nature human behaviour

SUPPLEMENTARY INFORMATION

VOLUME: 1 | ARTICLE NUMBER: 0145

In the format provided by the authors and unedited.

Selective overweighting of larger magnitudes during noisy numerical comparison

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

Supplementary Information

[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

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

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.

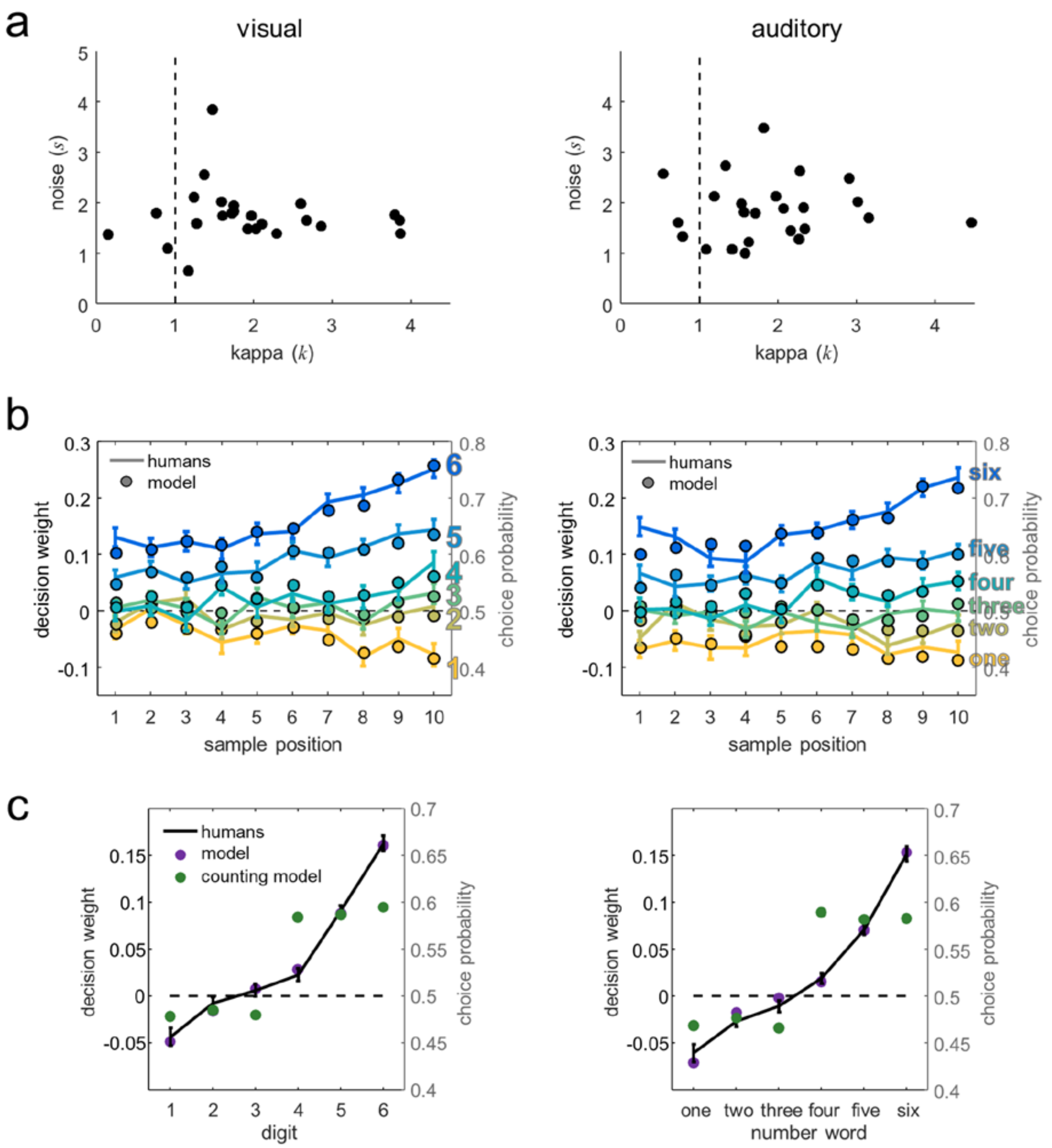
Supporting experiment

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.

