



This postprint was originally published by the APA as:
Clarmann von Clarenau, V., Appelhoff, S., Pachur, T., & Spitzer, B.
(2024). **Over- and underweighting of extreme values in decisions
from sequential samples.** *Journal of Experimental Psychology:
General*, 153(3), 814–826. <https://doi.org/10.1037/xge0001530>

The following copyright notice is a publisher requirement:

©American Psychological Association, 2024. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission.

The final article is available, upon publication, at:

<https://doi.org/10.1037/xge0001530>

Provided by:

Max Planck Institute for Human Development

Library and Research Information

library@mpib-berlin.mpg.de

**Over- and Underweighting of Extreme Values in
Decisions from Sequential Samples**

Verena Clarmann von Clarenau^{1,2,3,4}, Stefan Appelhoff^{1,4}, Thorsten Pachur^{*,1,2,5,6}, and
Bernhard Spitzer^{*,1,2,4}

¹Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin,
Germany

²Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin,
Germany

³Department of Psychology, Humboldt University, Berlin, Germany

⁴Research Group Adaptive Memory and Decision Making, Max Planck Institute for Human
Development, Berlin, Germany


⁵School of Management, Technical University of Munich


⁶Science of Intelligence, Research Cluster of Excellence, Germany


* Shared senior authorship

Author Note

Verena Clarmann von Clarenau  <https://orcid.org/0000-0002-9055-4220>

Stefan Appelhoff  <https://orcid.org/0000-0001-8002-0877>

Thorsten Pachur  <https://orcid.org/0000-0001-6391-4107>

Bernhard Spitzer  <https://orcid.org/0000-0001-9752-932X>

This research was supported by the Adaptive Center for Rationality at the Max Planck Institute for Human Development, the International Max Planck Research School on Computational Methods in Psychiatry and Ageing Research (IMPRS COMP2PSYCH), the Max Planck Dahlem Campus of Cognition (MPDCC), European Research Council Consolidator Grant ERC-2020-COG-101000972 (BS), DFG (German Research Foundation) grants 444526808, 462752742 (BS) and PA1925/2-1 (TP), as well as under Germany's Excellence Strategy – EXC 2002/1 “Science of Intelligence” – project number 390523135. Part of this research was presented at the Tagung experimentell arbeitender Psychologen 2021 (TeaP; Virtual) and the 62nd Annual Meeting of the Psychonomic Society 2021 (Virtual). An early version of this manuscript was posted as a preprint on the PsyArXiv repository (<https://doi.org/10.31234/osf.io/6yj4r>). The data, code, and experimental materials for this research are available in the Open Science Framework repository at <https://osf.io/x83pk/>. This work was not preregistered. We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Verena Clarmann von Clarenau, Center for Adaptive Rationality, Max Planck Institute of Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: clarmann@mpib-berlin.mpg.de or Bernhard Spitzer, Research Group Adaptive Memory and Decision Making, Max Planck Institute of Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: spitzer@mpib-berlin.mpg.de

Author contributions:

OVER- AND UNDERWEIGHTING OF EXTREME VALUES

VCvC: Conceptualization, Investigation, Data curation, Formal analysis, Project administration, Software, Visualization, Writing – original draft

SA: Data curation, Formal analysis, Software, Validation, Writing – review & editing

TP: Conceptualization, Methodology, Project administration, Resources, Supervision, Writing – review & editing

BS: Conceptualization, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing

OVER- AND UNDERWEIGHTING OF EXTREME VALUES

Abstract

People routinely make decisions based on samples of numerical values. A common conclusion from the literature in psychophysics and behavioral economics is that observers subjectively compress magnitudes, such that extreme values have less sway over people's decisions than prescribed by a normative model (underweighting). However, recent studies have reported evidence for anti-compression, that is, the relative overweighting of extreme values. Here, we investigate potential reasons for this discrepancy in findings and propose that it might reflect adaptive responses to different task requirements. We performed a large-scale study ($n = 586$) of sequential numerical integration, manipulating (i) the task requirement (averaging a single stream or comparing two interleaved streams of numbers), (ii) the distribution of sample values (uniform or Gaussian), and (iii) their range (1 to 9 or 100 to 900). The data showed compression of subjective values in the averaging task, but anti-compression in the comparison task. This pattern held for both distribution types and for both ranges. In model simulations, we show that either compression or anti-compression can be beneficial for noisy observers, depending on the sample-level processing demands imposed by the task. This suggests that the empirically observed patterns of over- and underweighting might reflect adaptive responses.

Public Significance Statement

In decisions based on numbers, people tend to either over- or underweight extreme values. This study provides a new framework to explain why sometimes overweighting and sometimes underweighting is observed. In simulations, we show that either of the two types of distortion can be performance-maximizing for noisy observers, depending on the processing demands of the task. This framework is empirically supported by a large-scale study showing that the type of distortion (over- or underweighting) displayed by participants varied with task demands, but not with other experimental factors. The results address long-standing questions as to why humans make seemingly irrational

OVER- AND UNDERWEIGHTING OF EXTREME VALUES

decisions, and reconcile discrepant findings in the previous literature.

Keywords: decision making, numerical cognition, computational modeling, adaptive cognition

Over- and Underweighting of Extreme Values in

Decisions from Sequential Samples

Many decisions are based on sampling numerical values—for example, when glancing through the prices in the menu of a new restaurant that has opened up in the neighborhood to judge whether it is affordable. A common observation in experimental studies is that decision makers tend to distort sample values away from their objective magnitude, for instance, giving relatively greater weight to mid-range values, and relatively less weight to more extreme values. Such subjective *compression* (underweighting of extreme values) is observed not only in psychophysical judgments of perceptual information (as described by the Weber–Fechner law; Fechner, 1860), but also in economic decisions involving numerical or monetary values (Bernoulli, 1954; Kellen et al., 2016; Tversky & Kahneman, 1992). On a neurocognitive level, compression is assumed to optimize information transfer in capacity-limited channels (efficient coding; e.g., Barlow, 1961; Bhui & Gershman, 2018) and to enable robust ensemble judgments (de Gardelle & Summerfield, 2011; Vandormael et al., 2017) that may protect against the deleterious effects of “late” decision noise (see below; e.g., Juechems et al., 2021; Li et al., 2017).

However, several recent studies of sample-based decision making have observed the opposite type of distortion, namely *anti-compression*. With anti-compression, extreme values are given relatively greater weight than implied by their objective magnitude (overweighting of extreme values; Kunar et al., 2017; E. A. Ludvig et al., 2018; E. A. Ludvig et al., 2014; Prat-Carrabin & Woodford, 2020; Shevlin et al., 2022; Spitzer et al., 2017; Tsetsos et al., 2012; Tsetsos et al., 2016; Vanunu et al., 2019). The reasons for this discrepancy in findings are currently unclear. Complicating matters, both types of distortion have been associated with performance benefits in the face of “late” decision noise (i.e., noise that occurs downstream of sensory-perceptual encoding—for instance, when evidence from multiple samples is combined into a binary choice; Li et al., 2017; Spitzer et al., 2017; Tsetsos et al., 2016). Based on model simulations, one proposal has been that compression maximizes the perfor-

mance of noisy observers when the sample values are normally (i.e., Gaussian) distributed, whereas anti-compression may be beneficial when the distribution is uniform (e.g., Li et al., 2017; Spitzer et al., 2017; Summerfield & Li, 2018).

Alternatively, whether participants show compression or anti-compression may also be influenced by other task aspects (see also Summerfield & Parpart, 2021). Many of the experimental tasks in which anti-compression has been observed posed relatively high cognitive demands, for instance, requiring evaluation of relational information in rapid sequential displays (e.g., Tsetsos et al., 2012; Tsetsos et al., 2016), and the degree of anti-compression has been found to increase with task complexity (Spitzer et al., 2017; Vanunu et al., 2019). Compression, in contrast, has often been observed in more direct perceptual (and arguably simpler) judgments of ensemble information, such as the average magnitude or orientation of a stimulus set (de Gardelle & Summerfield, 2011; Katzin et al., 2021; Li et al., 2017; Vandormael et al., 2017). The type of distortion observed (i.e., compression or anti-compression) may thus hinge on the processing resources required to evaluate the sample information in the context of the specific decision task at hand.

Here, we show in simulations that compression confers performance benefits under noise (see also Juechems et al., 2021; Li et al., 2017) only when the individual samples in a set can be evaluated with relatively little processing effort. In contrast, when their evaluation is more demanding—such that limited processing resources need to be traded off between the different samples in a trial—we find the optimal policy to be anti-compression. In other words, whether compression or anti-compression is the favorable policy for noisy observers may depend on the sample-wise processing requirements imposed by the task. We tested this proposal in a large participant sample ($N = 586$), manipulating the task requirement (simple averaging vs. comparison of number sequences; see *Methods*). We additionally manipulated the distribution of sample values (uniform vs. Gaussian, see above; Li et al., 2017; Spitzer et al., 2017), and also their range (1 to 9 vs. 100 to 900), as there

is some evidence that nonlinear distortions might be more pronounced with higher than with lower numbers (Birnbaum & Chavez, 1997; Wakker & Deneffe, 1996). Our results show that whether participants' decisions reflected compression or anti-compression was determined solely by the processing requirement of the task, and that they adopted the favorable weighting policy in the respective task context. Thus, we present a theoretical framework supported by empirical data to explain a previously puzzling heterogeneity in the literature, namely, that decision makers sometimes overweight and sometimes underweight extreme values in sample-based decisions.

Methods

Transparency and Openness

All data and analysis code as well as experimental materials are available at <https://osf.io/x83pk> (Clarmann von Clarenau et al., 2022). The data were analyzed using Matlab version R2020b (The MathWorks, 2020a), with the Statistics and Machine Learning Toolbox (The MathWorks, 2020e), the Econometrics Toolbox (The MathWorks, 2020b), the Parallel Computing Toolbox (The MathWorks, 2020c), the Optimization Toolbox (The MathWorks, 2020d), and the BayesFactor Toolbox (Krekelberg, 2022). The study design and analyses were not preregistered. The study was approved by the ethics committee of the Max Planck Institute for Human Development.

