Selective overweighting of larger magnitudes during noisy numerical comparison

Bernhard Spitzer*[1,2], Leonhard Waschke[3], and Christopher Summerfield[1]

[1]Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK

[2]Department of Education and Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

[3]Department of Psychology, University of Lübeck, 23562 Lübeck, Germany

Published in Nature Human Behaviour **1,** 0145 (2017)
https://doi.org/10.1038/s41562-017-0145

*to whom correspondence should be addressed:
Corresponding author: Bernhard Spitzer (bernardodispitz@gmail.com), Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK Phone +44(0)1865 271321

Humans are often required to compare average magnitudes in numerical data, for example when comparing product prices on two rival consumer websites. However, the neural and computational mechanisms by which numbers are weighted, integrated and compared during categorical decisions are largely unknown[1–5]. Here, we show a systematic deviation from "optimality" in both visual and auditory tasks requiring averaging of symbolic number. Participants comparing numbers drawn from two categories selectively overweighted larger numbers when making a decision, and larger numbers evoked disproportionately stronger decision-related neural signals over the parietal cortex. Representational Similarity Analysis[6] showed that neural (dis)similarity in patterns of EEG activity reflected numerical distance, but that encoding of number in neural data was systematically distorted in a way predicted by the behavioural weighting profiles, with greater neural distance between adjacent larger numbers. Finally, using a simple computational model, we show that although suboptimal for a lossless observer, this selective overweighting policy paradoxically maximizes expected accuracy, by making decisions more robust to noise arising during approximate numerical integration[2]. In other words, although selective overweighting discards decision information, it can be rational for limited-capacity agents engaging in rapid numerical averaging.

Healthy humans (N=24) viewed sequentially occurring symbolic numbers (samples, n = 10; range 1-6, uniformly sampled) drawn from two categories, indicating with a key press which category had the larger average (**Fig. 1a**). Categories were distinguished by their font color (red vs. green; visual condition) or the voice in which they were spoken (male vs. female; auditory condition). Fully informative performance feedback followed each response. Discrimination performance (visual: 68.3%±0.9%, auditory: 69.8%±1.2%) did not differ between auditory and visual conditions (Wilcoxon signed-rank test: p=0.17).

We used a simple psychophysical model to understand the rational policy for performing noisy numerical averaging in our task (see Methods). Model input X_i was the number occurring on each sample i normalized (for convenience) within the range [-1,1]. The model was parameterized to allow two potential sources of loss during averaging. The first, kappa (k), encoded a potential compression of the number line, allowing numbers to carry different weight in the decision: each sample X_i was transformed to a momentary decision value via a sign-preserving exponential function of the form $(X+b)^k$, where b is an additive offset parameter. Where $k < 1$, the transfer function has a sigmoidal form that compresses outlying values (e.g. $X_i >> 0$ or $X_i << 0$) relative to inliers (e.g. $X_i \approx 0$; **Fig. 1b**, light grey lines); the converse is true when $k > 1$ (see **Fig. 1b,** dark grey lines). The second source of loss was assumed to occur after processing of each sample, i.e. during numerical averaging or at the response itself[7,8]: to generate simulated model choices, we passed the difference in cumulative decision values for each category through a sigmoidal function with slope sigma (s), where higher values of s (e.g. low slopes) indicate more noise in neural computation.

We then used our simulations to explore how the accuracy-maximizing policy changes under different values of compression k and decision noise s. In the absence of noise (e.g. perfect averaging), the optimal policy leaves the numbers uncompressed ($k = 1$); other policies discard

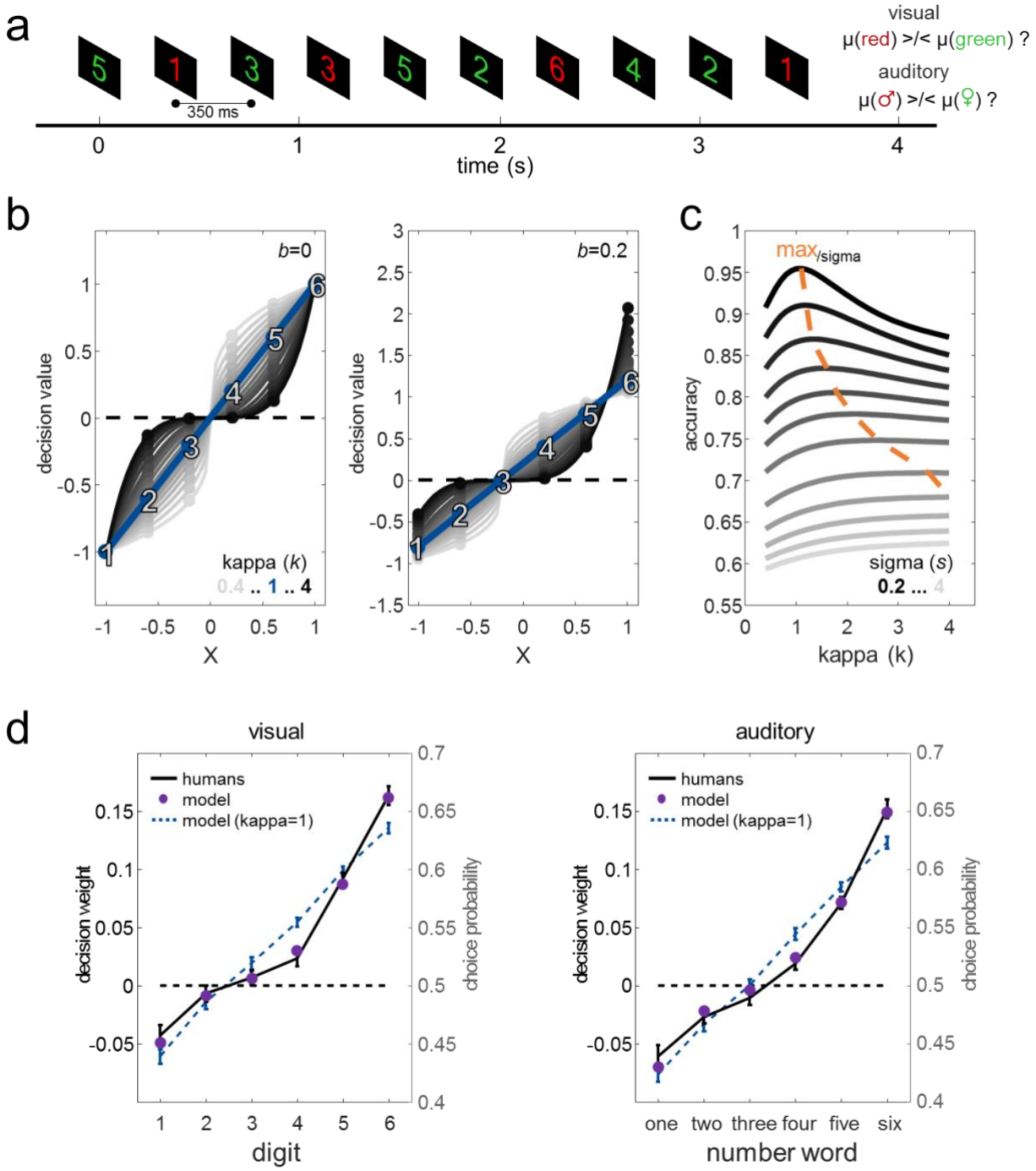


Figure 1. Task, model simulations, and human behaviour. a, Example trial sequence from the visual task. Ten numbers appeared in red or green font, separated by 350 ms. The task was to report whether the average of the red or green numbers was higher. **b**, Function mapping sensory inputs *X* onto a decision value *DV* = $(X+b)^k$ for different values of *k* (light grey lines, $k < 1$; dark grey lines, $k > 1$) and $b = 0$ (left panel) or $b = 0.2$ (right panel). **c**, Predicted accuracy under different values of *k* (x-axis) and integration noise *s* (lines), where light grey lines correspond to larger values of *s* (e.g. noisier decisions). Simulations are shown for $b = 0$ (see Fig. 2 for simulations with $b > 0$). **d**, Left panel: Decision weights for numbers 1-6 in the visual task. Black line shows human data (n=24) and purple dots show the predictions of the best-fitting model with $k = 1.95$ and $s = 1.75$ (mean estimates over subjects). Blue dashed line shows the fitted model predictions for $k = 1$. Right panel: as left, but for the auditory condition.

numerical information before averaging (**Fig. 1c,** dark grey lines). However, as integration noise increases (s >> 0), the accuracy-maximizing value of k increases (**Fig. 1c**, light grey lines); in other words, accuracy is maximized by giving more weight to outlying numbers (e.g. 1 and 6) than inlying numbers (e.g. 3 or 4), just as 'selective integration' has been shown to maximize accuracy in the presence of higher integration noise[2]. This was the case both under no bias (b = 0; **Fig. 1c**) and a bias towards overweighting larger numbers (b > 0; cf. **Fig. 2**). In the latter case, the accuracy-maximizing policy will give especially high weight to large outlying numbers (e.g., 6; right panel in **Fig. 1b**). In both cases, during noisy numerical averaging (e.g. where capacity is limited, and integration is leaky or imperfect), the best policy is to base choices principally on more extreme (outlying) values in the numerical sequence (i.e., k>1).