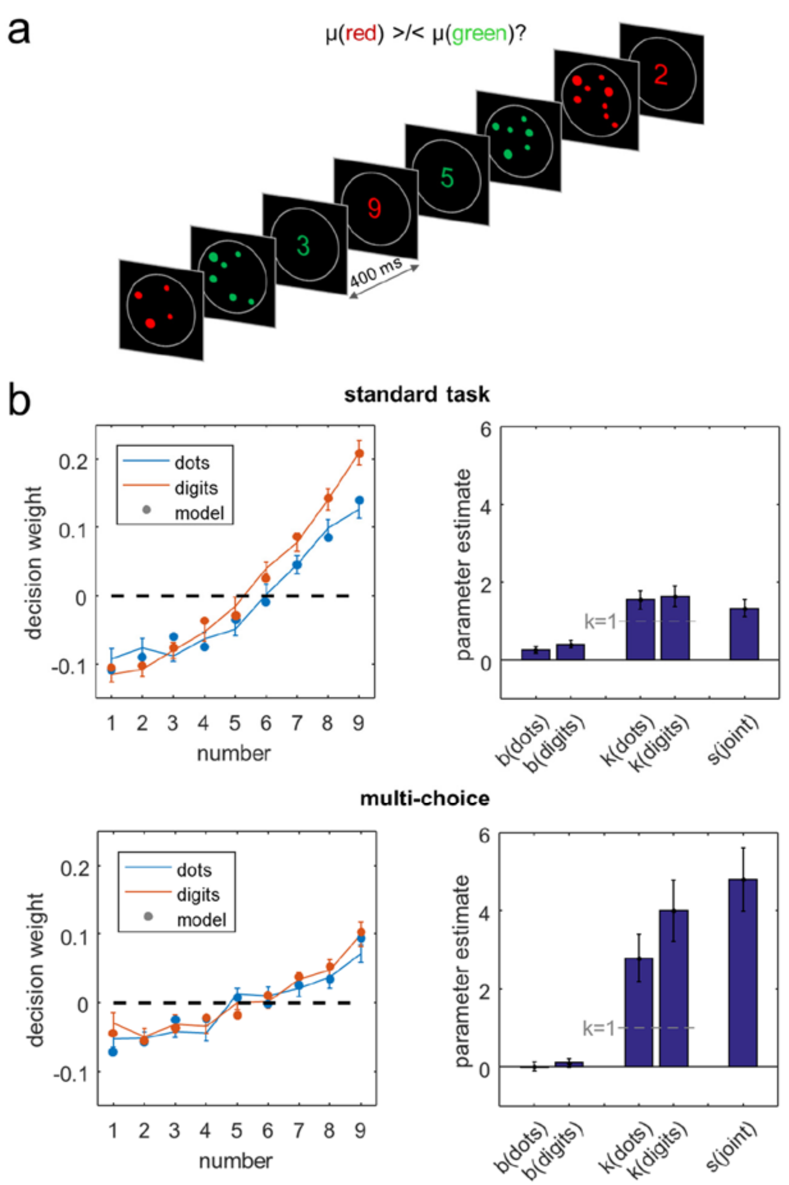
Stimuli, task, and procedure. On each trial, 8 visual samples were presented in sequence at a rate of 400 ms (**Supplementary Fig. 2a**). Each sequence contained 4 symbolic (digits, font Arial, approx. 3° visual angle) and 4 non-symbolic (dots displays) number samples in random serial order, each drawn 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