Participants

We aimed at recruiting 800 young adults ($n = 100$ per condition) via Prolific (<https://www.prolific.co>). The target sample size was based on model simulations and previous lab studies of psychometric distortions (see Introduction) and was increased to accommodate anticipated drop-outs in online testing. Eventually, $N = 778$ participants (222 female, 442 male, 114 data on sex unavailable; mean age 24.83 ± 5.02 years, range 18–41 years) took part in the experiment. All participants gave written informed consent prior to performing the experiment and received a basic reimbursement of 5.40 GBP per hour and a performance-dependent bonus of up to 0.9 GBP. Participants were excluded if they failed on attention

checks (see below) or if their performance in the main task was not significantly above chance ($p < 0.01$, binomial test against 0.5, corresponding to at least 56% correct responses). We further excluded 13 subjects who had participated repeatedly (in different conditions). After exclusion, $n = 586$ participants (185 female, 390 male, 11 data on sex unavailable; $M_{age} = 24.74 \pm 4.85$ years) were retained in the analysis (mean $n = 73.4$ per experimental condition; min: 68, max: 78).

Stimuli, Task, and Procedure

The experiment was conducted online using Qualtrics (<https://www.qualtrics.com>). In each of the 8 experimental conditions, participants were asked to judge sequences of 8 number samples displayed in red or blue (Figure 1a). The beginning of each trial was indicated by a fixation cross displayed in the middle of the screen for 1 second. Subsequently, 8 numbers (Arabic digits; 4 in red and 4 in blue font, presented in random order) were sequentially displayed at fixation at a rate of 350ms per sample. Each sampled number was softly faded out after 270ms to smoothen stimulus transitions. The samples were drawn randomly (see below) and independently from the range of either 1 to 9 (in steps of 1; *small-range conditions*) or 100 to 900 (in steps of 100; *large-range conditions*).

After the last sample had been presented, participants were asked to enter a binary decision via key press. In the single-stream “averaging” task, they were asked to indicate whether the average of all samples (regardless of color) was larger or smaller than a reference value (5 in the small-range conditions; 500 in the large-range conditions). In the dual-stream “comparison” task, they were asked to indicate whether the red samples had a higher average value than the blue samples or vice versa. Thus, both tasks required participants to evaluate all 8 sample values. The two tasks differ in the processing requirements associated with each sample: in the single-stream task, all numbers are evaluated within the same frame of reference (i.e., a larger number always evidences “larger”). In the dual-stream task, in contrast, the decision value of a sample flips depending on its color—that is, a red number evidences “red” when it is large, but “blue” when it is small, and vice versa for blue numbers

(see *Computational Modeling*, Eq. 2). Arguably, this renders the evaluation of the individual samples inherently more effortful than in the single-stream task. In both tasks, a response had to be given within 3 seconds; otherwise a time-out message was displayed and the trial was discarded from analysis (this applied to 0.43% of trials, on average).

Note that the differences between sample values (i.e., red vs. blue) that are to be evaluated in the dual-stream task can be numerically larger (up to twice as large) than the differences of the sample values from the reference value (e.g., “5”) in the single-stream task (see Figure 1a). The dual-stream task might thus be considered to be easier than the single-stream task. However, in sequential integration, the evidence in favor of one or the other choice (e.g., “>” or “<”) provided by the individual samples in a trial can be assumed to be identical in both tasks (see *Psychometric Model*). In fact, our empirical results confirmed that the dual-stream task was *harder* (not easier) than the single-stream task (see *Human Participant Results*).

We also manipulated the distribution from which the sample values were drawn (Figure 1b). In *uniform conditions* (Figure 1b, upper panel), the sample values were drawn from a uniform distribution. In *Gaussian conditions* (Figure 1b, lower panel), the values were drawn from truncated normal distributions with a standard deviation of $\sigma = 3$ (small-range conditions) or $\sigma = 300$ (large-range conditions). To compensate for anticipated differences in task difficulty, we moderately shifted the Gaussian mean away from the midpoint of the sample range (by 0.6 in the small-range and by 60 in the large-range conditions); we derived the magnitude of the shift from pilot data and model simulations, which examined which shift would be necessary to approximate similar performance levels as with uniform samples. In the single-stream task, positive/negative shifts were randomly varied across trials. In the dual-stream task (where red and blue samples had opposite decision values, Figure 1b), the shifts were applied with opposite signs (positive/negative) to the distributions from which the red and blue samples in a trial were drawn (with random assignment across trials). Trials

on which no objectively correct response could be given because the mean of the number stream was exactly 5 (single-stream condition; 5.67% of trials) or identical for red and blue samples (dual-stream condition, 5.32% of trials) were excluded from analysis.

The three experimental factors (i.e., “task”, “range”, and “distribution”) were fully crossed in a $2 \times 2 \times 2$ between-subjects design. In each condition, participants performed 250 trials in blocks of 50. After each block, summary performance feedback was provided (percentage of correct responses). After every 25th trial, participants were asked to perform a brief attention check task consisting of 4 trials. Here, they were shown a geometric shape (square or circle) and asked to indicate its name via key press. If a participant passed fewer than 3 (75%) of the trials on any given attention check, the experimental session was terminated and their data were discarded (see *Participants*). After the main task, participants in the Gaussian conditions additionally performed the Berlin Numeracy Test (Cokely et al., 2012) and a brief number line estimation task (“number-to-position”; Siegler & Opfer, 2003), the results of which are not reported here. Participants who successfully completed the experiment were paid a performance-dependent bonus (up to 0.9 GBP) on top of their basic reimbursement.

Descriptive Analysis

We used a reverse correlation approach (Neri et al., 1999; Spitzer et al., 2016) to calculate decision weights for each sample value (1-9). Specifically, in the single-stream task conditions, we calculated for each sample value (e.g., 3) the proportion of times a sample of this value was followed by a “larger” decision. In the dual-stream task conditions, decision weights for each sample value were computed analogously, as the proportion of times the color of the sample (i.e., red or blue) was subsequently chosen to be larger. In either of the two tasks, the choice proportions express how much influence a sample value had on a participant’s decision; this yields a psychometric weighting function over sample values, where a choice proportion of 0.5 (indifference) corresponds to zero decision weight (see Figure 3, left and right y-axis labels). For comparison with model predictions (see below), we

computed analogous weighting functions also from the choice probabilities (Eq. 5) predicted under the best-fitting model parameterization for each participant. For additional analysis of the weighting functions' local slopes (see *Results, Complementary Analyses*), we calculated the difference in decision weight between neighbouring sample values (e.g., 1 vs. 2) and compared it between outlying (1 vs. 2 and 8 vs. 9) and inlying value pairs (4 vs. 5 and 5 vs. 6).

Psychometric Model

We modeled decisions in our tasks using a simple psychometric model as reported in Spitzer et al. (2017). In the model, each sample value X (normalized to range from -1 to 1) is transformed into a subjective decision value dv according to a sign-preserving power function of the form:

$$dv = \frac{X + b}{|X + b|} \times |X + b|^k, \quad (1)$$

where $k < 1$ implies underweighting (i.e., compression) and $k > 1$ implies overweighting (i.e., anti-compression) of extreme values, relative to a linear transformation (i.e., $k = 1$; Figure 2a). Non-zero values of parameter b indicate an offset bias in terms of a shift of the function's indifference point. The dvs of the $N = 8$ individual samples in a trial are accumulated to yield a trial-level decision value DV :

$$DV = \sum_{i=1}^N \frac{dv_i \times c_i}{g}, \quad (2)$$

where c_i is an indicator variable ($+1$ or -1) that codes the color category of the sample (i.e., red or blue). In the single-stream conditions (where color is irrelevant), c_i is fixed at $+1$. In the dual-stream conditions, c_i effects a sign-flip of dvs for one color relative to the other, effectively implementing a comparison between the two color categories.

g is a scaling factor that normalizes the gain of the transformation in Eq. 2 (quantified

by the integral of the decision values; Figure 2a–c) to be constant for any values of k and b :

$$g = \frac{\sum |f + b|^k}{\sum |f|}. \quad (3)$$

We considered two variants of this normalization. In one variant (Figure 2b), we defined f as the 9 possible input values of X (see Eq. 1), which normalizes the gain over the range of stimuli that could potentially occur in the experiment (e.g., 1 to 9 in the small-range condition). In a second, more refined variant, we computed g on the individual trial level, where f refers to the concrete sequence of input values X presented on a given trial (see Figure 2c). This second type of normalization ensures that equivalent gain of processing is available for each individual number sequence, for every parameterization of Eq. 2 (see *Simulation Results* for details).

To capture potential recency effects (i.e., greater weighting of samples occurring closer to the decision; Hogarth & Einhorn, 1992), we also included a leakage parameter l that modulates the weight of a sample as a function of its temporal position $i = 1 \dots N$ (with $N = 8$ samples) in the number stream:

$$DV = \sum_{i=1}^N \frac{dv_i \times c_i}{g} \times (1 - l)^{N-i}, \quad (4)$$

where larger values of l indicate a stronger recency effect. The trial-level DV was then transformed into a choice probability CP using a logistic choice rule,

$$CP = \frac{1}{1 + e^{\frac{-DV}{s}}}, \quad (5)$$

where CP is the probability of responding “larger” (single-stream condition) or “red > blue” (dual-stream condition) and s reflects decision noise (with higher values indicating more random responses).

Parameter Estimation

The psychometric model was fitted to the experimental data of each participant individually by minimizing the negative log-likelihood of the model given the observed responses. Parameter values were restricted to ranges derived from previous work (k : 0–5, s : 0–5, b : –1–1, l : –0.5–1, see also Spitzer et al., 2017). Model performance was assessed using mean BICs on the group level and examined statistically using conventional inferential statistics (two-tailed). For our main analyses of the empirical data (Figures 3-5), we used the model without gain normalization (i.e., g was set to 1). Note that g is not a free model parameter but acts as a scaling factor on noise parameter s , which does not systematically affect the other model parameters or the model’s goodness of fit (Spitzer et al., 2017, see also Parameter Recovery below). When comparing the empirical estimates of k and s with those values that optimized performance in our simulations (Figures 2e–f), we fitted the model with the same gain-normalization settings as were used in the respective simulations to warrant comparability.