Turning to the human data, we examined choice probabilities to estimate the influence of each sample (number, 1-6) on the decision (**Fig. 1d**). In terms of absolute decision weight, in both the auditory and visual tasks, participants overweighted the higher numbers 5 (relative to 2; Wilcoxon signed-rank tests: both p < 0.001) and 6 (relative to 1; both p < 0.001) when making their choices. Furthermore, the weight functions deviated significantly from linearity (visual: $F_{5,115}$ = 16.60; auditory: $F_{5,115}$ = 17.35; both p < 0.001). This behaviour was captured by fitting the psychophysical model to the human data, maximizing the likelihood of choices at the single trial level (purple dots). Estimated values of *s* (integration noise) averaged about 1.8 (see below), a value at which values of k > 1 will maximize accuracy (**Fig. 1c**; see **Fig. 2a** for detailed simulations). Consistent with this policy, the obtained best-fitting values of k (**Supplementary Fig. 1a**) significantly exceeded 1 (visual: k = 1.95±0.19, auditory: k = 1.91±0.18; Wilcoxon signed-rank tests: both p < 0.001; difference between conditions, p=0.80) indicating an overweighting of outliers, with a positive offset bias (visual: b = 0.44±0.07, p < 0.001; auditory: b = 0.27±0.06, p < 0.001, which was slightly greater in the visual condition, p=0.013), confirming the preference for large (e.g. 6) over small (e.g. 1) numbers. Note that this "anti-compression" for large numbers is the opposite of what would be expected from scalar variability, i.e. if numbers were weighted according to Weber's law[9–12].

For comparison, the weights from an equivalent simulated observer with k = 1 (no compression) are shown in blue (**Fig. 1d**, dashed lines); quantitative model comparison indicated that these fit the data more poorly (Wilcoxon signed-rank test on AIC values: visual, p=0.002; auditory, p=0.004). Quantitative analysis also ruled out a model in which participants simply "counted" the larger numbers (see **Supplementary Information SI**, **Supplementary Fig. 1c**). The introduction of one further parameter encoding a leak in the integration process allowed the psychophysical model to capture the full pattern of decision weighting as a function of sample position (1-10) and numerical value (1-6) (i.e. 60 data points with 5 parameters; (**Supplementary Fig. 1b**). Inclusion of the leak both reliably improved the overall fits (Wilcoxon signed-rank tests on AIC values: both p<0.001) and reduced the best-fitting estimates of s (visual: 1.11 ± 0.08 vs. 1.75 ± 0.12; auditory: 1.29 ± 0.09 vs. 1.85 ± 0.12; Wilcoxon signed rank tests: both p < 0.001), suggesting that imperfect memory is itself a contributor to the cost of integration. Statistical tests with model and human as fixed factors showed no overall differences or interactions with experimental factors (all F < 3, all p > 0.05, corrected), confirming the ability of the model to capture human performance.

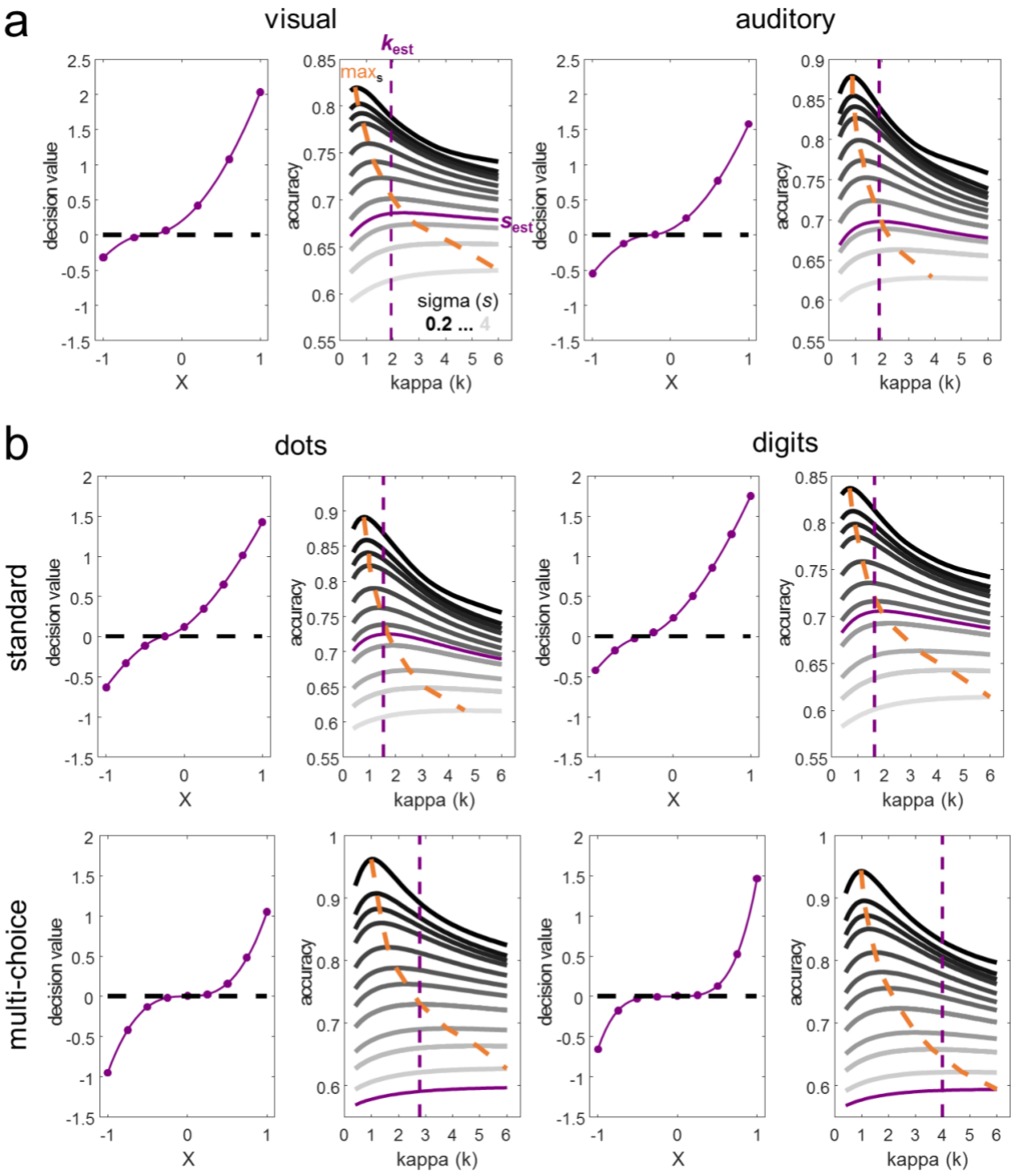


Figure 2. Overview of model results and simulations for each experiment and condition. Left panels illustrate best-fitting mapping function. Right panels show simulated model accuracy (same conventions as Fig. 1c) under the best-fitting parametrization in humans. Human kappa (k_{est}, dashed vertical) and sigma (s_{est}, solid) for each condition are highlighted in purple. **a**, Main experiment (cf. Fig. 1a,d). **b**, Supporting experiment (cf. Supplementary Fig. 2). In all conditions, k_{est} was increased (>1) towards maximum predicted accuracy under the estimated noise level (s_{est}). Note that in cases where k_{est} fell short of the theoretical maximum (dashed purple vs. dashed orange), the associated differences in predicted accuracy were relatively minor (< 1%).

These findings suggest that during numerical averaging, decision values are "anti-compressed" in precisely a way that will compensate for "late" noise in the integration process, and consequently maximise rewards (**Fig. 2a**). To directly test whether human decision policies adapt to the level of late noise in the task, we conducted a new experiment in which the cost of integration was manipulated directly in two distinct conditions. A fresh cohort of participants (N=21) viewed sequential number samples (**Supplementary Fig. 2a**, red and green digits or dots displays, n=8 samples, range 1-9) with instructions to compare averages along a single axis (e.g. red vs. green) or multiple axes (e.g. both red vs. green and digits vs. dots, see **Supplementary Information SI** for details). Fitting these data with the psychophysical model described above indicated that "late" noise was indeed lower in the former condition (the "standard" task: s=1.33 ± 0.22) than in the latter ("multi-choice" task: s=4.80 ± 0.82) with a significant difference between the two (Wilcoxon signed-rank test: p<0.001).