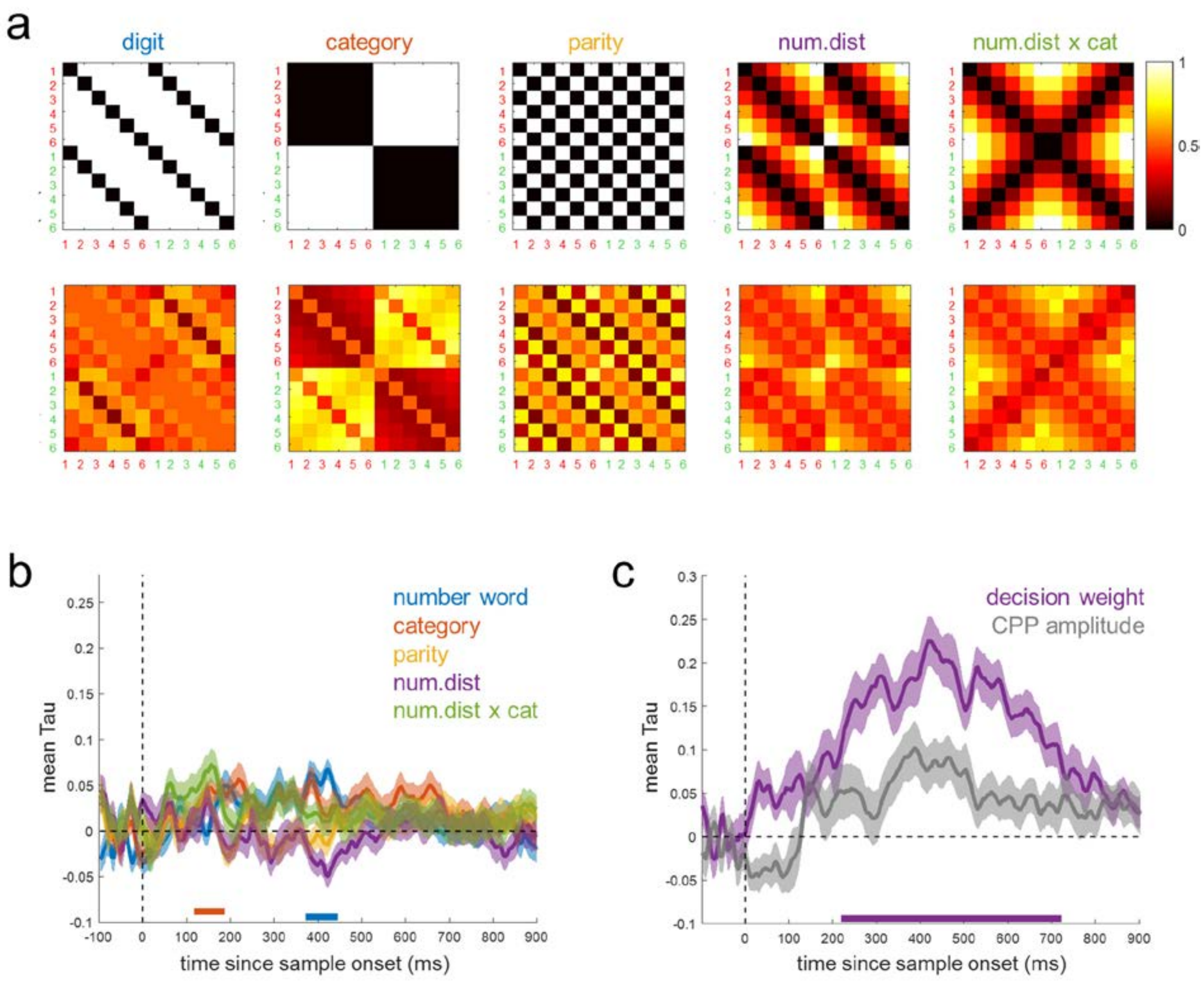


Supplementary Figure 1: Supplementary behavioural results (main experiment, N=24). Left panels: visual condition (digits, cf. Fig. 1a), right panels: auditory condition (number words) **a**, Maximum-Likelihood estimates of k and s in each individual subject. **b**, Predicted (dots) and observed (lines) 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).