Parameter Recovery

To ensure that our estimated model parameters were valid, we performed parameter recovery simulations. Specifically, we iteratively simulated group data sets analogous to those obtained in our experiment. Across iterations, we varied the value of each parameter (in steps of 0.2) within its range (see above) while fixing the remaining parameters at their empirical estimates. Binary responses were generated by drawing for each trial from a binomial distribution with $p = CP$. We then fitted our model to the simulated data, using the same procedure as in the modeling of the empirical data. The recovered mean parameter values mostly correlated strongly with their respective values in the simulation (single-stream conditions: mean $r = 0.84$, min: $r = 0.71$, max: $r = 0.92$; dual-stream conditions: mean $r = 0.60$, min: $r = 0.45$, max: $r = 0.77$), while cross-correlations between different parameters were generally lower (single-stream condition: mean $r = 0.03$, min: $r = -0.04$, max: $r = 0.19$; dual-stream condition: mean: $r = 0.05$, min: $r = -0.05$,

max: $r = 0.17$). The parameter of our main interest (k) was recovered well both in the single-stream (mean $r = 0.85$) and in the dual-stream task conditions (mean $r = 0.54$), without major distortions by other model parameters (single-stream task: min $r = -0.03$, max $r = 0.09$; dual-stream task: min $r = -0.01$, max $r = 0.17$).

Performance Simulations

We used the psychometric model to simulate task performance under different parameter settings. Accuracy was inferred from the predicted choice probabilities CP (if the correct response was “larger”; $1 - CP$ otherwise). We simulated performance across different values of k (0 to 2.5 in steps of 0.01) and s (0 to 2 in steps of 0.01). For each value of k , we examined the difference in accuracy relative to linear (i.e., undistorted) transformation (i.e., $k = 1$) at any given noise level s (see Figure 2d-f). In our *a priori* model simulations shown in Figures 2d-f, we set $l = 0$ (i.e., no leakage) and $b = 0$ (i.e., no offset bias). However, qualitatively very similar results were obtained when using the empirical estimates of l and b derived from our participants, both in the single-stream and in the dual-stream tasks. Note that in our model, the single- and dual-stream tasks are formally equivalent (see Eq. 2), except for a difference in how bias (b) affects the response (see also Results). Thus, the simulation results illustrated in Figures 2d-f hold *a priori* for the single-stream and dual-stream tasks alike. While we present the simulation results for sequences of 8 samples (using the same sequences that had been used in the experiments with human participants), qualitatively identical results were also obtained in exploratory simulations with shorter (e.g., 4 samples) or longer sequences (e.g., 10 samples).

Results

Simulation Results

We examined choice behavior in variants of a sample-based decision task (Figure 1) where observers judge a sequence of 8 numbers. We used a generic psychometric model that formalizes single-stream (i.e., mean $>/< 5$) and dual-stream judgments (i.e., mean[red] $>/<$ mean[blue]) in the same way (see *Methods*). The simulation results reported

in the following thus hold identically for both types of task.

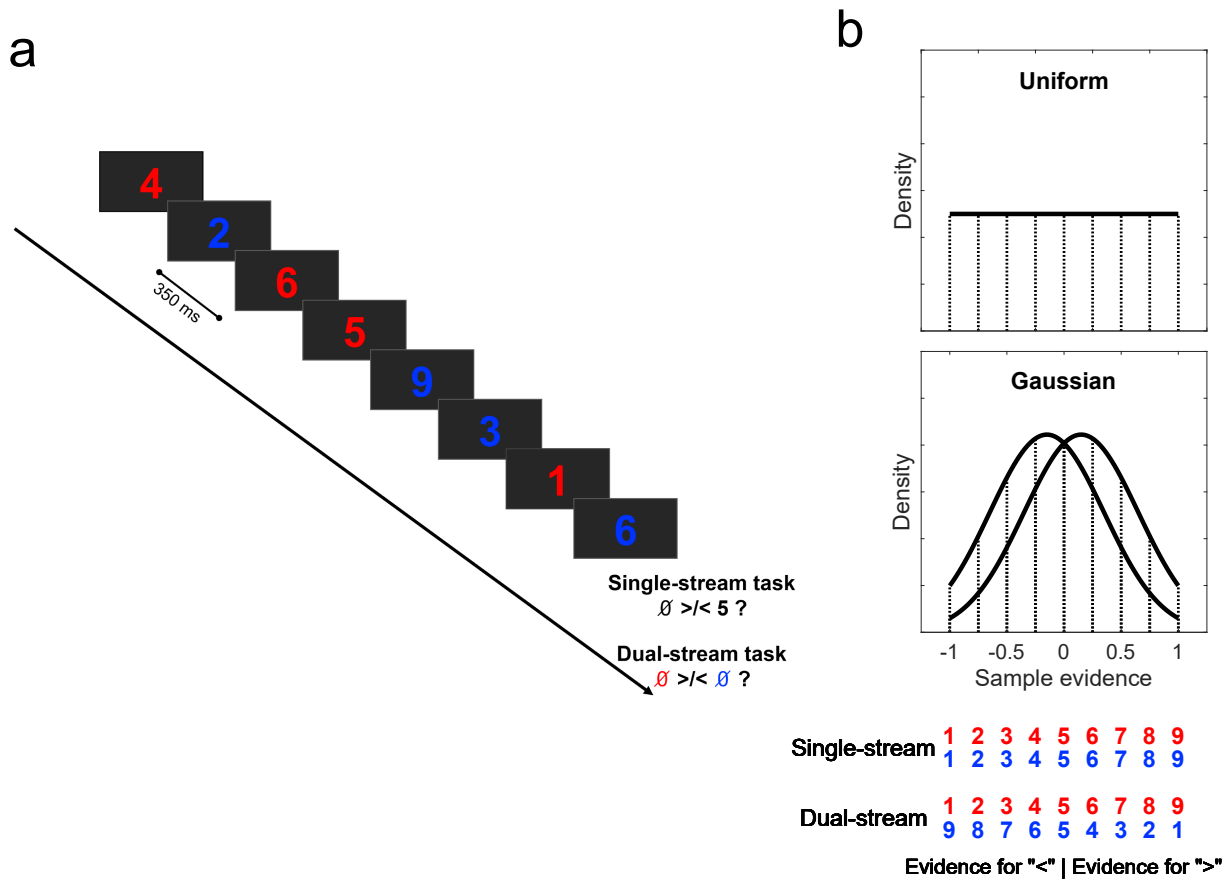


Figure 1. Task and Key Experimental Manipulations. *a*, Example stimulus sequence used in both tasks (single-stream averaging and dual-stream comparison). On each trial, eight number samples (drawn from 1 to 9 in the small-range conditions; from 100 to 900 in the large-range conditions) appeared in either red or blue font at a rate of 350ms/sample. In the single-stream task, participants were asked to indicate whether the average of all samples (regardless of color) was larger or smaller than 5 (in the small-range conditions) or than 500 (in the large-range conditions). In the dual-stream task, participants were asked to indicate whether the red samples had a larger or a smaller average value than the blue samples. *b*, Distribution of sample values in the uniform (top) and Gaussian (bottom) conditions. Digits on bottom illustrate the mapping (x-axis) onto red (top row) and blue (bottom row) sample stimuli (cf. *a*) in the respective task conditions. The panel for the Gaussian condition shows two distributions because we varied trial-by-trial within each task whether the mean of the distribution was in favor of “smaller” or “larger” (see Methods).

We simulated the effects of compressive or anti-compressive distortions of sample values (here 1 to 9) on task performance, taking into account different types of processing limitations. We started by replicating previous findings showing that, without further assumptions, compressive weighting policies ($k < 1$) are performance-maximizing in the face of decision noise ($s \gg 0$; Figure 2a and d; see also e.g., Juechems et al., 2021; Li et al., 2017). However, compressive transformations are also characterized by overall larger decision values (in terms of absolute decision values $\sum |dv|$) than linear or anti-compressive transformations (Figure 2a, inset bar graph). In other words, with compressive weighting, the sample values are transformed with a greater “gain” of processing (Li et al., 2017), which can be assumed to be costly (e.g., in terms of metabolic resources) for biological observers (Baddeley et al., 2000; Kostal et al., 2013). The beneficial effect of compression on performance in Figure 2d can thus be explained by the recruitment of greater processing resources, resulting in an overall steeper weighting curve (Figure 2a) which counteracts the effects of decision noise (Li et al., 2017).

It is commonly assumed that neural gain is a finite resource (Cowan, 2001; Lennie, 2003; Marois & Ivanoff, 2005; Tombu et al., 2011). Thus, in some task contexts (e.g., when sample processing is computationally demanding), giving a higher decision weight to one sample may come at the cost of other samples. For instance, selectively focusing on one type of stimulus may lead to reduced processing of other stimuli (Alonso et al., 2011; Eldar et al., 2013). In previous work, such processing limit was modeled by normalizing the gain of the transformations (in terms of the integral $\sum |dv|$ over the range of input values; here, 1 to 9) for every value of k (Figure 2b, see Eq. 3; see Li et al., 2017; Spitzer et al., 2017). When repeating the simulation with this normalization, compression ($k < 1$) maximized performance only when the samples were drawn from a Gaussian distribution (Figure 2e, lower panel; see also Li et al., 2017). When the samples were drawn from a uniform distribution, in contrast, performance was maximized under anti-compression ($k > 1$; Figure 2e, upper panel; see also Spitzer et al., 2017).