Replicating the finding of anti-compression in the presence of noise (s>>0), we found the best-fitting estimates of k in both tasks to be larger than 1 (**Fig. 2b,** see **Supplementary Information SI** for details), for both sample formats (digits and dots; all k >1.5, Wilcoxon signed-rank tests: all p<0.05, uncorrected). More importantly, in the multi-choice task, the estimates of k were significantly larger (mean 3.40 ± 0.48) compared to the standard task (mean 1.60 ± 0.20; 2x2 repeated measures ANOVA: main effect of task $F_{1,20}$=12.92, p=0.002). In other words, as we increased integration noise, the observed anti-compression also increased. We also again found evidence for a positive offset bias (see **Supplementary Information SI** for details), indicating that participants especially overweighted larger numbers (**Supplementary Fig. 2b).** Lastly, the analysis revealed no significant differences in any of the above effects between digits and dots displays (all $F_{1,20}$<3.88, all p>0.05), confirming that selective number integration occurred independent of presentation format (symbolic/non-symbolic, see above visual/auditory). Together, across all conditions under study, humans adopted a non-linear sampling policy that drives accuracy near to the model-predicted maximum, given their estimated noise level and bias (**Fig. 2a-b**).

In order to explore these effects at the neural level, we recorded EEG while participants performed the first experiment (**Fig. 1a**). All sequential samples were fully statistically independent, allowing us to analyze neural responses to individual numbers in the stream (see Methods). Consistent with previous research[13,14], we observed differences in the centro-parietal positive (CPP) response following the onset of each number, in the periods 290-700 ms (visual) and 500-800 ms (auditory) post-onset (**Fig. 3a**, all time bins p < 0.01, FDR corrected). In the visual modality (**Fig. 3b,** left panel), the CPP was relatively larger for number 6, reduced for sample 5, and smallest for all other numbers, as indicated by post-hoc tests (Wilcoxon signed-rank tests: 6 vs. 5, p = 0.008; 5 vs. 4, p < 0.001; whereas 4 vs. 3, 3 vs. 2, 2 vs. 1, all p > 0.70; Bonferroni-corrected). The auditory condition followed a similar pattern (**Fig. 3b**, right panel), albeit with noisier and lower-amplitude CPP effects (6 vs. 5, p = 0.28; 5 vs. 4, p = 0.004; 4 vs. 3, 3 vs. 2, 2 vs. 1, all p > 0.40; corrected). We fitted the neural amplitude modulations with the absolute decision weights obtained from the non-linear model (cf. purple dots in Fig. 1d), which provided a better fit than the linear model for both auditory and visual conditions (both p<0.02, Wilcoxon signed-rank test on regression deviances). Expanding the CPP analysis to encompass sample order, we observed no interactions between order and number (both F < 1.4, both p > 0.20), suggesting that the overweighting of larger numbers is invariant across sample positions (1…10).

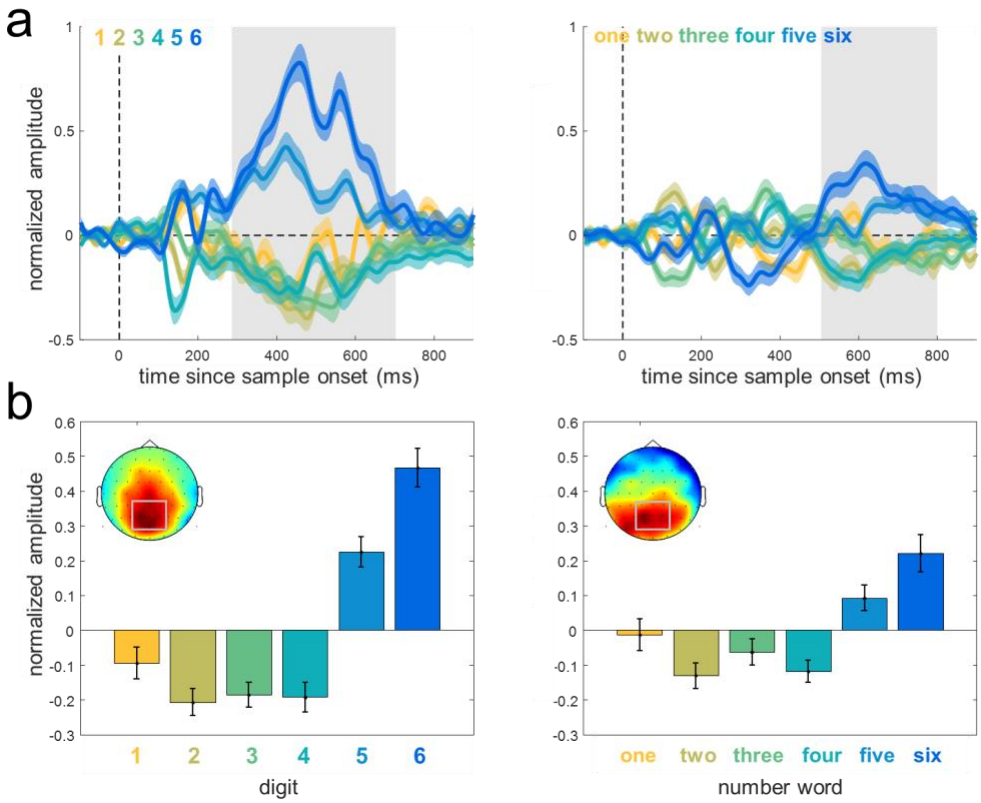


Figure 3. CPP analysis. Left panels: visual, right panels: auditory. **a**, Mean-subtracted EEG signals evoked by numbers 1-6 over centro-parietal electrodes, averaged across sample positions. Colored shadings show SEM. Gray shaded area indicates time window of significant CPP modulations identified by leave-one-out permutation (see Methods). **b**, Mean CPP amplitudes in the time windows identified in a. Error bars show SEM. Insets illustrate scalp topography for sample number 6.

Next, we employed a multivariate approach (RSA)[6,15] to probe the neural encoding of number in more detail, and link the neural representations back to categorization behaviour. We computed for each post-sample time-point the representational (dis)similarity in EEG signals for each number 1-6 in each of the two categories (12 x 12 representational dissimilarity matrix or RDM, based on Mahalanobis distance; see Methods). We then compared this to predicted RDMs that were created under the assumption that neural distance depended on (i) the physical properties of the digit, (ii) category membership (e.g. red vs. green font), (iii) parity (odd vs. even), (iv) numerical distance (i.e., the pairwise numerical difference between any two numbers, independent of category), and (v) category-dependent numerical distance (see **Supplementary Fig. 3a** for details). We used recursive orthogonalization (see Methods) to ensure that each model RDM explains unique variance in the observed neural RDM from human subjects.

In **Fig. 4a**, we plot the time course of correlations (Kendall's Tau) between the 5 model RDMs and the human RDM for each subject in the visual condition. The neural patterns were dominated by a category-independent numerical distance effect (**Fig. 4a** purple) that was significant from approx. 200-700 ms post sample onset ($p_{cluster} < 0.001$; cluster-based permutation test); this can also be seen in the grand mean EEG RDM (**Fig. 4b**). However, additional effects of category (i.e., font color), parity, and a category-specific numerical distance effect (number x category) were also observed

