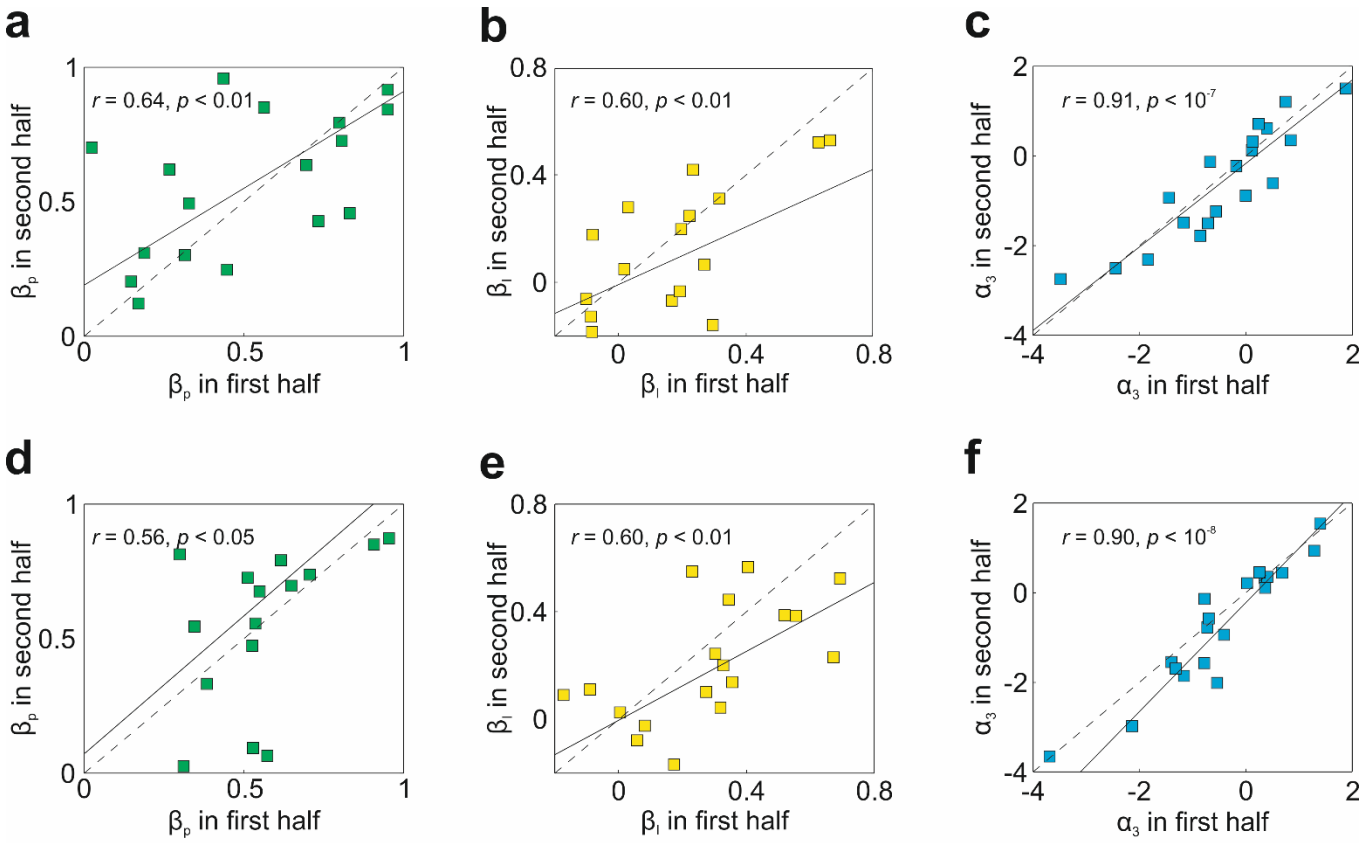
Supplementary Figure 6. Influence of $\hat{p}(\text{correct})$ and Fisher information on reaction times and confidence reports. a) Mean reaction times (mRT) averaged across participants for each variance condition, separated into correct and incorrect trials. Horizontal lines show the s.e.m. We observed a significant effect of outcome (correct vs. incorrect, F(1,29)=40.6, p=10^{-7}), a non-significant main effect of variance (F(3,29)=0.49, p=0.69), and a significant interaction (F(3,29)=4.3, p=0.007). **b)** We regressed reaction times against $\hat{p}(\text{correct})$ and Fisher information. To do this, we used Equation [10], except with reaction time rather than confidence on the left hand side,

$$\log\left(\frac{p(RT > j)}{1 - p(RT > j)}\right) = -\upsilon_j + \eta_p Z_p + \eta_I Z_I \qquad \text{[S10]}$$

where $p(RT > j)$ stands for the probability of observing a reaction time larger than the j^{th} sextile in the distribution. The influence of $\hat{p}(\text{correct})$ on confidence (β_p, x-axis) was significantly correlated with the influence of $\hat{p}(\text{correct})$ on reaction times (η_p, y-axis) (r=-0.61, p=10^{-4}). c) The influence of Fisher information on confidence (β_I, x-axis) was significantly correlated with the influence of Fisher information on reaction times (η_I, y-axis) (r=-0.49, p=0.005). We also observed a non-significant correlation between β_p and η_I (r=-0.06, p=0.75) and between β_I and η_p (r=-0.15, p=0.41). These findings suggest that the contribution of $\hat{p}(\text{correct})$ and Fisher information to confidence is not simply reflected in confidence reports, but also in reaction times. The negative correlation between the regressors is consistent with the idea that confidence might be, at least partially, based on decision time[3].



Supplementary Figure 7. Analysis of confidence across domains. Same as the main panel in Fig. 3 of the main text, except that both tasks of Experiment 3 are also included here. Regression weights on confidence for different individuals. x-axis: weight of the probability of being correct (β_p); y-axis: weight of information (β_I). Each marker (circle, diamond, or square) represents one experiment. The colour codes for significance (at the 0.05 level) are as follows: dark green, only β_p was significant; light green, both β_p and β_I were significant; yellow, only β_I was significant; grey, neither was significant. Circles: 30 participants performing the visual task in Experiment 1. Diamonds: 20 other participants performing the visual task in Experiment 3. Squares: the same 20 participants of Experiment 3 performing the numerical task.



Supplementary Figure 8. Stability in Experiment 3. Stability within each experiment for the visual (**a-c**) and numerical (**d-f**) task. For each half of the experiment (200 trials each), we decomposed confidence in terms of the weight of $\hat{p}(\text{correct})$ (β_p), the weight of information (β_I), and the overall confidence (α_3). Correlation across halves for β_p (**a**/**d**), β_I (**b**/**e**), and α_3 (**c**/**f**). Each square is a different participant, the dotted line is the identity, and the value of *r* given in each box is the Pearson correlation coefficient. All three variables are stable within each experiment for both the visual and numerical task.

References

1. Nelder, J. A. & Baker, R. J. Generalized linear models. *J. R. Stat. Soc.* **135**, 370-384 (1972).

2. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Frontiers in human neuroscience* **8** (2014).

3. Kiani, R., Corthell, L., & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron*, **84**, 1329-1342 (2014).