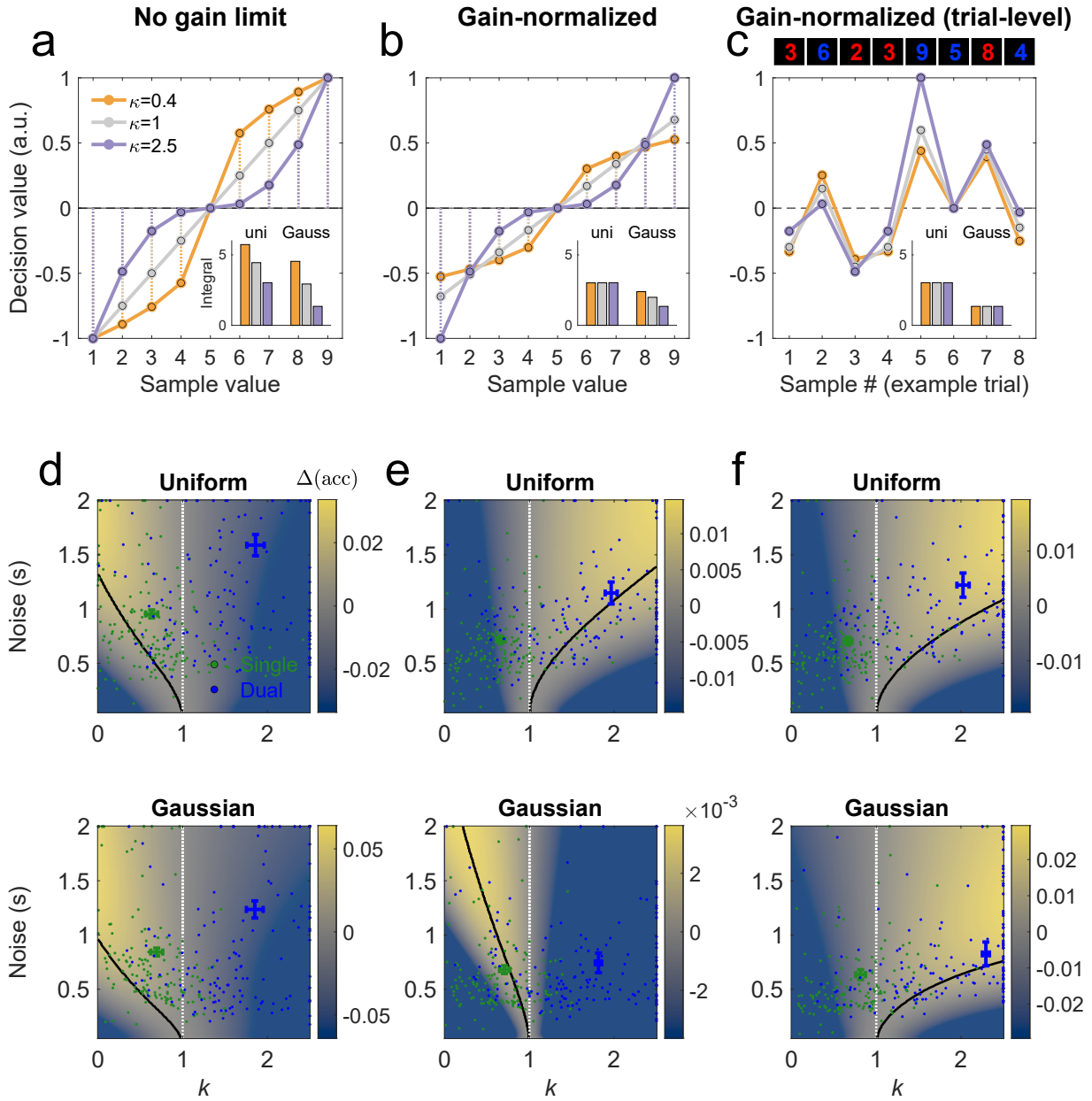


Figure 2. Simulated task performance under processing limitations. *a*, Functions mapping numerical sample values (1 to 9) onto decision values dv , for exemplary values of exponent k (see Equation 1). Inset bar graph shows the mean integral ($\sum |dv|$ across the samples in a trial) for trials with uniformly (left) and Gaussian (right) distributed samples (see Figure 1b). *b*, same as *a*, but normalized to have equivalent gain ($\sum |dv|$ across the range of sample values 1 to 9) for each value of k . *c*, Decision values normalized for equivalent gain on any given trial ($\sum |dv|$ across the samples occurring in a trial) for each value of k .

*Illustrated is an example trial with 8 samples in the single-stream task. **d–f**, Performance simulations. Colormaps show the difference in choice accuracy $\Delta(\text{acc})$ relative to linear weighting ($k = 1$) across values of k and s , for the different types of gain normalization illustrated in **a–c**. Solid black lines indicate the maximal performance improvement under each noise level. Dotted white lines indicate $k = 1$. Dots show empirical parameter estimates in the single-stream and dual-stream tasks. Large dots: mean estimates, error bars show SEM. Small dots: individual participants. Note that fitting with gain normalization in **e** and **f** yields numerically smaller estimates of decision noise s , but equivalent patterns of distortions k . Parameter estimates outside the axis limits are plotted at the plot boundaries. Upper panels: uniform conditions (see Figure 1b, upper panel); lower panels: Gaussian conditions (see Figure 1b, lower panel).*

Critically, for samples from a uniform distribution, the normalization shown in Figure 2b is equivalent to limiting the processing resources that are available to an observer (on average) on any given trial (see bar graph in Figure 2b). For samples from a Gaussian distribution, in contrast, compressive distortions ($k < 1$) will still require greater resources overall, since mid-range samples (e.g., 4 or 6), which are more resource-intensive under $k < 1$ (see Figure 2b), will occur relatively more frequently. To resolve this imbalance, we implemented gain normalization on the single-trial level (Figure 2c), such that each trial sequence was transformed with equivalent gain of processing for any value of k (see bar graph in Figure 2c). The trial-level normalization implements a hypothetical scenario where an observer would process each trial with a fixed amount of resources, such that different distortions k would distribute these resources differently across the individual samples in a trial. This type of normalization thus implements the idea that limited processing resources are traded off between the individual samples in a trial, such that giving extra weight to some samples comes at the cost of discarding others. In a sequential integration framework (cf. Eq. 2), an observer may discard (or downweight) the other samples also retroactively, for

instance, by discarding (downweighting) the evidence accumulated thus far ($\sum dv$) in favor of a later sample (cf. also Eq. 4). When we repeated our simulations with this alternative type of normalization, the performance under noise ($s > 0$) was always maximized by anti-compression ($k > 1$), both when the samples were uniformly or Gaussian distributed (Figure 2f). In other words, for both distributions, anti-compression ($k > 1$) allocated the gain of processing more efficiently among the individual samples in a sequence, resulting in higher task performance without recruiting greater resources.

To summarize our simulation results, whereas compression can improve the performance of a noisy observer by allocating overall more resources to the samples in a trial, anti-compression is favorable when these resources are exceeded, as anti-compression yields higher performance with equivalent resources on the trial level. Importantly, our simulations predict that which type of distortion is favorable in a given task will not primarily depend on the samples' distribution (i.e., uniform or Gaussian; see Li et al., 2017; Spitzer et al., 2017; Summerfield and Li, 2018). Instead, it should hinge on the extent to which limited processing resources need to be traded off between the different samples in a trial. We assume this trade-off to be weak (favouring compression as an optimal policy) in tasks where evaluating the decision value of a sample is relatively easy, and to be stronger (favoring anti-compression) when sample evaluation is more resource-intensive, so that a capacity-limited observer would not be able to integrate every sample in full.

Human Participant Results

We tested whether human participants would adopt performance-maximizing weighting policies in an online experiment ($n = 586$) where we manipulated the sample-level processing demand of the task (Figure 1a). The participants were either asked to judge the average of the whole stream (single-stream task) or to decide whether the red or the blue numbers had a larger average (dual-stream task). In the dual-stream task, whether a given number sample supports one or the other decision depends on its color (see *Methods, Psychometric Model*), posing an additional processing requirement for each sample. Across subgroups of

participants in each task, we further manipulated the range of sample values (1 to 9 or 100 to 900) as well as their distribution (uniform or Gaussian, see Figure 1b).

Descriptive results

Participants' mean accuracy was $79.80 \pm 0.44\%$ in the single-stream task and $75.52 \pm 0.44\%$ in the dual-stream task. A $2 \times 2 \times 2$ ANOVA with the factors "task" (single, dual), "distribution" (uniform, Gaussian), and "range" (small, large) showed a main effect of "task" [$F(1, 578) = 51.67, p < 0.001, \eta^2 = 0.08$], confirming that the dual-stream task was more difficult. In addition, there was a main effect of "distribution" [$F(1, 578) = 34.25, p < 0.001, \eta^2 = 0.05$], indicating higher performance in the Gaussian ($M = 79.41 \pm 0.44\%$) than in the uniform conditions ($M = 75.96 \pm 0.45\%$). No other main effects or interactions were significant (all $F < 1.2$, all $p > 0.28$, all $\eta^2 < 0.002$).

We inspected descriptive weighting functions (see *Methods*) to gauge how strongly each numerical sample value contributed to participants' choices (see Figure 3, solid lines). The weighting functions showed different shapes depending on the task (single- or dual-stream). While a compressive curve (i.e., relatively shallower local slopes near extreme values than near intermediate values) was evident in the single-stream task, an anti-compressive curve (i.e., steeper local slopes near extreme values) emerged in the dual-stream task. Descriptively, this pattern was evident for both small and large sample ranges (see Figure 3a,b and Figure 3c,d, respectively) and for both distribution types (see Figure 3 top and bottom rows).

Modeling results

We next fitted our psychometric model to the empirical data. On average across all participants, the full model showed a better performance (mean BIC = 430.27 ± 6.42) than simpler variants that did not include a leakage ($l = 0$; mean BIC = $449.46 \pm 6.05, Z = 4.48, p < 0.001, r = 0.02$; Wilcoxon signed-rank test) or bias parameter ($b = 0$; mean BIC = $444.73 \pm 6.22, Z = 6.91, p < 0.001, r = 0.02$; Wilcoxon signed-rank test). All subsequent analyses are therefore based on the full model, which includes a leakage and a bias parameter (see

Eq. 1–4).