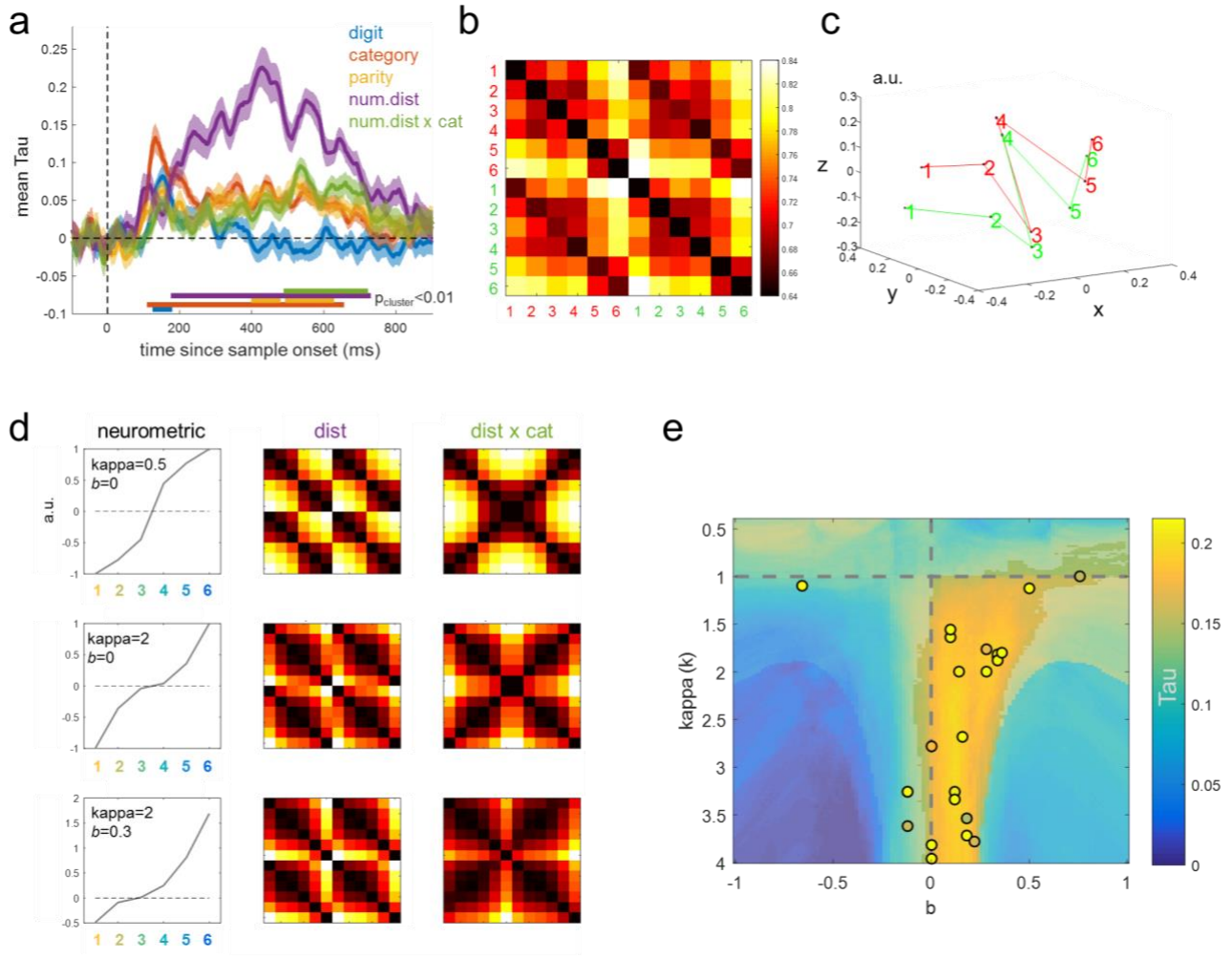


Figure 4. Representational similarity analysis (RSA). **a**, Time course of correlations (Kendall's Tau) between orthogonalized model RDMs for different sample features (see Supplementary Fig. 3a) and the observed EEG-RSA patterns following each sample in the visual condition. Colored shadings show SEM. Marker lines on bottom indicate significant differences from zero. For auditory results, see Supplementary Fig. 3b. **b**, Grand-mean EEG-RDM for a representative time window (200-600 ms) in the visual condition. **c**, 3D-illustration of the first three dimensions of a multidimensional scaling of the EEG-RDM shown in b. **d-e**, Neurometric mapping functions estimated from EEG-RSA**. d**, Unidimensional model RDMs (*middle, right*) predicted under different parametrizations of hypothetical neurometric mapping functions (*left*). **e**, Grand mean correlations (collapsed over distance- and distance x category RDMs from d) between model RDMs and observed EEG-RDM over values of kappa *k* and bias *b*. Dashed grey lines delineate the parameter space for $b = 0$ (symmetric mapping) and $k = 1$ (linear mapping). Maximum mean correlations were observed for values of $b > 0$ and $k > 1$ (see Results). Transparent mask highlights parameter space of significant positive correlation with both the distance- and the distance x category RDM in d (Wilcoxon signed-rank tests; all pixels $p_{conjunction}<0.001$, uncorrected, and exceeding the symmetric-linear model). Dots show maximum mean Tau for each participant (some dots are covered by others).

(all $p_{cluster} < 0.01$), with the latter peaking late (**Fig. 4a**, green), consistent with a response-mapped representation. To further visualize these effects, we reduced the dimensionality of the dissimilarity matrix via multidimensional scaling. Visualizing the first 3 dimensions showed clear effects of numerical distance (x-dimension), category (z-dimension) and parity (y-dimension) (**Fig. 4c**). Visual inspection of the grand mean EEG RDM (**Fig. 4b**) may suggest that a numerical distance effect might

have arisen mostly by dissimilarity of numbers 6 and 5 from the remaining numbers (1-4; cf. CPP analysis, Fig. 3). Interestingly, however, we found a statistically significant effect even when restricting the analysis to numbers 1-4 (324-574 ms, $p_{cluster} < 0.001$). In other words, multivariate RSA disclosed aspects of a number-line representation that were invisible to conventional parietal evoked signals (cf. **Fig. 3**; see also **Supplementary Fig. 3c).** However, we observed no systematic EEG-RSA effects in the auditory condition (**Supplementary Fig. 3b**).

Having established a numerical distance effect in multivariate EEG patterns, we next asked whether the neural data predicted the distortions in numerical coding observed in the behavioural weighting profiles (cf. **Fig. 1d**). To test this, we estimated the best-fitting "neurometric" mapping function predicted from the EEG-RSA patterns, by generating model RDMs from hypothetical mapping functions parameterized by k and b (**Fig. 4d**). We exhaustively searched over values of k (0.4 to 4) and b (-1 to 1) and correlated the predicted model RDMs (both for distance- and distance x category effects) with the EEG-RSA pattern in each participant. The best-fitting parametrizations (in terms of maximum mean Kendall's Tau correlation) were characterized by values of $k > 1$ (mean 2.52 ± 0.20, Wilcoxon signed-rank test: $p < 0.001$) and $b > 0$ (mean 0.15 ± 0.05, p=0.005) (**Fig. 4e**). In other words, the neurometric number mapping inferred from EEG-RSA mirrored both key aspects of the psychometric mapping inferred from choice behaviour: (i) exponential overweighting of outlying samples (i.e., anti-compression), and (ii) an overall weighting bias towards large numbers. Together, these results show a strong correspondence of model simulations, choice behaviour, and sample-level neural representations in demonstrating "optimal irrationality"[2] or "rational inattention"[16] in the presence of noise during sequential information integration.

The present findings build upon recent work in which humans compare the average magnitude in two streams of visual items occurring in parallel (e.g. side-by-side on the screen). In this setting, humans tend to ignore or downweight the locally weaker of the two simultaneously-occurring samples[1], and this behaviour can similarly be accounted for with a selective weighting policy that systematically discards decision information. Given a selective weighting policy, it is possible to construct equally-valued streams A, B and C such that participants will systematically choose A > B, B > C and C > A, i.e. show a classic violation of the rational axiom of transitivity[2]. Nevertheless, there as here, selective weighting maximizes accuracy – i.e. is rational – if one assumes that noise corrupts information integration[2]. We note in passing that the selective weighting policy observed in our experiments tended to overweight larger numbers (e.g. 6) rather than all outliers (e.g. 1 and 6) as would be predicted by the rational policy under late noise. We leave it to future research to determine whether the offset bias that was observed in both experiments (although not in multi-choice conditions) depends in part on the framing of the task. Finally, recent work has extended the selective weighting framework to provide a normative account of the "robust averaging" (i.e. downweighting of outliers, not inliers) of decision information that occurs when stimulus feature values are distributed in an approximately Gaussian fashion across the experiment[17]. In all of these cases, humans seem to have evolved policies that discard information in order to increase the robustness of decisions in the face of noise corrupting the neural computations associated with information integration.

Behavioural signatures of decision weighting were also reflected in neural signals. The CPP is an evoked centro-parietal potential that has previously been shown to build up during information integration with an amplitude that reflects the strength of available decision information[13,14]. Here,

we observed a relatively larger CPP for numbers 5 and 6 in both the auditory and visual domains. The CPP is most likely related to the well-described P300 potential[18–20] and it may relate here to the detection of the information that is being used to form a decision[21] or to evaluation processes that unfold at the level of each individual sample[19]. The effects were discernable, but considerably more noisy, in the auditory domain compared to visual; this is probably related to unavoidable time-varying differences in the specific physical input associated with each speech sample. Together, these findings offer independent corroborating evidence for the strategy of selective overweighting that we observed in human behaviour.

The RSA results revealed a neural representation of an ordered 'number line' for numeric visual symbols. Similar representations of numerical magnitude have previously been reported for non-symbolic number, e.g. a dots display[22–24], but not for number symbols/digits[22–26]. It is striking that neural patterns recorded at the scalp encode numerical magnitude even when other potentially correlated factors have been accounted for (e.g. visual similarity in digits themselves), and future researchers may wish to harness this finding to reveal other aspects of human numerical cognition. Here, however, we emphasize that in the visual domain, the neural representation of the number line was distorted in exactly the way predicted by behavioural data. Interestingly, although it is theoretically possible that the CPP could account for the pattern similarity effects revealed with RSA, it seems unlikely that this is the case in our data. For example, we found that participants' "number line" in decision weighting explained RSA variance not accounted for by the CPP (**Supplementary Fig. 3c**). Secondly, we observed no RSA effects for the auditory condition, despite statistically reliable differences in the CPP. The reasons for this latter discrepancy is unclear; it could be related to the difficulty in establishing high-precision neural patterns in time-locked data for time-varying speech items, or it might reflect computational differences in the processing of auditory stimuli.

Methods

Participants.

24 healthy volunteers (12 females, 12 males, age 26.6 ± 2.8) participated in the experiment after giving written informed consent. The study was approved by the ethics commission of the Free University Berlin and was conducted in accordance with the Human Subjects Guidelines of the Declaration of Helsinki.