Our main interest was in the k parameter, which indicates whether there is underweighting ($k < 1$; compression) or overweighting ($k > 1$; anti-compression) of extreme values. The best-fitting estimates of k (Figure 4a) corroborate our observations with the psychometric weighting functions (see Figure 3). In the single-stream task, k was signifi-

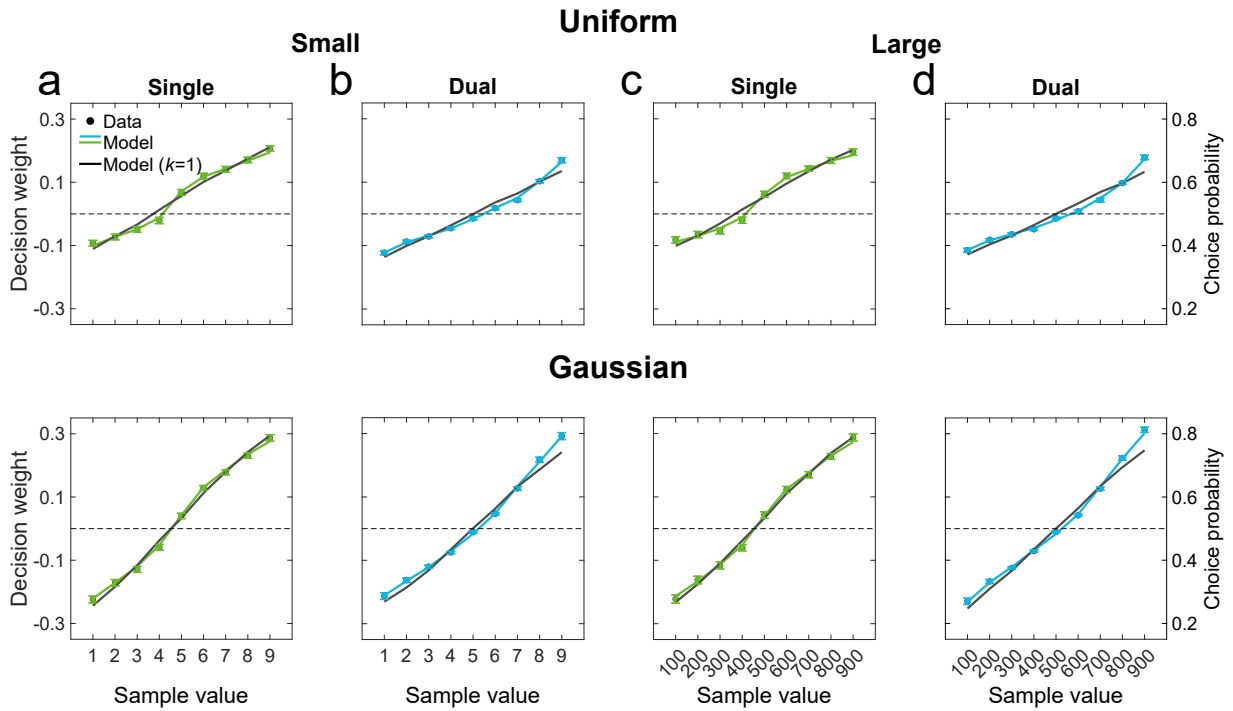


Figure 3. Descriptive weighting functions and model fits. Decision weights (see Methods) for numbers 1 to 9 (small range condition, **a,b**) and 100 to 900 (large range condition, **c,d**). Single-stream task conditions in panels **a** and **c**; Dual-stream task conditions in panels **b** and **d**; Dots: behavioral data; error bars show SEM. Colored (curved) lines: predictions from the fitted psychometric model (see Figure 4). Black (straight) lines show the model predictions for $k = 1$. Dashed horizontal lines indicate indifference (i.e., decision weight = 0 or choice probability = 0.5, see left and right y-axes). Upper panels: uniform conditions; lower panels: Gaussian conditions.

cantly smaller than 1 ($M = 0.66$, $SE = 0.03$; $Z = -11.21$, $p < 0.001$, $r = -0.04$, Wilcoxon signed-rank test against 1), indicating compression. In contrast, in the dual-stream task, k was significantly larger than 1 ($M = 1.86$, $SE = 0.06$; $Z = 11.77$, $p < 0.001$, $r = 0.04$, Wilcoxon signed-rank test against 1), indicating anti-compression. For further inspection, we performed a $2 \times 2 \times 2$ ANOVA (specified as for accuracy above) of the estimates of k in each experimental condition. The analysis showed a main effect of task [single vs. dual: $F(1, 578) = 311.28$, $p < 0.001$, $\eta^2 = 0.35$], but no other main effects or interactions (all $F < 3.46$, all $p > 0.06$, all $\eta^2 < 0.004$). In other words, the type of distortion (compression or anti-compression) was significantly modulated only by the “task” requirement (averaging a single stream or comparing dual streams), and not by the other factors under study (“range” and “distribution”).

We next examined effects of the manipulations on the decision noise parameter s (Figure 4b). A $2 \times 2 \times 2$ ANOVA (specified as above) showed a main effect of “task” [$F(1, 578) = 33.98$, $p < 0.001$, $\eta^2 = 0.05$], reflecting that the dual-stream task was more difficult than the single-stream task (see also *Descriptive Results*; mean s : 1.41, $SE = 0.07$ vs. 0.9, $SE = 0.05$). Further, also mirroring the results for accuracy, there was a main effect of “distribution” [$F(1, 578) = 7.08$, $p = 0.008$, $\eta^2 = 0.01$], with lower s in the Gaussian than in the uniform conditions (mean s : 1.04, $SE = 0.06$, vs. 1.26, $SE = 0.06$). The difference in s (and accuracy) between the two distribution types likely reflects that their difficulty levels could be pre-experimentally matched only in approximation (see *Methods*) based on smaller pilot samples. No other main effects or interactions were significant (all $F(1, 578) < 2.92$, all $p > 0.09$, all $\eta^2 < 0.005$).

Analogous analyses for the bias (b) and leakage (l) parameters showed no differences between conditions (all $F < 1.71$, all $p > 0.19$, all $\eta^2 < 0.003$), with the exception that b differed between tasks [$F(1, 578) = 84.88$, $p < 0.001$, $\eta^2 = 0.13$; see Figure 4]. We refrain from interpreting this latter effect because, for technical reasons, the estimates of b are not

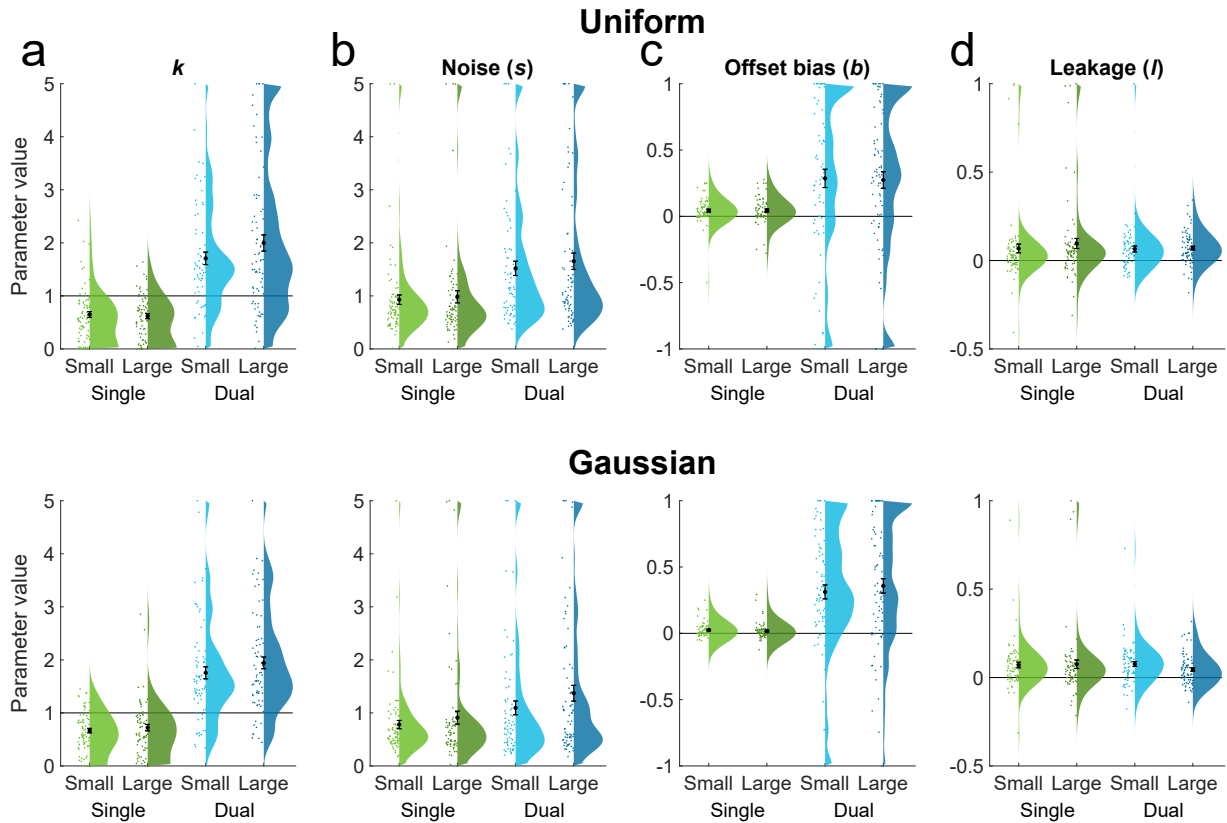


Figure 4. Parameter estimates. Parameter estimates for exponent k (a), noise parameter s (b), offset bias b (c) and leakage l (d). Upper panels: uniform conditions; lower panels: Gaussian conditions. Black dots show the average across individuals, error bars show SEM. Small dots show the parameter estimates for the individual participants. The shaded half-violin outline illustrates the probability density of the parameters, smoothed by a kernel density estimator. Left two violin plots per panel: single-stream task; Right two violin plots per panel: dual-stream task. Light colors: small-range conditions, dark colors: large-range conditions. The black horizontal lines indicate $k = 1$ (no distortion) in a, $b = 0$ (no bias) in c, and $l = 0$ (no leakage) in d.

directly comparable between the two tasks (e.g., b can effect an overall displacement of the psychometric functions in the single- but not in the dual-stream task; see Eq. 1 and 2). For completeness, we report that the b parameter was significantly positive (> 0) in both

tasks (single-stream: $Z = 8.52$, $p < 0.001$, $r = 0.03$, Wilcoxon signed-rank test against 0; dual-stream: $Z = 9.19$, $p < 0.001$, $r = 0.03$, Wilcoxon signed-rank test against 0), consistent with previous findings (Spitzer et al., 2017). Lastly, the leakage parameter l was significantly larger than 0 (indicating greater weighting of later samples) in both tasks (single-stream: $Z = 8.95$, $p < 0.001$, $r = 0.03$, Wilcoxon signed-rank test against 0; dual-stream: $Z = 9.95$, $p < 0.001$, $r = 0.03$, Wilcoxon signed-rank test against 0). Thus, there were generally recency effects in our tasks, consistent with previous findings (Anderson, 1964; Appelhoff et al., 2022a; Cheadle et al., 2014; Hubert-Wallander & Boynton, 2015; Spitzer et al., 2017; Summerfield & Tsetsos, 2015; Weiss & Anderson, 1969; Yashiro et al., 2020).

Correlations between degree of distortion and decision noise

Our simulations not only indicated that the ideal type of distortion (i.e., compression or anti-compression) should depend on the task requirements, but also that the degree of distortion (of either type) should increase with the level of decision noise (s) (see Figure 2d and f). Hence, if participants adopted ideal weighting policies given their individual noise levels, we would expect to observe opposite correlations between the distortion parameter k and noise s in the two tasks. Specifically, for participants with higher noise levels (s), estimates of k should be *lower* ($k \ll 1$, stronger compression) in the single-stream task and *higher* ($k \gg 1$, stronger anti-compression) in the dual-stream task. Our data support this prediction: There was a negative correlation between k and s in the single-stream task (Figure 5a, $r = -0.32$, $p < 0.001$), but a positive correlation in the dual-stream task (Figure 5b, $r = 0.24$, $p < 0.001$). Both correlations were robust to the exclusion of outliers near the parameter boundaries (excluding data points < 0.1 or > 4.9 in either k or s : single-stream: $r = -0.25$, $p < 0.001$; dual-stream: $r = 0.18$, $p = 0.003$). As the correlations were of opposite signs, they are unlikely to be due to parameter interdependencies (Krefeld-Schwab et al., 2022, see *Methods, Parameter Recovery*). Together, these results empirically corroborate the complex relationship between psychometric distortions and decision noise that we identified in our simulations of theoretically ideal policies (Figure 2).

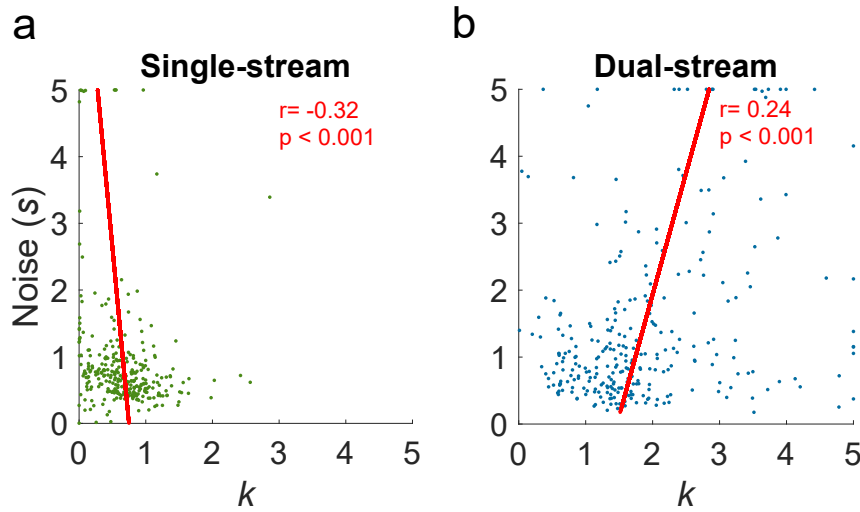


Figure 5. *Correlations between k and noise (s) across participants in the (a) single-stream and (b) dual-stream tasks. Lines show the linear trend. Higher levels of decision noise were associated with stronger compression in the single-stream task, but with stronger anti-compression in the dual-stream task.*

Comparing experimental results with predicted optimal distortions

To compare participants' behavior with the results of our simulations (Figure 2d–f), we repeated our model fitting with the respective normalizations illustrated in Figure 2a–c (see also *Methods*, Eq. 3). The results for the single-stream task matched reasonably well with our simulations without gain limitations (2d), as we would expect given that sample evaluation in this task was relatively easy. In the dual-stream task, in contrast, there was a trend towards the ideal solution under a trial-level gain limit (for the uniform and Gaussian conditions alike; Figure 2f), as we would expect given that the higher demands of this task forced participants to trade off processing resources between samples. Quantitatively, the degree of either type of distortion (compression or anti-compression) fell short of the model-predicted optimum under the respective noise level (see Figures 2e and f). However, our model simulations only delineate the endpoints of a continuum (from no to full exhaustion of sample-level resources) on which we assume our experimental tasks to differ. It would be

relatively straightforward to extend our framework to simulate any position a task may take between these two extremes [e.g., through parameterization of g : $g_{\text{partial}} = 1 + \alpha(g - 1)$, where α (ranging from 0 to 1) would reflect the extent to which processing resources are exhausted].

Complementary Analyses

We performed supplementary analyses to back up our key empirical finding of compression in the single-stream task and anti-compression in the dual-stream task. First, we used Bayesian t-tests to corroborate that the distortion parameter k was only modulated by the factor “task” (single- or dual-stream) and not by the other factors under study (“distribution” and “range”). The Bayes factors showed extreme evidence for an effect of the factor “task” (single- or dual-stream; $BF_{10} > 100$), but moderate evidence for the null hypothesis (no effect) when examining the factor “distribution” ($BF_{10} = 0.11$), and anecdotal evidence for the null hypothesis when examining the factor “range” ($BF_{10} = 0.604$).

Second, we directly tested for differences in the descriptive weighting functions’ local slopes, to examine whether they showed the hallmarks of compression or anti-compression in the single- and dual-stream conditions, respectively (collapsed across the other manipulations). Specifically, we compared the mean local slope of outlying values (e.g., 1-2 and 8-9) against that of inlying values (e.g., 4-5 and 5-6) using paired t-tests. As expected, the difference in local slopes was significantly negative (i.e., steeper for inlying values) in the single-stream task ($M = -0.041$, $SE = 0.003$, $t(295) = -12.31$, $p < 0.001$), but significantly positive (i.e., steeper for outlying values) in the dual-stream task ($M = 0.017$, $SE = 0.003$, $t(289) = 5.55$, $p < 0.001$). Thus, the key features of compression or anti-compression were evident even in model-free analyses of the data.

Finally, given the observation of an overall bias towards larger numbers (i.e., $b > 0$; see also Spitzer et al., 2017), we asked whether the above slope differences were only driven by large sample values. That did not seem to be the case, as the same pattern was evident

when restricting the analysis to the lower end of the value range (e.g., comparing 1-2 vs. 3-4; single-stream task: $M = -0.01$, $SE = 0.004$, $t(295) = -2.23$, $p = 0.026$; dual-stream task: $M = 0.01$, $SE = 0.005$, $t(289) = 2.22$, $p = 0.027$). Together, these supplementary results support the main findings from our computational modeling analysis.

Discussion

Evaluating samples of magnitude, such as in decisions based on numbers, is integral to adaptive human behavior. Previous research has found evidence for opposite types of distortion of numerical values—compression and anti-compression—in tasks requiring the integration of number sequences (Li et al., 2017; E. Ludvig et al., 2014; E. A. Ludvig et al., 2014; Spitzer et al., 2017; Vanunu et al., 2019). Here, we showed that whether people subjectively compress or anti-compress numerical values depends on whether they are asked to assess the average value of a single stream or to compare the values of two interleaved streams. Arguably, the latter task is cognitively more effortful, because evaluating a sample’s decision value for the comparison requires more cognitive operations (see also Appelhoff et al., 2022b). The pattern of results matches the predictions of our simulations with a psychometric model, which showed that compression yields a performance benefit for noisy observers when tasks are within their processing limit, whereas anti-compression improves performance in computationally demanding tasks (i.e., where evaluating a sample properly comes at the cost of missing the decision information in other samples). Taken together, our results suggest that participants adopted a favorable weighting policy in the respective task context, given their capacity limitations—in other words, that their choice of weighting policy was adaptive.

Our findings speak to the long-standing question as to *why* people distort objective magnitude information in decision making. It has recently been proposed that well-known distortions, such as those of outcome and probability information in choices between monetary gambles (Kellen et al., 2016; Tversky & Kahneman, 1992), may serve a rather basic goal, namely to maximize objective returns (Juechems et al., 2021; Spitzer et al., 2017). A

central insight from this recent work has been that for noisy observers (e.g., humans and other biological agents), distortions of sample information can lead to *higher* returns than a normatively correct linear mapping (Juechems et al., 2021; Li et al., 2017). Here, we extended this approach by showing that very different types of distortion (i.e., compression or anti-compression) can optimize the performance of noisy observers, depending on the extent to which a task taxes their processing capacities.

Importantly, the basic shape of distortion (compression or anti-compression) was not related to the overall difficulty of making a choice. In our simulations, the ideal extent of either type of distortion increased with higher levels of late decision noise, that is, with overall declining accuracy. Likewise, in our empirical data, noisier participants (with lower accuracy) showed stronger compression in simple numerical averaging, but stronger anti-compression in the more effortful dual-stream comparison task. As an explanation for this pattern, we propose that limitations (or bottlenecks) at different stages of the processing hierarchy may impact differently on the shape of psychometric distortions.