Stimuli, task, and procedure

On each trial, 10 numbers ("samples") were presented in sequence at a rate of 350 ms (**Fig. 1a**). In the visual condition, digits (font Arial, approx. 1.8° visual angle) were presented at fixation in either a green or a red font for 280 ms, followed by a 70 ms blank period. In the auditory condition, German number words were played with either a female or male voice. Speech samples were taken from a public repository (http://www.freesound.org), time-compressed to a common length of 350 ms (using the PSOLA algorithm in Adobe Audition®, San Jose, USA) and loudness-normalized. Each

sample was independently and randomly drawn with uniform probability from a pool of 12 possible items, consisting of the numbers 1…6 in each of the two categories (i.e. the number of samples drawn from each category was fully randomized). Following the offset of the final item in the sequence, participants were given 2s to indicate by key press (left/right hand, counter-balanced between participants) which of the two sample categories contained the higher average numerical value. Median response times averaged at 487 (visual) and 484 ms (auditory). Correct responses were after 100 ms rewarded with a bell ("bling") sound, errors were fed back by a "buzz" sound. After a brief wait period (500-1500 ms, randomly varied), the next trial started. The onset of a trial (500 ms before sequence onset), as well as the response periods (350 ms after the onset of the last item in a sequence), were signalled by a small central fixation point briefly changing its color (grey/white). Participants were instructed to maintain fixation throughout all trials (including the auditory condition), aided by a head support (SR Research Ltd., Ottawa, Canada) to avoid movements. After several practice runs, each participant performed 6 blocks of 100 trials (3 in each modality condition, in alternating order), providing 3000 sample presentations per modality and participant.

Psychophysical model

In our simulations, for convenience we defined X to range between -1 and 1 in six equidistant steps, corresponding to the six numerical magnitudes (1…6) used in the experiment. We characterized the mapping of sample information X onto a subjective decision value (dv) as a family of (sign-preserving) exponential functions

$$dv = \frac{X + b}{|X + b|} \cdot |X + b|^k \qquad \text{(eq. 1)}$$

where $k < 1$ implies a relative downweighting, and $k > 1$ a relative upweighting of outlying samples (**Fig. 1b**, *left*). The special case where $k = 1$ corresponds to a linear mapping, i.e. $dv = X$. Parameter b accounts for a potential asymmetric weighting bias, in terms of an offset of the mapping function relative to its indifference point ($dv = 0$). A value of $b = 0$ corresponds to a point symmetric mapping, whereas $b \neq 0$ implies an up- or downward shifted, asymmetric function (cf. **Fig. 1b**, *right*).

Our initial goal was to evaluate model performance across different values of k and b. Here, it is important to consider that different parameterizations of the mapping function in eq. 1 differ in the absolute decision value that is obtained by transformation of perceptual inputs X. This absolute decision value is in turn related to the probability of a correct choice being made (see eq. 4 below). Assuming that decisional gain is a limited cognitive resource (i.e., a quantity that should not change between the models we test, for example reflecting an upper limit on the number of spikes produced by the relevant neurons), we computed for each transformation (eq. 1) its multiplicative gain factor

$$g = \frac{\sum |f + b|^k}{\sum |f|} \qquad \text{(eq. 2)}$$

which quantifies the extent to which a feature space f (here, the six equiprobable values of X) is transformed into a dv-space whose absolute values are larger (or smaller) than the absolute values in f. Using g as normalization factor, the trial-level decision value DV is given by the sum over the 10 samples of each sequence (cf. **Fig. 1a**)

$$DV = \sum_{i=1}^{10} \frac{dv_i \cdot c_i}{g}$$

(eq. 3a)

where c_i is a dummy variable that encodes the category of a sample [e.g., *c(red)* = 1; *c(green)*= -1]. To additionally model a potential leak of decision value over time (**Supplementary Fig. 1b**), we extended formula (eq. 3a) by a simple exponential function over samples i[27]

$$DV = \sum_{i=1}^{10} \frac{dv_i \cdot c_i}{g} \cdot l^{10-i}$$

(eq. 3b)

Lastly, the trial-level DV was transformed into a choice probability using a logistic choice function with noise term sigma (s)

$$CP = \frac{1}{1 + e^{\frac{-DV}{s}}}$$

(eq. 4)

We refer to s as "late" or integration noise, denoting noise that occurs at processing stages downstream to perceptual sample encoding. Such noise could arise during integration or at the response itself, but we note that the compact parametrization of late noise in (eq. 4) is equivalent to adding a (constant) noise term to each dv_i.

To simulate model accuracy (**Fig. 1c**), model choices were generated by randomly drawing from a binomial distribution with p = CP, where CP was computed trial-by-trial according to (eqs. 1-4). When fitting human choice data, we included a constant term to account for potential motor biases (e.g. towards left vs. right responses). To avoid parameter instabilities, we fitted the model without gain normalization and rescaled s by dividing it by g, which warrants formal equivalence to eqs. 1-4. Parameter estimates (MLEs) were obtained by minimizing the negative log-likelihood of the model given each participant's single-trial responses across values of k(0.1 to 10), b(-1 to 1), s(0.01 to 8, unnormalized), and (in models with leakage) l (0 to 1). In two participants in the visual condition, the model without leakage (eq. 3a) yielded exceedingly large raw estimates of s. However, group level results were robust to either inclusion or exclusion of these participants. Quantitative model comparisons (e.g. between exponential and linear models) were corrected for model complexity based on the Akaike information criterion (AIC). To evaluate model performance against human choice behaviour, we again generated binomial model choices (cf. simulations), but now using the individual best-fitting model parametrizations, and the exact same sample sequences as presented in the human experiment. We then compared the choice data of human and model

observers using conventional statistical analyses. Choice probabilities associated with each sample number (1…6) were inferred from the relative frequency of choosing a sample's category (i.e. its color or speaker) at the end of a trial, and were transformed into estimates of (signed) decision weight with an indifference point at zero (i.e. decision weight = choice probability - 0.5; cf. **Fig. 1d**, dual y-axes). Evaluation against model-predictions was complemented by model-free tests for symmetry (comparing the absolute decision weights of 1 vs. 6, 2 vs. 5, and 3 vs. 4) and linearity (omnibus test of linear regression residuals across numbers 1…6) of the human weighting functions.

EEG recording and analysis.

We recorded 64-channel EEG (BioSemi ActiveTwo, Amsterdam, Netherlands) configured according to the extended 10–20 system. Ocular activity was recorded via adhesive electrodes placed in bipolar montages around the eyes (horizontal and vertical), and was additionally monitored using an Eyelink 1000 camera (SR Research Ltd.). EEG signals were digitized at 2048 Hz, off-line referenced to common average, filtered (0.5-45 Hz), and down-sampled to 256 Hz. The EEG was corrected for eye blinks using adaptive spatial filtering[28] and epoched around each individual sample (-100…900 ms relative to sample onset). Bad channels were identified by visual inspection. Residual artefacts were rejected by excluding epochs with amplitudes > 80 µV from analysis. The artifact-free epochs were baseline-subtracted (-100-0 ms) and smoothed with a sliding 50 ms Gaussian kernel. EEG analysis was performed using functions from SPM12 (build 6470) for M/EEG (www.fil.ion.ucl.ac.uk/spm/), FieldTrip (http://www.ru.nl/neuroimaging/fieldtrip), and custom MATLAB code (The Mathworks, Inc., Natick, USA).

Centro-parietal evoked responses (CPP)

All EEG analysis was performed on the individual sample level. In each participant and modality condition, the mean waveform (averaged over all samples) was subtracted from each individual epoch, effectively eliminating the stimulus-onset response to the current and subsequent samples (note that epochs overlapped with up to two subsequent sample onsets, cf. **Fig. 1a**). Epochs of the same sample type were averaged and subjected to conventional statistical analysis. Based on previous CPP findings[14,20,29] signals were pooled over centro-parietal channels (CP1, P1, POz, Pz, CPz, CP2, P2). Time windows for CPP analysis were identified using a leave-one-out procedure to preclude circular inference. For each participant, CPP amplitudes were averaged over those adjacent significant time bins ($p<0.01$; false-discovery rate corrected, FDR) that exhibited the strongest overall amplitude modulations in the remaining 23 participants, based on non-parametric omnibus tests over sample types.

Representational similarity analysis (RSA).

The pre-processed channel data were projected onto principal components retaining 99% of the variance[30]. Multivariate (dis-)similarity was assessed in terms of the pair-wise Mahalanobis distance between the mean-subtracted component patterns associated with each sample type (numbers 1…6 per sample category, yielding a 12x12 RDM at each time point), using the residual single-trial variance at each time point for noise normalization[31].

To test the extent to which sample information was encoded in the time-course of the EEG-RDM, we created hypothetical model RDMs for the following features of interest (**Supplementary Fig. 3a**):

(i) physical number, with minimum dissimilarity between identical numbers, and maximum dissimilarity between all other pairs, regardless of sample category; (ii) sample category, with minimum/maximum dissimilarity between same/different category samples; (iii) numerical parity, with minimum/maximum dissimilarity between numbers of the same/different parity (even, uneven); (iv) numerical distance, with dissimilarity linearly increasing as a function of the numerical difference between any two numbers, independent of sample category; (v) category-dependent numerical distance, where the encoding of numerical distance within each sample category is the same as in iv, but is inverted between the two categories (in terms of a numerical distance x category interaction, i.e. a 6 in one category is predicted to be similar to a 1 in the other category). The latter is expected to occur if numerical value representations were response-mapped, that is, if they systematically differed in driving left (e.g., "red") vs right (e.g., "green") key choices.