Previous work has suggested a distinction between “early” sensory noise (e.g., due to limits in sensory acuity; Lavie & Fockert, 2003; Pelli, 1991; Treisman & Geffen, 1967) and “late” decision noise (e.g., due to the difficulty of integrating multiple feature values into a binary response; Baek & Chong, 2020; Drugowitsch et al., 2016; Findling & Wyart, 2021; Juslin & Olsson, 1997; Solomon, 2020; Summerfield & Parpart, 2021) as limiting factors in human decision making. The present findings highlight another processing bottleneck intermediate to these early and late processing stages: the difficulty of evaluating the decisional *meaning* of a sample within the context of a given task. We assume this processing to occur after sensory encoding and prior to combining the information from different samples into a final response. Tasks that load strongly on this intermediate (sample-by-sample) bottleneck may enforce a trade-off of processing resources *between* samples and promote selective integration (Tsetsos et al., 2016) of extreme values (i.e., anti-compression) as a performance-maximizing

policy. Protection against late decision noise, in contrast, can be achieved through stronger distortions of either type (see also Li et al., 2017; Spitzer et al., 2017), depending on task context. In this framework, the optimal weighting policy under late noise may even be linear (undistorted; as observed in, e.g., Kang & Spitzer, 2021), namely, if a task poses moderate sample-level demands.

Contrary to expectations based on previous work (Li et al., 2017; Prat-Carrabin & Woodford, 2020; Spitzer et al., 2017; Summerfield & Li, 2018), neither the distribution of sample values (uniform or Gaussian) or their “range” (small or large) had an impact on participants’ weighting policy. This observation is consistent with the results from our model simulations, which assume an upper bound on the processing resources available to an observer on any given *trial* (cf. Li et al., 2017; Spitzer et al., 2017). It remains to be shown whether our experimental results with symbolic numbers generalize to other input formats (e.g., sensory-perceptual information Liu et al., 2015; Marinova et al., 2020; Pekár & Kinder, 2019; Rosenbaum et al., 2021; Sato & Motoyoshi, 2020), where the distribution and range of input values may play an additional role. For instance, while the discrete numerical samples in our experiment were easily readable (i.e., early sensory noise was presumably negligible), other sensory-perceptual inputs may be more prone to, for example, range adaptation effects (Brenner et al., 2000; Fairhall et al., 2001; Smirnakis et al., 1997; Wark et al., 2007), which might also impact the shape of psychometric weighting.

An alternative explanation for our empirical results warrants consideration. In our single-stream task, numbers had to be evaluated against a fixed reference value (e.g., 5), whereas in our dual-stream task, two streams of numbers had to be contrasted. One might thus argue that the finding of compression or anti-compression did not depend on sample-level demands, but on whether the samples had to be evaluated against a fixed or a variable reference value. In light of recent studies by Rosenbaum et al. (2021) and Vanunu et al. (2019), such alternative explanation for our results seems unlikely. These studies examined

decision behavior in tasks that required evaluation within a fixed reference frame (like our single-stream task) but under conditions that were computationally demanding. The authors observed anti-compression, which is consistent with our interpretation of the present results in terms of processing demands.

For completeness we note that in addition to our main finding of adaptive distortions, participants' decisions also showed characteristics that were not encompassed by our model simulations: a "leakage" of sample information over time (i.e., a "recency" bias towards later presented samples), and an overall bias towards larger numbers (e.g., choices were more strongly driven by sample values "9" than "1", although the latter provided equally strong objective evidence). Both of these biases have been reported repeatedly in previous work (Anderson, 1964; Appelhoff et al., 2022a, 2022b; Cheadle et al., 2014; Hubert-Wallander & Boynton, 2015; Luyckx et al., 2019; Spitzer et al., 2017; Summerfield & Tsetsos, 2015; Weiss & Anderson, 1969; Yashiro et al., 2020), but their precise origin and functional role remain unclear. These biases were not modulated in interpretable ways by our present experimental manipulations, leaving their further investigation to future work.

A limitation of our simulation framework is that it cannot be used to predict the extent to which a given task will exhaust an observer's sample-level processing capacities. It thus remains difficult to determine *a priori* which kind of distortion (compression or anti-compression) would maximize a noisy observer's returns. A interesting direction for future research could be to quantify, using simulations and/or neuroscientific approaches, the extent to which processing resources are expected to be traded off between the samples in a given task, as hypothesized here. A related question for future work is how participants may have learned to use different weighting policies in different tasks contexts. Whereas our model simulations identified ideal policies through objective performance maximization, the computations by which human participants select their weighting policy for a given task may be different. Furthermore, while our study highlights sample-level processing demands

as a key determinant of psychometric distortions, other factors may also play a role (Pachur et al., 2018; Pachur et al., 2017; Rosenbaum et al., 2021; Vanunu et al., 2020). For instance, Rosenbaum et al. (2021) showed that the type of stimulus information (numerical or sensory-perceptual; which is unaccounted for in our model) can alter the weighting of samples in ensemble judgments. Finally, while our framework formally describes subjective distortions as parameterized psychometric functions, similar behaviors might arise from (mixtures of) heuristic policies (Gigerenzer et al., 2011), such as selective counting of sample values that a participant deems diagnostic in a given trial.

In conclusion, our work offers a theoretical framework and empirical data to explain why decision makers may over- or underweight extreme values in decisions based on sequential samples. Rather than reflecting idiosyncratic quirks of the human mind, subjective distortions of decision information may improve the objective performance of capacity-limited observers. Our results reconcile conflicting findings about the form of such performance-maximizing distortions and suggest that human participants intuitively adopt a decision policy that is beneficial for the task at hand.

Constraints on Generality

Our study examined a large international sample of adult participants aged between 18 and 41 years who in all likelihood possessed basic numeracy skills. Participants completed the study remotely via a web browser. We have no reason to assume that different results would be obtained in a laboratory setting. In terms of materials and stimuli, we used ranges of symbolic numbers with a limited granularity of 9 discrete steps. Whether or not the results generalize to more finely sampled number ranges, to non-symbolic numbers, and/or to non-numerical magnitudes yet remains to be shown.

References

- Alonso, R., Brocas, I., & Carrillo, J. (2011). Resource allocation in the brain. *Review of Economic Studies*, *81*(2), 501–534. <https://doi.org/10.1093/restud/rdt043>
- Anderson, N. H. (1964). Test of a model for number-averaging behavior. *Psychonomic Science*, *1*(7), 191–192. <https://doi.org/10.3758/BF03342858>
- Appelhoff, S., Hertwig, R., & Spitzer, B. (2022a). Control over sampling boosts numerical evidence processing in human decisions from experience. *Cerebral Cortex*, *33*(1), 207–221. <https://doi.org/10.1093/cercor/bhac062>
- Appelhoff, S., Hertwig, R., & Spitzer, B. (2022b). Eeg-representational geometries and psychometric distortions in approximate numerical judgment. *PLOS Computational Biology*, *18*(12), 1–19. <https://doi.org/10.1371/journal.pcbi.1010747>
- Baddeley, R., Hancock, P., & Földiák, P. (Eds.). (2000). *Information theory and the brain*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511665516>
- Baek, J., & Chong, S. C. (2020). Distributed attention model of perceptual averaging. *Attention, Perception, & Psychophysics*, *82*, 63–79. <https://doi.org/10.3758/s13414-019-01827-z>
- Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). MIT Press. <https://doi.org/10.7551/mitpress/9780262518420.003.0013>
- Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, *22*(1), 23–36. <http://www.jstor.org/stable/1909829>
- Bhui, R., & Gershman, S. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, *125*(6), 985–1001. <https://doi.org/10.1037/rev0000123>
- Birnbaum, M. H., & Chavez, A. (1997). Tests of theories of decision making: Violations of branch independence and distribution independence. *Organizational Behavior and human decision Processes*, *71*(2), 161–194.

- Brenner, N., Bialek, W., & Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, *26*(3), 695–702. [https://doi.org/10.1016/S0896-6273\(00\)81205-2](https://doi.org/10.1016/S0896-6273(00)81205-2)
- Cheadle, S., Wyart, V., Tsetsos, K., Myers, N., de Gardelle, V., Castañón, S. H., & Summerfield, C. (2014). Adaptive gain control during human perceptual choice. *Neuron*, *81*(6), 1429–1441. <https://doi.org/10.1016/j.neuron.2014.01.020>
- Clarmann von Clarenau, V., Appelhoff, S., Pachur, T., & Spitzer, B. (2022). *Over- and underweighting of extreme values [data set]*. <https://doi.org/10.17605/OSF.IO/X83PK>
- Cokely, E., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*(1), 25–47.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. <https://doi.org/10.1017/S0140525X01003922>
- de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, *108*(32), 13341–13346. <https://doi.org/10.1073/pnas.1104517108>
- Drugowitsch, J., Wyart, V., Lodeho-Devauchelle, A.-D., & 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>
- Eldar, E., Cohen, J., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. <https://doi.org/10.1038/nn.3428>
- Fairhall, A., Lewen, G., Bialek, W., & Steveninck, R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412*, 787–792. <https://doi.org/10.1038/35090500>
- Fechner, G. (1860). *Elemente der Psychophysik [Elements of psychophysics]*. Breitkopf und Härtel. <https://books.google.de/books?id=6rINAAAYAAJ>

- Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: Origin, impact, function. *Current Opinion in Behavioral Sciences*, *38*, 124–132. <https://doi.org/10.1016/j.cobeha.2021.02.018>
- Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*. Oxford University Press.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. [https://doi.org/10.1016/0010-0285\(92\)90002-J](https://doi.org/10.1016/0010-0285(92)90002-J)
- Hubert-Wallander, B., & Boynton, G. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision*, *15*, Article 5. <https://doi.org/10.1167/15.4.5>
- Juechems, K., Balaguer, J., Spitzer, B., & Summerfield, C. (2021). Optimal utility and probability functions for agents with finite computational precision. *Proceedings of the National Academy of Sciences*, *118*(2), Article e2002232118. <https://doi.org/10.1073/pnas.2002232118>
- Juslin, P., & Olsson, H. (1997). Thurstonian and brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological review*, *104*, 344–66. <https://doi.org/10.1037/0033-295X.104.2.344>
- Kang, Z., & Spitzer, B. (2021). Concurrent visual working memory bias in sequential integration of approximate number. *Scientific Reports*, *11*, Article 5348. <https://doi.org/10.1038/s41598-021-84232-7>
- Katzin, N., Rosenbaum, D., & Usher, M. (2021). The averaging of numerosities: A psychometric investigation of the mental line. *Attention, Perception & Psychophysics*, *83*(3), 1152–1168. <https://doi.org/10.3758/s13414-020-02140-w>
- Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards? *Cognition*, *157*, 126–138. <https://doi.org/10.1016/j.cognition.2016.08.020>