Each model RDM was recursively orthogonalized with respect to all other model RDMs using the Gram-Schmidt process (**Supplementary Fig. 3a**). Next, each model RDM was correlated with the EEG RDM at each peri-sample time point using Kendall's Tau correlation coefficient. All correlations were computed over the upper RDM triangle, excluding all redundant elements and the diagonal. Significant correlations were identified using cluster-based permutation tests[32] over time points (1000 iterations, cluster-defining threshold $p < 0.01$, Wilcoxon signed-rank tests, uncorrected). Subsequent analyses where performed on the mean EEG RDM in a representative time window (200-600 ms). For dimensionality-reduced visualization we used classical multidimensional scaling as implemented in MATLAB, selecting the dimensions with the largest three eigenvalues in explaining the grand mean RDM.

Supporting experiment.

Methods and additional analyses for the supporting experiment are presented in the **Supplementary Information (SI)**.

Statistics

Behavioral- and modelling results were analyzed using non-parametric tests (two-sided) as detailed in Methods and Results. Time windows of significant effects in neural data were identified using leave-one-out and permutation procedures as explained in Methods (CPP and RSA analysis). Complementary ANOVA results are based on Greenhouse-Geisser corrected degrees of freedom where appropriate.

Code availability

Custom code used in the analysis is made available at
https://github.com/summerfieldlab/Spitzer_etal_2017

Data availability

The data that support the findings of this study are available at https://github.com/summerfieldlab/Spitzer_etal_2017. Raw EEG files are available from the corresponding author upon request.

References

1. Tsetsos, K., Chater, N. & Usher, M. Salience driven value integration explains decision biases and preference reversal. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 9659–9664 (2012).
2. Tsetsos, K. *et al.* Economic irrationality is optimal during noisy decision making. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 3102–3107 (2016).
3. Brezis, N., Bronfman, Z. Z., Jacoby, N., Lavidor, M. & Usher, M. Transcranial Direct Current Stimulation over the Parietal Cortex Improves Approximate Numerical Averaging. *J. Cogn. Neurosci.* **28,** 1700–1713 (2016).
4. Brezis, N., Bronfman, Z. Z. & Usher, M. Adaptive Spontaneous Transitions between Two Mechanisms of Numerical Averaging. *Sci. Rep.* **5,** 10415 (2015).
5. Malmi, R. A. & Samson, D. J. Intuitive averaging of categorized numerical stimuli. *J. Verbal Learn. Verbal Behav.* **22,** 547–559 (1983).
6. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17,** 401–412 (2013).
7. Scott, B. B., Constantinople, C. M., Erlich, J. C., Tank, D. W. & Brody, C. D. Sources of noise during accumulation of evidence in unrestrained and voluntarily head-restrained rats. *eLife* **4,** e11308 (2015).
8. Wyart, V. & Koechlin, E. Choice variability and suboptimality in uncertain environments. *Curr. Opin. Behav. Sci.* **11,** 109–115 (2016).
9. Gibbon, J. Scalar expectancy theory and Weber's law in animal timing. *Psychol. Rev.* **84,** 279–325 (1977).
10. Moyer, R. S. & Landauer, T. K. Time required for judgements of numerical inequality [47]. *Nature* **215,** 1519–1520 (1967).
11. Dehaene, S., Dupoux, E. & Mehler, J. Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *J. Exp. Psychol. Hum. Percept. Perform.* **16,** 626–641 (1990).
12. Van Opstal, F., de Lange, F. P. & Dehaene, S. Rapid parallel semantic processing of numbers without awareness. *Cognition* **120,** 136–147 (2011).
13. Kelly, S. P. & O'Connell, R. G. Internal and external influences on the rate of sensory evidence accumulation in the human brain. *J. Neurosci. Off. J. Soc. Neurosci.* **33,** 19434–19441 (2013).
14. O'Connell, R. G., Dockree, P. M. & Kelly, S. P. A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nat. Neurosci.* **15,** 1729–1735 (2012).
15. Nili, H. *et al.* A toolbox for representational similarity analysis. *PLoS Comput Biol* **10,** e1003553 (2014).
16. Woodford, M. Prospect Theory as Efficient Perceptual Distortion. *Am. Econ. Rev.* **102,** 41–46 (2012).
17. Li, V., Herce Castanon, S., Solomon, J. A., Vandormael, H. & Summerfield, C. Robust averaging protects decisions from noise in neural computations (in revision).
18. Twomey, D. M., Murphy, P. R., Kelly, S. P. & O'Connell, R. G. The classic P300 encodes a build-to-threshold decision variable. *Eur. J. Neurosci.* **42,** 1636–1643 (2015).

19. Donchin, E. & Coles, M. G. H. Is the P300 component a manifestation of context updating? *Behav. Brain Sci.* **11,** 357–374 (1988).
20. Sutton, S., Braren, M., Zubin, J. & John, E. R. Evoked-Potential Correlates of Stimulus Uncertainty. *Science* **150,** 1187–1188 (1965).
21. Picton, T. W. The P300 wave of the human event-related potential. *J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc.* **9,** 456–479 (1992).
22. Bulthé, J., De Smedt, B. & Op de Beeck, H. P. Visual number beats abstract numerical magnitude: format-dependent representation of Arabic digits and dot patterns in human parietal cortex. *J. Cogn. Neurosci.* **27,** 1376–1387 (2015).
23. Eger, E. *et al.* Deciphering Cortical Number Coding from Human Brain Activity Patterns. *Curr. Biol.* **19,** 1608–1615 (2009).
24. Harvey, B. M., Klein, B. P., Petridou, N. & Dumoulin, S. O. Topographic representation of numerosity in the human parietal cortex. *Science* **341,** 1123–1126 (2013).
25. Dehaene, S. The organization of brain activations in number comparison: event-related potentials and the additive-factors method. *J. Cogn. Neurosci.* **8,** 47–68 (1996).
26. Libertus, M. E., Woldorff, M. G. & Brannon, E. M. Electrophysiological evidence for notation independence in numerical processing. *Behav. Brain Funct.* **3,** 1 (2007).
27. Wyart, V., Myers, N. E. & Summerfield, C. Neural mechanisms of human perceptual choice under focused and divided attention. *J. Neurosci. Off. J. Soc. Neurosci.* **35,** 3485–3498 (2015).
28. Ille, N., Berg, P. & Scherg, M. Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies. *J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc.* **19,** 113–124 (2002).
29. Spitzer, B., Blankenburg, F. & Summerfield, C. Rhythmic gain control during supramodal integration of approximate number. *NeuroImage* **129,** 470–479 (2016).
30. Grootswagers, T., Wardle, S. G. & Carlson, T. A. Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *J. Cogn. Neurosci.* 1–21 (2016). doi:10.1162/jocn_a_01068
31. Nili, H., Walther, A., Alink, A. & Kriegeskorte, N. Inferring exemplar discriminability in brain representations. *bioRxiv* 080580 (2016). doi:10.1101/080580
32. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164,** 177–190 (2007).

Acknowledgements

This work was supported by grants from the German Research Foundation to B.S. (DFG SP 1510/1-1; DFG SP 1510/2-1) and by a European Research Council (ERC) Starter Grant (281628) to C.S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Vickie Li, Timo Flesch, and Hamed Nili for helpful suggestions and scripts, Felix Blankenburg for resources, Andreea Epure for help with data acquisition, and Rogier Kievit and two anonymous reviewers for helpful comments on a previous version of the manuscript.

Contributions

B.S. designed the experiments with contributions by L.W. and C.S. L.W. and B.S. conducted the experiments. B.S. and C.S. developed the analysis approach. B.S. analyzed the data with contributions by C.S. B.S. and C.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Corresponding author

Correspondence to Bernhard Spitzer (bernardodispitz@gmail.com), Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK Phone +44(0)1865 271321

Supplementary Information (SI): *Selective overweighting of larger magnitudes during noisy numerical comparison* (Spitzer, Waschke, Summerfield)

Supplementary Methods

Counting model

It is theoretically possible that rather than averaging numerical values (e.g., red, green), participants might have adopted a "selective counting" strategy to achieve above-chance level performance in our task. For instance, participants might simply have counted samples that exceeded a certain threshold (e.g. >3) and compared this tally between categories (e.g. red-green, cf. Fig. 1a). To test whether our participants might have used such strategy, we fitted a "selective counting" model where the psychometric mapping (cf. Methods, eq. 1) is formulated as a step-function with individually estimated threshold (1-6) and offset parameters, for direct comparability with eq. 1. Critically, the selective counting model fitted the human data substantially less well than our non-linear integration model (eq. 1), in both the visual and the auditory conditions (both $p<0.001$, Wilcoxon signed-rank tests on AIC values; see **Supplementary Fig. 1c** for graphical illustration). Furthermore, fitting the model predictions of the selective counting model with our non-linear integration model (eq. 1) yielded compression parameters (k) considerably smaller than those obtained from human data (visual: 0.68 vs. 1.91; auditory: 0.38 vs. 1.95; cf. Results). In other words, the selective counting model yielded a poor fit of the human data and was unable to predict key aspects of our modelling results ($k>>1$), rendering it unlikely that selective counting was a dominant strategy in our experiments.