- Kostal, L., Lánský, P., & McDonnell, M. D. (2013). Metabolic cost of neuronal information in an empirical stimulus-response model. *Biological Cybernetics*, *107*(3), 355–365. <https://doi.org/10.1007/s00422-013-0554-6>
- Krefeld-Schwalb, A., Pachur, T., & Scheibehenne, B. (2022). Structural parameter interdependencies in computational models of cognition. *Psychological Review*, *129*(2), 313.
- Krekelberg, B. (2022). *Bayesfactor: Release 2022 (v2.3.0)* (Version v2.3.0). Zenodo. <https://doi.org/10.5281/zenodo.7006300>
- Kunar, M. A., Watson, D. G., Tsetsos, K., & Chater, N. (2017). The influence of attention on value integration. *Attention, Perception, & Psychophysics*, *79*(6), 1615–1627. <https://doi.org/10.3758/s13414-017-1340-7>
- Lavie, N., & Fockert, J. (2003). Contrasting effects of sensory limits and capacity limits in visual selective attention. *Perception & Psychophysics*, *65*(2), 202–212. <https://doi.org/10.3758/BF03194795>
- Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*(6), 493–497. [https://doi.org/10.1016/S0960-9822\(03\)00135-0](https://doi.org/10.1016/S0960-9822(03)00135-0)
- Li, V., Hecce Castañón, S., Solomon, J. A., Vandormael, H., & Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. *PLOS Computational Biology*, *13*(8), Article e1005723. <https://doi.org/10.1371/journal.pcbi.1005723>
- Liu, A. S., Schunn, C. D., Fiez, J. A., & Libertus, M. E. (2015). Symbolic integration, not symbolic estrangement, for double-digit numbers. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1404–1409). The Cognitive Science Society. <https://mindmodeling.org/cogsci2015/papers/0246/index.html>
- Ludvig, E., Madan, C., & Spetch, M. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making*, *27*, 146–156. <https://doi.org/10.1002/bdm.1792>

- Ludvig, E. A., Madan, C. R., McMillan, N., Xu, Y., & Spetch, M. L. (2018). Living near the edge: How extreme outcomes and their neighbors drive risky choice. *Journal of Experimental Psychology: General*, *147*(12), 1905.
- Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making*, *27*(2), 146–156.
- Luyckx, F., Nili, H., Spitzer, B., & Summerfield, C. (2019). Neural structure mapping in human probabilistic reward learning (D. Lee, J. I. Gold, D. Lee, & M. Chafee, Eds.). *eLife*, *8*, e42816. <https://doi.org/10.7554/eLife.42816>
- Marinova, M., Sasanguie, D., & Reynvoet, B. (2020). Numerals do not need numerosities: Robust evidence for distinct numerical representations for symbolic and non-symbolic numbers. *Psychological Research*, *85*(2), 764–776. <https://doi.org/10.1007/s00426-019-01286-z>
- Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, *9*(6), 296–305. <https://doi.org/10.1016/j.tics.2005.04.010>
- Neri, P., Parker, A. J., & Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. *Nature*, *401*(6754), 695–698.
- Pachur, T., Schulte-Mecklenbeck, M., Murphy, R. O., & Hertwig, R. (2018). Prospect theory reflects selective allocation of attention. *Journal of Experimental Psychology: General*, *147*(2), 147–169. <https://doi.org/10.1037/xge0000406>
- Pachur, T., Suter, R. S., & Hertwig, R. (2017). How the twain can meet: Prospect theory and models of heuristics in risky choice. *Cognitive Psychology*, *93*, 44–73. <https://doi.org/10.1016/j.cogpsych.2017.01.001>
- Pekár, J., & Kinder, A. (2019). The interplay between non-symbolic number and its continuous visual properties revisited: Effects of mixing trials of different types. *Quarterly Journal of Experimental Psychology*, *73*(5), 698–710. <https://doi.org/10.1177/1747021819891068>

- Pelli, D. (1991). Noise in the visual system may be early. In M. A. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 147–151). MIT Press.
- Prat-Carrabin, A., & Woodford, M. (2020). Efficient coding of numbers explains decision bias and noise. *bioRxiv*. <https://doi.org/10.1101/2020.02.18.942938>
- Rosenbaum, D., de Gardelle, V., & Usher, M. (2021). Ensemble perception: Extracting the average of perceptual versus numerical stimuli. *Attention, Perception, & Psychophysics*, *83*, 956–969. <https://doi.org/10.3758/s13414-020-02192-y>
- Sato, H., & Motoyoshi, I. (2020). Distinct strategies for estimating the temporal average of numerical and perceptual information. *Vision Research*, *174*, 41–49. <https://doi.org/10.1016/j.visres.2020.05.004>
- Shevlin, B. R., Smith, S. M., Hausfeld, J., & Krajbich, I. (2022). High-value decisions are fast and accurate, inconsistent with diminishing value sensitivity. *Proceedings of the National Academy of Sciences*, *119*(6), e2101508119.
- Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14* (3), 237–250. <https://doi.org/10.1111/1467-9280.02438>
- Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W., & Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. *Nature*, *386*, 69–73. <https://doi.org/10.1038/386069a0>
- Solomon, J. (2020). Five dichotomies in the psychophysics of ensemble perception. *Attention, Perception, & Psychophysics*, *83*, 904–910. <https://doi.org/10.3758/s13414-020-02027-w>
- Spitzer, B., Blankenburg, F., & Summerfield, C. (2016). Rhythmic gain control during supramodal integration of approximate number. *NeuroImage*, *129*, 470–479. <https://doi.org/https://doi.org/10.1016/j.neuroimage.2015.12.024>

- Spitzer, B., Waschke, L., & Summerfield, C. (2017). Selective overweighting of larger magnitudes during noisy numerical comparison. *Nature Human Behaviour*, *1*, Article 0145. <https://doi.org/10.1038/s41562-017-0145>
- Summerfield, C., & Li, V. (2018). Perceptual suboptimality: Bug or feature? *Behavioral and Brain Sciences*, *41*, Article e245. <https://doi.org/10.1017/S0140525X18001437>
- Summerfield, C., & Parpart, P. (2021). Normative principles for decision-making in natural environments. *PsyArXiv Preprints*. <https://doi.org/10.31234/osf.io/s2wvz>
- Summerfield, C., & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences*, *19*(1), 27–34. <https://doi.org/10.1016/j.tics.2014.11.005>
- The MathWorks, I. (2020a). *9.9.0.2037887 (r2020b)*. Natick, Massachusetts, United States. <https://www.mathworks.com>
- The MathWorks, I. (2020b). *Econometrics toolbox*. Natick, Massachusetts, United State. <https://de.mathworks.com/help/econ/>
- The MathWorks, I. (2020c). *Optimization toolbox*. Natick, Massachusetts, United State. <https://de.mathworks.com/help/parallel-computing/>
- The MathWorks, I. (2020d). *Optimization toolbox*. Natick, Massachusetts, United State. <https://de.mathworks.com/help/optim/>
- The MathWorks, I. (2020e). *Statistics and machine learning toolbox*. Natick, Massachusetts, United State. <https://de.mathworks.com/help/stats/>
- Tombu, M., Asplund, C., Dux, P., Godwin, D., Martin, J., & Marois, R. (2011). A unified attentional bottleneck in the human brain. *Proceedings of the National Academy of Sciences*, *108*(33), 13426–13431. <https://doi.org/10.1073/pnas.1103583108>
- Treisman, A., & Geffen, G. M. (1967). Selective attention: Perception or response? *Quarterly Journal of Experimental Psychology*, *19*(1), 1–17. <https://doi.org/10.1080/14640746708400062>

- Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences*, *109*(24), 9659–9664. <https://doi.org/10.1073/pnas.1119569109>
- Tsetsos, K., Moran, R., Moreland, J. C., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, *113*(11), 3102–3107. <https://doi.org/10.1073/pnas.1519157113>
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. <https://EconPapers.repec.org/RePEc:kap:jrisku:v:5:y:1992:i:4:p:297-323>
- Vandormael, H., Castañón, S., Balaguer, J., Li, V., & Summerfield, C. (2017). Robust sampling of decision information during perceptual choice. *Proceedings of the National Academy of Sciences*, *114*(10), 2771–2776. <https://doi.org/10.1073/pnas.1613950114>
- Vanunu, Y., Hotaling, J., & Newell, B. (2020). Elucidating the differential impact of extreme-outcomes in perceptual and preferential choice. *Cognitive Psychology*, *119*, Article 101274. <https://doi.org/10.1016/j.cogpsych.2020.101274>
- Vanunu, Y., Pachur, T., & Usher, M. (2019). Constructing preference from sequential samples: The impact of evaluation format on risk attitudes. *Decision*, *6*(3), 223–236. <https://doi.org/10.1037/dec0000098>
- Wakker, P., & Deneffe, D. (1996). Eliciting von neumann-morgenstern utilities when probabilities are distorted or unknown. *Management Science*, *42*(8), 1131–1150.
- Wark, B., Lundstrom, B., & Fairhall, A. (2007). Sensory adaptation. *Current Opinion in Neurobiology*, *17*(4), 423–429. <https://doi.org/10.1016/j.conb.2007.07.001>
- Weiss, D. J., & Anderson, N. H. (1969). Subjective averaging of length with serial presentation. *Journal of Experimental Psychology*, *82*(1), 52–63. <https://doi.org/10.1037/h0028028>

Yashiro, R., Sato, H., Oide, T., & Motoyoshi, I. (2020). Perception and decision mechanisms involved in average estimation of spatiotemporal ensembles. *Scientific Reports*, *10*, Article 1318. <https://doi.org/10.1038/s41598-020-58112-5>