Supporting experiment

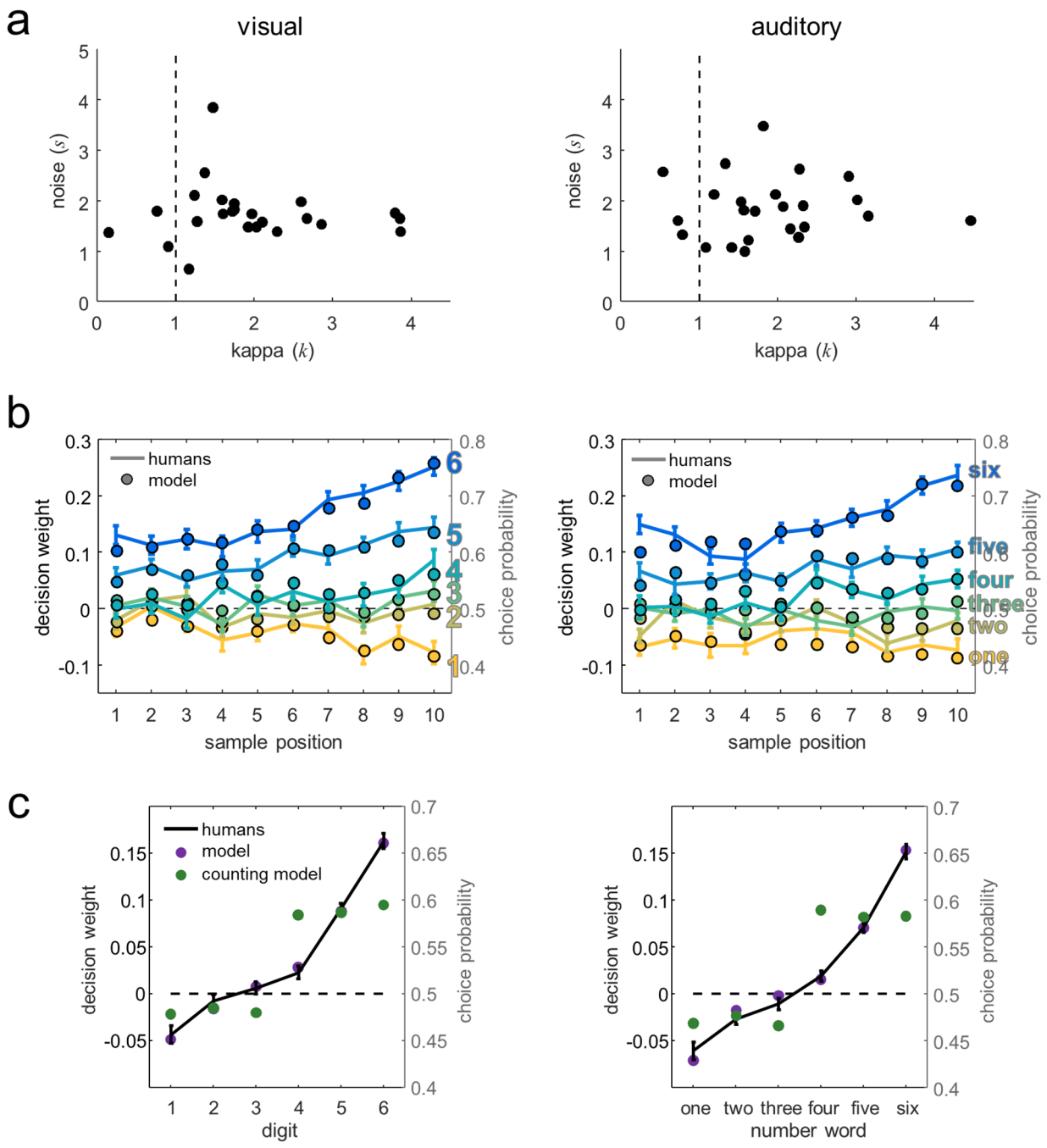
Participants. A new sample of healthy volunteers (N=22, 13 females, 9 males, age 27.9 ± 9.7) participated in the supporting experiment after giving written informed consent. One participant failed to complete all task conditions, leaving N=21 for analysis. The experiment was approved by the Oxford University Medical Sciences Division Research Ethics Committee.

Stimuli, task, and procedure. On each trial, 8 visual samples were presented in sequence at a rate of 400 ms (**Supplementary Fig. 2a**). Each sequence contained 4 symbolic (digits, font Arial, approx. 3° visual angle) and 4 non-symbolic (dots displays) number samples in random serial order, each drawn with uniform probability from numbers 1-9. The locations of dots in non-symbolic samples were varied randomly and independent of number within a circular display area of approx. 7.2° visual angle, with individual dot sizes of approx. 0.25, 0.36, or 0.5° (randomly assigned). A grey circle around the display area was shown during the entire sequence for spatial reference, together with a thin grey fixation cross (not shown in Supplementary Fig. 2a). Each sample was displayed for 200 ms followed by a 200 ms blank period. Half of the samples in each sequence was displayed in red, the other half in green color (randomly assigned across all 8 samples, independent of format). In the "standard" task condition, participants indicated with a key press whether the red or the green samples had the larger average (i.e., integrating across number formats, dots and digits). In another, more difficult task condition ("multi-choice" task), participants made (i) the exact same judgment as above (i.e., red>/<green) but were additionally asked to indicate (ii) whether the dots or the digits had the larger average, and (iii) whether the average of the entire sequence was larger or smaller than five. In this task, after each sequence, participants were sequentially probed with cues (red><green, dots><digits, >five<, in random serial order) to enter their choices (i-iii) one after the other. Each participant performed both tasks (standard, multi-choice) on the same day, but in separate sessions.

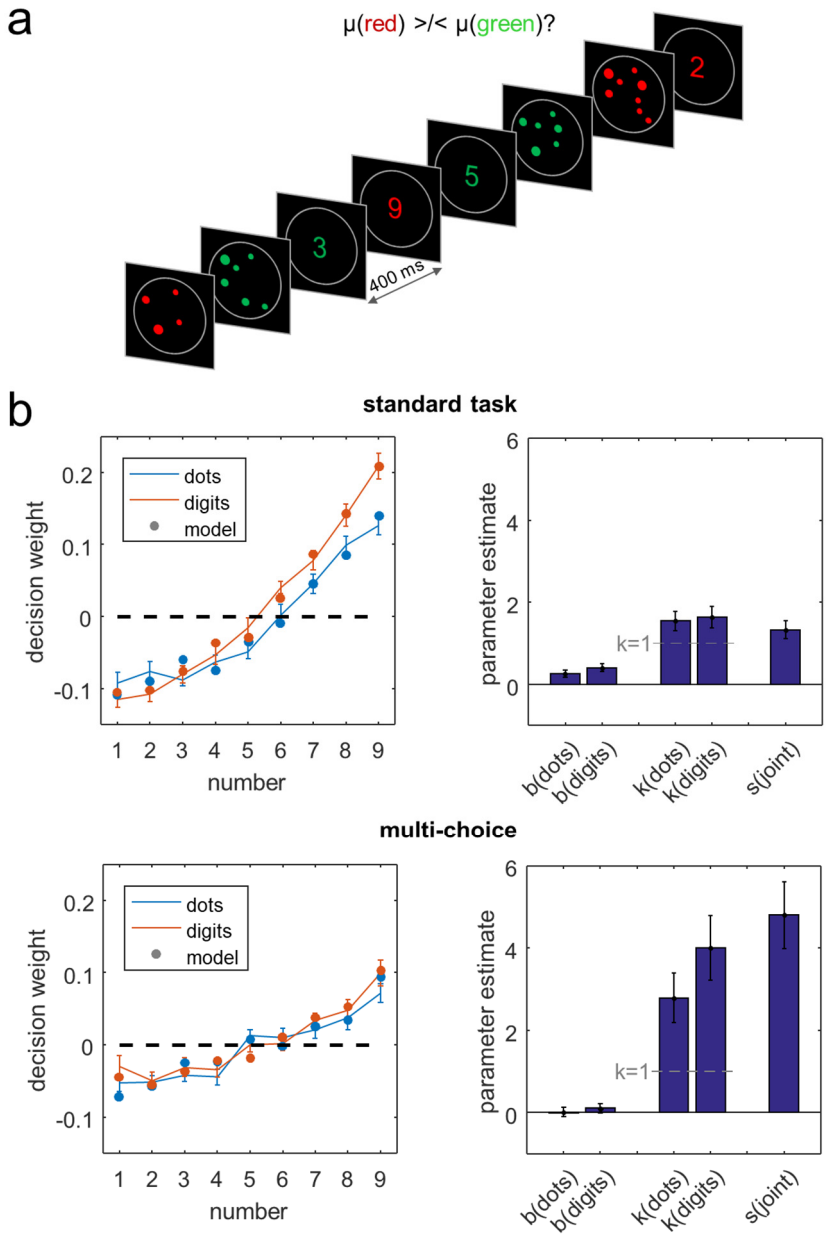
In each session, after several practice runs, 4 blocks of 65 trials were performed, providing 2080 sample presentations per task and participant.

Supplementary results. As expected, discrimination performance (red >< green) was significantly lower in the multi-choice task compared to the standard task (60.4% ± 1.9% vs 73.5% ± 1.6%, Wilcoxon signed-rank test: $p<0.001$). We fitted the non-linear gain model analogous to the main experiment (see Methods, *Psychophysical model*), but using separate parametrizations (b, k) of the mapping function (see Methods, eq. 1) for digits and dots displays, respectively (i.e. fitting 2x9 data points with 6 parameters including a constant term). The best-fitting model parameters are shown in **Supplementary Fig. 2b** right panels (see main text for analysis of s and k parameters). The estimates of b showed significantly positive offset biases in the standard task (mean 0.33 ± 0.09; Wilcoxon signed-rank tests: both $p<0.05$) but not in the multi-choice task (mean 0.06 ± 0.11; both $p>0.30$). However, a 2x2 repeated measures ANOVA failed to show reliable differences in b across tasks and sample formats (all $F_{1,20}<3.9$, all $p>0.05$). As in the main experiment, inclusion of a leak parameter l (see Methods, eq. 3b) further improved the model fit, although the improvement was statistically significant only in the multi-choice task (Wilcoxon signed-rank test on AIC values: $p<0.01$; standard task, $p=0.14$). In direct comparison, l was significantly larger in the multi-choice task than in the standard task (0.38 ± 0.07 vs 0.06 ± 0.01; Wilcoxon signed-rank test: $p<0.001$), corroborating a contribution of memory leakage to overall integration noise (see also main experiment).

Supplementary Figures

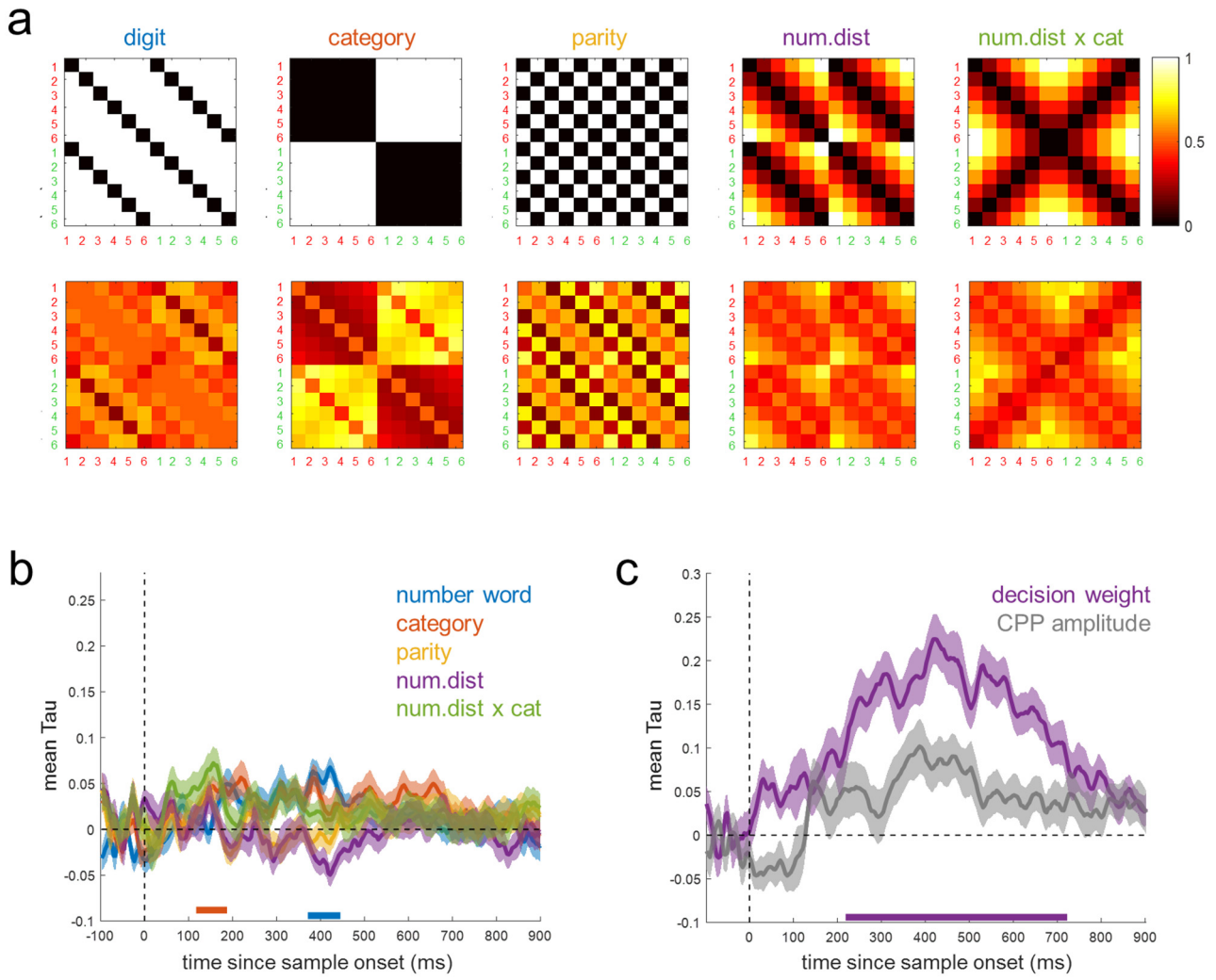


Supplementary Figure 1: Supplementary behavioural results (main experiment, N=24). Left panels: visual condition (digits, cf. Fig. 1a), right panels: auditory condition (number words) **a**, Maximum-Likelihood estimates of k and s in each individual subject. **b**, Predicted (dots) and observed (lines) mean weights as a function of sample position (x-axis) and number (colorscale, yellow-blue). Error bars show SEM. **c,** Counting model. Black: human data; error bars show SEM. Purple: predictions of the best-fitting non-linear integration model (eq. 1). Green: predictions of the best-fitting counting model (see Supplementary Methods). Same axis conventions as Fig. 1d. Left panel: visual; right panel: auditory.



Supplementary Figure 2: Supporting experiment. **a**, Example trial sequence. In each task condition (standard, multi-choice), participants decided whether the red or the green samples had the larger average. In the multi-choice condition, participants were additionally required to simultaneously evaluate other dimensions of the sequence, rendering red>/<green integration more difficult. **b,** Results (N=21), left panels: mean decision weights for numbers 1-9, plotted separately for dots and digits samples. Lines show human data, filled circles show predictions of the best-fitting non-linear model. Right panels: best-fitting parameter estimates. Note that separate parametrizations of the mapping function (bias b, kappa k, see eq. 1) were used for digits and dots samples (cf. a). Error bars show SEM. We note that compared to the main experiment (Fig. 1, Fig. 2a), the modelling analysis of the supporting experiment gave a less accurate description for very small numbers (cf. Fig. 2b). This might be attributable to particularities in encoding numbers < 4 in non-symbolic samples ("subitizing")[1] which were not included in the main experiment.

1. Kaufman, E. L., Lord, M. W., Reese, T. W. & Volkmann, J. The Discrimination of Visual Number. *Am. J. Psychol.* **62,** 498–525 (1949).



Supplementary Figure 3: Supplementary RSA methods and results (main experiment, N=24). **a**, Model RDMs encoding individual sample features before (top row) and after (bottom row) recursive Gram-Schmidt orthogonalisation (see Methods, Representational similarity analysis). **b**, Correlations (Kendall's Tau) between orthogonalized model RDMs and the observed EEG-RSA patterns in the auditory condition. Same conventions as Fig. 4a. **c**, Mean correlation (Kendall's Tau) between the EEG-RSA pattern in the visual condition and orthogonalized model RDMs predicted from the (i) psychometric weight functions of the nonlinear gain model fitted to the behavioural data (purple; cf. Fig. 1d left) or (ii) CPP amplitudes (grey; cf. Fig. 3b left). The model-predicted weights explained a substantial portion of RSA variance that was not explained by CPP amplitude (220-715 ms, $p_{cluster}<0.001$), indicating that multivariate EEG patterns were predicted by participants' "number line" in decision weighting (cf. Fig. 1d) over and above the influence of univariate CPP-modulations in the same data epochs (cf. Fig. 3b).